Tangent Bundle Curve Completion with Locally Connected Parallel Networks

Guy Ben-Yosef
guybeny@cs.bgu.ac.il
Ohad Ben-Shahar
ben-shahar@cs.bgu.ac.il
Computer Science Department and Zlotowski Center for Neuroscience,
Ben-Gurion University, Beer-Sheva 84105, Israel

We propose a theory for cortical representation and computation of visually completed curves that are generated by the visual system to fill in missing visual information (e.g., due to occlusions). Recent computational theories and physiological evidence suggest that although such curves do not correspond to explicit image evidence along their length, their construction emerges from corresponding activation patterns of orientation-selective cells in the primary visual cortex. Previous theoretical work modeled these patterns as least energetic 3D curves in the mathematical continuous space $R^2 \times \mathcal{S}^1$, which abstracts the mammalian striate cortex. Here we discuss the biological plausibility of this theory and present a neural architecture that implements it with locally connected parallel networks. Part of this contribution is also a first attempt to bridge the physiological literature on curve completion with the shape problem and a shape theory. We present completion simulations of our model in natural and synthetic scenes and discuss various observations and predictions that emerge from this theory in the context of curve completion.

1 Introduction and Problem Formulation

Visual curve completion is a phenomenon in which the visual system fills in the missing parts between observed boundary fragments to facilitate a perception of whole objects. This core aspect of perceptual organization has been studied by vision scientists for over a century, where initial demonstrations of the phenomenon, dubbed *illusory contours* (Schumann, 1904), *anomalous contours* (Lawson & Gulick, 1967), and *cognitive contours* (Gregory, 1972), have evolved into a relatively coherent discussion in Kanizsa's subjective contours (see Figures 1A and 1B) and his theory of modal completion (Kanizsa, 1979). Kanizsa (1979) also emphasized another form of visual completion, amodal completion (see Figures 1C and 1D), and leveraged his demonstrations to argue that the completion process is,

Neural Computation **24, 3277–3316** (2012) © 2012 Massachusetts Institute of Technology

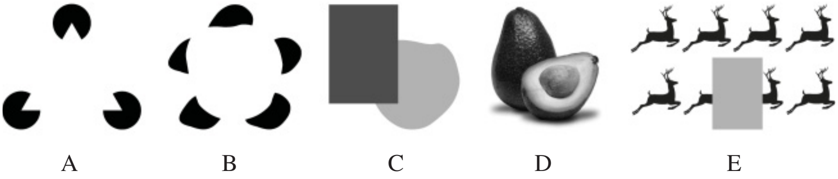

Figure 1: Phenomenology of visual curve completion. (A, B). Modal completion gives rise to "contours without gradient" (Kanizsa, 1979) and to particular interpretations of stimuli that involve occlusion by an illusory figure. In these cases, the illusory triangle and apple-like shapes are clearly perceived, although more than 50% of their contours are formed by zero luminance contrast. (C, D) Amodal completion refers to curve completion of object boundaries behind real occluders, or in Kanizsa's own words, to contours whose "perceptual existence is not verified by any sensory modality" (Kanizsa, 1979, p.6). (E) An important part of the completion mechanism is low level and bottom up. Here the perceptual result is of an excessively long deer, which no observer has likely seen before. In this case, completion based on good continuation seems to override a lifetime of visual experience. (Inspired by Kanizsa, 1979)

at least in part, low level, bottom up, and geometrically stable, in a way that often overrides context or visual experience (see Figure 1E). Since then, the problem of curve completion has taken a central role in early vision research, not only in terms of human perception but in computational vision and visual neuroscience.

As one can easily appreciate, visually completed curves are elicited by certain image contour fragments and depend on an appropriate grouping between pairs of contour fragments, or what is commonly referred to as *inducers*. Indeed, certain inducer pairs induce completion, while others do not (e.g., see Figure 2A). While the grouping problem (i.e., the problem of which pairs of inducers trigger completion) is a fundamental and difficult problem, here we focus on its companion problem, the shape problem, which deals with understanding, modeling, and predicting the shape of perceptually completed curves assuming that two inducers are given (see Figure 2B).

The first difficulty in addressing the shape problem lies in the proper characterization of the desired solution, the latter being confounded by the difficulty of measuring the exact shape of perceptually completed curves with psychophysical techniques. Furthermore, without reasonable assumptions or knowledge on the class of curves that perceptual completions could belong to, the interpretation itself of experimental data becomes problematic. Hence, at the very least, one needs to assume one or more generating principles, from which a rigorous mathematical description can be derived and experimentally measured data can be interpreted.

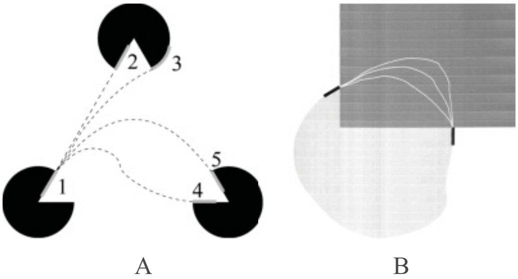

Figure 2: The grouping and the shape problems. (A). The grouping problem. Only specific pairs of contour fragments indeed induce completion. Here, for example, while fragment 1 appears to team up with fragment 2 to generate a completed curve, this pairing does not happen with fragments 3, 4, 5, or 6. (B). The shape problem. Suppose we were able to identify a pair of contour fragments that induce completion (in black). Out of infinite many possible curves connecting between these two fragments (some of them are plotted here in gray), what is the unique shape chosen by the visual system? You can test your perception in Figure 1C.

Alternatively, one may devise assumptions directly on the shape that is expected in the completion process, although to avoid the risk of being arbitrary, such proper shape axiomatization must rely on experimental data again. In either case, in our view, the exploration of the shape problem in curve completion must rely on the entire body of experimental findings accumulated in the past three decades, including those involving behavioral, psychophysical, single-cell recordings, fMRI, and EEG experiments, to name but a few. In other words, a good modeling approach must be multidisciplinary to combine evidence from all of the psychophysical, neurophysiological, and computational vision sciences.

Following such a multidisciplinary approach, we recently proposed a curve completion theory based on an abstraction of the completion process directly in the primary visual cortex (see section 2), namely the mathematical space $\mathbf{R}^2 \times \mathcal{S}^1$, also known in modern differential geometry as the *unit tangent bundle*, which is associated with the image plane (or retinal field) $\mathbf{R}^2$. Employing both physical and Gestalt arguments, we proposed that perceptually completed curves are formed by particular neural activation patterns in V1 that can be abstracted as 3D lifted curves of unique properties in the tangent bundle space (see below). Following this continuous abstraction, its mathematical analysis, and its corresponding curve completion algorithms, in this letter, we explore their biological plausibility and develop a network architecture that implements this approach, all motivated by a large number of reported findings in the physiological literature, some of which are linked here to rigorous shape completion theory for the first time.

To get a proper formulation of the shape problem and without denying that high-level and top-down factors may also play a role, we begin by asking what information is being conveyed by the inducers to the bottom-up components of the completion mechanism (cf. Kanizsa, 1979; Kellman & Shipley, 1991; section 2.3). Addressing this issue, the curve completion literature has focused on the local information that can be extracted around the points of occlusion, which in most previous studies has been taken to be the position and orientation of each fragment (Kellman & Shipley, 1991; Guttman & Kellman, 2004; Fulvio, Singh, & Maloney, 2008; Ullman, 1976; Horn, 1983; Williams & Jacobs, 1997b; Kimia, Frankel, & Popescu, 2003; Sharon, Brandt, & Basri, 2000). Formally, then, one can consider the following (admittedly ill-posed) problem definition:

Problem 1. Given the position and orientation of two inducers $p_0 = [x_0, y_0; \theta_0]$ and $p_1 = [x_1, y_1; \theta_1]$ in the image plane, find the "correct" shape of the perceptual curve that passes between these inducers.[1]

Naturally, the problematic term correct is a degree of freedom that obtains its meaning according to the selected completion principle. Still, any formalization of what is "correct" should clearly agree and be motivated by relevant experimental findings, including the constraints of multiple visual areas. Unfortunately, most of the proposed computational theories to date fall short of doing so.

2 Background and Related Work

2.1 Computational Theories. Perhaps the first computational model ever proposed in the context of problem 1 is the biarc curve model suggested by Ullman (1976). In his seminal work, Ullman suggested that the "correct" shape is one that satisfies four geometrical or perceptual properties: isotropy, smoothness, total minimum curvature, and extensibility. In seeking the curve that uniquely satisfies these axiomatic properties, Ullman further suggested that the completed shape between two given inducers consists of two circular arcs, each tangent to both an inducer and the other arc. Since the number of such biarc pairs for given inducers is infinite, the selected biarc is the one that generates the minimal total curvature.

Ullman did not present a closed-form solution to his biarc model (a mathematical solution based on a one-dimensional nonlinear optimization was introduced later by Rutkowski, 1979), but he did suggest a solver based on a parallel network of simple computational nodes, reminiscent of the

[1] Few psychophysical studies suggest that the curvature of the observed fragment at the point of occlusion also affects the completed shape (Takeichi, 1995; Singh & Fulvio, 2005). While this possibility is not covered by our computational model, it is definitely viable, and we discuss it in section 4.

computational infrastructure in the primary visual cortex. His network has three layers. The first two are designated for computing all possible arcs, leaving the two inducers. The third layer uses a simple summation to assign a degree of smoothness to each pair of arcs, that is, a weight that indicates its total curvature (note that in this way, pairs of arcs that are not tangent to each other get lower weight). Following this summation, nonmaxima suppression is used to select the pair with minimal total curvature, which defines the final outcome. Unfortunately, to our best knowledge, this model was never implemented or evaluated against perceptual data.

Ullman's model has inspired a wealth of shape completion studies, each adopting one of two aspects of his work. The first, and more frequently followed, aspect is what was called the axiomatic approach (Ben-Yosef & Ben-Shahar, 2012): a quest for the unique completed curve that satisfies a specified set of predefined desired perceptual characteristics. In addition to studying some of Ullman's axioms in a stricter and more rigorous fashion (e.g., minimum total curvature via elastica; Horn, 1983; Mumford, 1994), subsequent studies have also suggested a variety of other desired axioms, such as scale invariance (Weiss, 1988), roundedness (Kimia et al., 2003), minimum total change of curvature (Kimia et al., 2003), or specific combinations of them (Bruckstein & Netravali, 1990; Sharon et al., 2000). However, since some axioms conflict with others and some axioms conflict with psychophysical findings (see the elaborate discussion in Ben-Yosef & Ben-Shahar, 2012), the axiomatic approach has invigorated a continuous debate on the scope of each axiom and what should be the "correct" set of perceptual axioms to be used in the first place.

Though much less frequently, Ullman's work inspired others in yet a different way: that the completion process should be carried out with "early visual processing" (Ullman, 1976, p.1). Put differently, modeling curve completion should rely on early vision-inspired networks whose elements are capable of performing local computations. In particular, some subsequent research has tried to link the curve completion process to Hubel and Wiesel's (1977) findings that the early cortical area of mammals is constituted of orientations selective cells at all orientations for all retinal positions. Exploiting this idea is Williams and Jacobs' (1997b) stochastic completion field, which suggests that curve completion is the result of the most likely random walk in a 3D lattice of positions and orientations. Similar to Ullman (1976), Williams and Jacobs (1997a) implemented their model via a three-layer feed-forward neural network. Their first and second layers compute the probability that a random walk would go through a cell P in the 3D grid after starting from the first (source) and second (sink) inducers, respectively. Their third layer computes the product of the source and sink layers to represent the probability that a random walk between the source and the sink would pass through P. Unlike Ullman, Williams and Jacobs did not employ a nonmaxima suppression process (or other types of later interactivity; see section 3.4) and argued that higher probabilities in the third layer, which they identified with neurons in V2, correspond to the visually completed curve. Unfortunately, no follow-up tests were done to validate their theory, either perceptually or physiologically.

Although inspired by Ullman (1976), the approach suggested by Williams and Jacobs (1997b) was pioneering in its own right in the sense of not imposing any prior assumption of the desired shape of the completed curve in the image plane. However, their solution does possess other arguable features, such as an adhoc decay process, which makes longer paths exponentially less likely (Williams & Jacobs, 1997b). Since this particular mechanism is not easily verifiable psychophysically or physiologically, it may be desirable to obtain a preference for shorter paths in a more principled and parameter-free fashion.

Arguing against the axiomatic approach while advocating biologically inspired and low-level completion mechanisms, we have recently suggested pushing the abstraction of the primary visual cortex (including Williams and Jacobs's lattice of positions and orientations) one step further and move from the discrete domain to the continuous space $\mathbf{R}^2 \times \mathcal{S}^1$. As we elaborate below, this change of representation enables the use of rich tools from calculus, differential geometry, and the calculus of variation to investigate new completion principles that are nonaxiomatic (in the sense described above) but can reasonably be assumed to govern the behavior of biophysical (and, in particular, neural) systems.

2.2 The Tangent Bundle Approach. Hubel and Wiesel's (1977) findings of the structure and organization in the primary visual cortex are commonly captured by the so-called ice cube model, suggesting that V1 is continuously divided into full-range orientation hypercolumns (see Figure 3A), each associated with a different image (or retinal) position. Hence, an image contour is represented in V1 as an activation pattern of all those cells that correspond to the oriented tangents along the curve's arclength (see Figure 3B). The ice cube model has led researchers to abstract V1 via the unit tangent Bundle $T(I) \overset{\triangle}{=} \mathbf{R}^2 \times \mathcal{S}^1$ of the image plane $I = \mathbf{R}^2$ (Hoffman, 1989; Citti & Sarti, 2006; Ben-Shahar & Zucker, 2003, 2004; Petitot, 2003) and inspired few researchers to study curve completion directly in this space (Citti & Sarti, 2006; Ben-Yosef & Ben-Shahar, 2010b, 2012). In this letter, we further explore the latest theory and study and devise a biologically plausible network architecture that implements it on neural machinery, all motivated by a large number of reported findings in the neurophysiological literature.

Recall that an image curve $\alpha(t) = (x(t), y(t))$ is represented in V1 as an activation pattern of all those cells that correspond to the oriented tangents along the curve's arc length (see Figure 3B). In the limit, where V1 is abstracted as the unit tangent bundle $T(I)$, this image curve is represented by a "lifted" curve $\beta(t) = (x(t), y(t), \theta(t))$ where both position and tangent

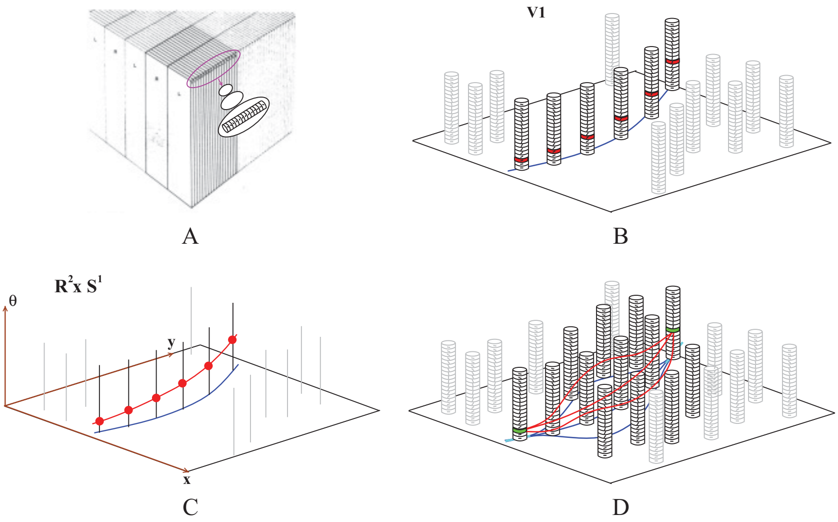

Figure 3: Curve completion in the tangent bundle. (A) The ice cube model (reproduced here with minor changes from Hubel & Wiesel, 1977) suggests that the primary visual cortex is continuously divided into hypercolumns, each covering the full range of orientation selectivity while being associated with a unique retinal position. (B) The association of each hypercolumn (abstracted here by one horizontal cylinder, which is rotated vertically) to its retinal (or image) position shows that a curve passing in the visual field is represented in V1 as an activation pattern of orientation-selective cells (in red) in the hypercolumns through which the curve passes. (C) Exploiting the continuous organization of both orientation and position tuning of V1 cells, we can further abstract orientation hypercolumns as infinitesimally thick fibers whose dense positioning in the image plane entails the 3D continuous unit tangent bundle $\mathbf{R}^2 \times \mathcal{S}^1$. In this abstraction, the activation pattern formed by the blue image curve becomes a 3D "lifted" continuous curve (plotted in red here). (D) Inspired by physiological findings (see section 2.3), we suggest that completed curves are represented in the early visual system similar to real image curves and therefore can be abstracted as lifted curves in the tangent bundle that link two end points (in green) representing the two image inducers (in cyan). Out of infinitely many possible tangent bundle curves (some of them are plotted here in red), the visual system selects one whose projection to the image plane (in blue) would be the perceived completed contour. We hypothesize that the selected tangent bundle curve is the one that corresponds to the activation pattern of least energy—the one that activates the minimal number of cells. According to our continuous abstraction, this implies the shortest admissible curve connecting the two green end points. The projection of this curve to the image plane would therefore define the "correct" shape from problem 1.

orientation are represented explicitly along the path (e.g., the red curve in Figure 3C). These insights suggest that if one cares about the biological plausibility of the curve completion process, perhaps the latter should be investigated directly in this space rather than in the image plane *I* itself. As Ben-Yosef and Ben-Shahar (2010b) noted, thinking of the curve completion problem in this way implies a construction of curves between boundary points in a three-dimensional space (each of which represents both the position and orientation of an inducer). Furthermore, the constructed curves in $T(I)$ cannot be arbitrary and must satisfy an admissibility constraint that forces their third coordinate to correspond to the tangential angle of their spatial progression in the image plane. More formally,

Definition 1. *Let $\alpha(t) = [x(t), y(t)]$ be a regular curve in I, and let $\beta(t) = [x(t), y(t), \theta(t)]$ be its corresponding curve in $T(I)$, which is created by lifting α to $\mathbf{R}^2 \times \mathcal{S}^1$. $\beta(t)$ is called admissible if and only if the coordinates of $\beta(t)$ satisfy*

$$tan\,\theta(t) = \frac{\dot{y}(t)}{\dot{x}(t)} \quad where \quad \dot{x}(t) \triangleq \frac{dx}{dt}, \;\; \dot{y}(t) \triangleq \frac{dy}{dt}. \tag{2.1}$$

The advantage of modeling the curve completion problem in $T(I)$ is rooted in the close relationship of this space to the first (though not necessarily the only) visual area that addresses this perceptual problem. This, in turn, also provides an opportunity to devise completion principles that are nonaxiomatic (in the sense of not assuming desired geometric properties in the image plane) but rather mechanistic, biological, or physical. Perhaps the simplest of the biological and physical completion principles is a "minimum energy consumption" or "least action" principle. According to this principle, the cortical tissue would attempt to link two boundary points (i.e., active cells) with the minimum number of additional active (i.e., energy-consuming) cells that give rise to the completed curve. In the abstract, this becomes a case of the shortest admissible path in $T(I)$ connecting two end points $[x_0, y_0, \theta_0]$ and $[x_1, y_1, \theta_1]$ (where admissibility must apply all along the curve, including its end points). Formally, this principle requires the minimization of

$$\mathcal{L}(\beta) = \int_{p_0}^{p_1} \sqrt{\dot{x}^2 + \dot{y}^2 + \hbar^2\dot{\theta}^2}dt \quad \text{subject to} \quad \tan\theta(t) = \frac{\dot{y}(t)}{\dot{x}(t)}, \tag{2.2}$$

where $p_0 = [x_0, y_0, \theta_0]$ and $p_1 = [x_1, y_1, \theta_1]$ represent the inducers, and $\hbar$ is a proportionality constant that balances the spatial and angular axes and makes them commensurable (Ben-Yosef & Ben-Shahar, 2010a). The curve β, which brings to a minimum the functional in equation 2.2 subject to admissibility, can be found by applying calculus of variations and nonlinear optimization methods. While most of the theory is discussed and derived elsewhere (Ben-Yosef & Ben-Shahar, 2010b, 2012), to make this letter as inclusive as possible appendix A provides a summary of the rigorous mathematical analysis of this model, its numerical solutions, and how its derived geometrical properties provide an excellent match to recent findings in the perceptual curve completion literature.

2.3 Neurophysiological Motivations. Since curve completion is a perceptual task performed by the visual system, it is important to examine more closely the link between rigorous shape models (whether biologically inspired or not) and the accumulated neurophysiological literature on the topic. Unlike most other work in the field, the findings we survey also provide key motivations for implementing the tangent bundle theory described in section 3.

In the past three decades, multiple studies have shown that participation of early visual neurons in the representation of curves is not limited to viewable curves only but extends to completed or illusory curves. Indeed, various tools to measure brain activity have shown that visual completion stimuli evoke activation in early visual areas V1 and V2, as well as in higher layers such as V4 and MT (V5). Studied on humans, monkeys, and cats, typical stimulation in these experiments constitutes illusory contours due to displaced gratings or modally completed curves induced by Kanizsa-type figures. As a baseline, in most experiments, the same subjects are also tested with luminance-defined shapes where real contours replace illusory curves from the critical conditions.

Among the first to show that illusory curve perception involves early visual cortical areas were von der Heydt, Peterhans, and Baumgartner (1984), who made single-cell neurophysiological recordings in the visual cortex of macaque monkeys during presentation of displaced gratings. They have found that approximately one-third of the orientation-selective cells in V2 fire when such illusory contours move across their receptive field (RF) and that many of these cells respond similarly to real and illusory contours. Follow-up studies reported a similar neural response in both areas V17 and V18 of a cat (Redies, Crook, & Creutzfeldt, 1986; Sheth, Sharma, Rao, & Sur, 1996) and area V1 in the macaque monkey (Grosof, Shapley, & Hawken, 1993). Sheth et al. (1996) also reported that cells that respond to both illusory and real curves are clustered in discrete columns and organized into maps of orientation preference.

One of the key questions that these studies triggered was how neurons in different visual layers interact during the completion process, and in particular, whether the completion process is bottom up or top down in terms of the visual cortical hierarchy. First insights into this issue were provided by Lee and Nguyen (2001), who examined the response of V1 and V2 neurons of the macaque to static illusory contours from Kanizsa-type figures. While they found that both V1 and V2 cells respond significantly to these illusory contours, they also reported a consistent temporal delay of V1 neural response to illusory curves (100 ms from stimulus onset, measured in superficial layers) compared to the response of real curves (approximately 45 ms) and the response of V2 cells to the same illusory curves (approximately 70 ms).

The development of functional brain imaging techniques in the early 1990s provided scientists opportunities to examine the curve completion process in humans too, where fMRI and PET have shown response to illusory contours not only in V1 and V2 (Hirsch et al., 1995; Ffytche & Zeki, 1996), but also in higher cortical areas V3A,V7,V4, MT, and V8 (Mendola, Dale, Fischi, Liu, & Tootell, 1999; Seghier et al., 2000; Kruggel, Herrmann, Wiggins, & von Cramon, 2001). Recently occlusion selective activity in V4 was reported using single-cell recordings as well (Bushnell, Harding, Yoshito, & Pasupathy, 2011). It was later suggested that the process includes feedback interactions from (typically large) RFs in the lateral occipital complex that encode whole illusory objects, down to V1 and V2 (Murray et al., 2002; Stanley & Rubin, 2003). At the same time, imaging studies suggest that the primary visual cortex maps both the inducers and the completed curves in a continuous fashion (Maertens, Pollman, Hanke, Mildner, & Moller, 2008).

While neurophsyiological findings may still be crude and indecisive, they do depict a particular computational chain that may guide the modeling of a neural computational process of curve completion in the visual cortex. In particular, it suggests that V1 responds to both the inducers and completed curves, though the former is stimulus driven while latter shows up later, after top-down feedback from higher visual areas. It also implies that V1 cells participating in the curve completion process exhibit some sort of continuous retinotopic organization linked to cells that respond to the inducers. All of these clues strongly motivate the configuration of our network computation for curve completion in the tangent bundle, as discussed next.

3 Neural Computation of the Shortest Admissible Path in the Tangent Bundle

The theory of curve completion via the shortest admissible path in the tangent bundle (Ben-Yosef & Ben-Shahar, 2010b, 2012) carries several important advantages. It is biologically inspired and based on a single physically and biologically plausible principle. It employs no shape axioms whatsoever and still makes shape predictions that closely match recent findings in the perception and psychophysics literature. Alas, the shape completion algorithm that emerges from it is hardly biologically plausible. Indeed, the theoretical analysis and derived numerical computations are based on the solution of differential equations (see appendix A) in a way that is difficult to reconcile with a distributed "network [of simple units] which performs simple local computations" (Ullman, 1976). While the latter may still be insufficient for demonstrating the full biological plausibility of an algorithm, it is at the very least a necessary condition for tightening the link between one's computational theory and the perceptual function it models. Naturally this link may be made even stronger if the neural computation as a whole supports the system-level organization found neurophysiologically and if its constituent functional units are inspired by those observed in the cortical tissue. In this section, we attempt to address both issues. We describe a network algorithm that exploits computational units observed in the visual cortex, reconstructs the shortest admissible path in the unit tangent bundle, and is configured in line with the organization and computational chain observed neurophysiologically (see section 2.3). We also discuss its computational complexity and show experimental results on various examples, including the temporal dynamics of the computation and its sensitivity (or lack thereof) to quantization and scale.

3.1 Setting the Infrastructure: Computational Building Blocks in the Visual Cortex. In the spirit of the above discussion and before we turn to describe the neural computation itself, we briefly discuss the main biologically plausible requirements and computational building blocks used in our network.

The ability to respond and represent oriented features will be used extensively, a basic capacity of early vision embodied in the orientation preference of simple cells (Hubel & Wiesel, 1977). We will assume that the linear operation of addition is fundamentally biologically plausible, as indicated by the summation properties of synaptic potential (Purves et al., 2004). Furthermore, we abstractly assume that various pieces of information and functional output of cells can be represented by their firing rate. For the proposed neural circuit, this is sort of a "biological requirement" that is putative in modeling, and although it is not necessarily a confirmed form of representation in the visual cortex, it is common in many neural computational models, including in the context of curve completion (Williams & Jacobs, 1997a).

Another elementary building block used in our network are lateral weighted links between orientation-selective units, a standard feature in virtually all computational neural networks whose biological justification is related to various factors such as myelination, amount of neurotransmitters released into synapses, and the number of connections formed between cells. Interestingly, in the visual systems, the last factor may be particularly influential as the number of long-range horizontal connections that connect cells in different hypercolumns (Gilbert & Wiesel, 1983; Rockland & Lund, 1982) is known to depend on the differences in their orientation preference (Bosking, Zhang, Schofield, & Fitzpatrick, 1997). But regardless of the the specific mechanism, another necessary condition for implementating our network is that lateral weighted links can represent distance in both orientation dimension (distance within hypercolumns) and spatial dimension (distance between hypercolumn). This is quite plausible since the orientation columnar organization dictates that both dimensions are represented as lateral cortical distance.

Finally, the last computational building block of interest in our case is the nonlinear MIN (resp. MAX) operator that returns the smallest (resp. largest) of its inputs. While physiologically it was studied less extensively, a number of studies have suggested that certain early visual neurons carry this computation (Lampl, Ferster, Poggio, & Riesenhuber, 2004; Finn & Ferster, 2007), while others have modeled it computationally (Yu, Gisse, & Poggio, 2002). Combined with other nonlinearities such as shunting inhibition (Torre & Poggio, 1978; Koch, Poggio, & Torre, 1983; Borg-Graham, Monier, & Frégnac, 1998; Frégnac, Monier, Chavane, Baudot, & Graham, 2003), previous studies have also modeled nonlinear gating operations where the neural circuit filters its data input unless it is equal or smaller (resp. larger) than its control signal (Ben-Yosef & Ben-Shahar, 2008). The availability of such a MIN and MIN-GATING operations is thus the last "biological requirement" in our computation.

3.2 A Distributed Network for Generating the Shortest Admissible Path in the Unit Tangent Bundle. Consider a discretized version of the unit tangent bundle that, as depicted in Figure 3B, more closely resembles the finite and discretized nature of the primary visual cortex. With proper neighborhood and edge structures (described below), we can think of our discretized space as a graph or network of interconnected cells (or vertices). As discussed in section 2.2 (and illustrated in Figure 3D), the curve completion problem in this graph receives as input two active cells that represent the boundary conditions and then seeks an admissible population of other cells that would constitute the completed curve between these end points. The minimum-energy (or least action) principle dictates that this sought-after population will be minimal in size (to consume the minimum possible energy), which in the abstract formulation corresponds to the shortest admissible path—the one that minimizes equation 2.2.

More formally, consider a three-dimensional grid graph $G = (V, E)$ in which each vertex $v = [x, y, \theta] \in G$ represents an orientation-selective neuron whose orientation preference is θ and spatial tuning is the coordinate $[x, y]$ in the visual field. Let $C_{[x,y|}$ denote the hypercolumn of all vertices with the same spatial tuning $[x, y]$, and each vertex v in hypercolumn C is connected by edges to all vertices u in all other hypercolumns with spatial distance r or less from C, as perhaps is dictated by the maximum distance that long-range horizontal connections can travel (see section 3.1). Our goal is to devise an algorithm that constructs the (discretized version of the) minimum length admissible path in this graph. (See Figure 4 for an illustration of the problem.)

Suppose for a moment that the weight of each edge $(u, v) \in E$ in our graph is "preprogrammed" to the length $w(u, v) = \mathcal{L}(u, v)$ of the shortest

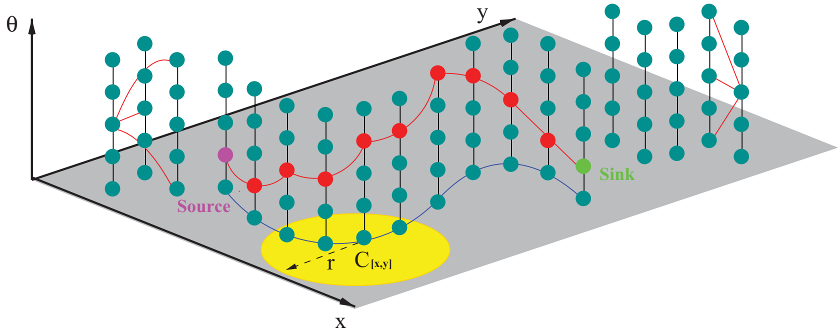

Figure 4: The discretized version of the unit tangent bundle can be formally considered as a three-dimensional grid graph $G = (V, E)$ in which each vertex $v = [x, y, \theta] \in V$ represents a neuron (or an entire orientation column) in V1 whose orientation preference is θ and spatial tuning is the coordinates $[x, y]$ in the visual field. Each edge $e \in E$ represents a horizontal connection between two adjacent neurons (several such edges are plotted here in red) whose synaptic weight reflects the length of shortest admissible segment between them. Each neuron is linked by edges only to its neighbors in a small given radius r (in yellow). The two inducers are assumed to be two "active" neurons due to the stimulation, and are represented here as source vertex (in magenta) and sink vertex (in green). With this graph structure in place, our task is to find a set of vertices (neurons) that reflect the shortest weighted path between the inducing source and sink (e.g., the set of vertices plotted here in red) in a parallel, cortically plausible fashion. The projection of this path to the image plane (in blue) would become the perceived completed contour. Note that in this illustrative sketch, admissibility may not necessarily hold.

admissible segment between points u and v in the unit tangent bundle.[2] With this graph structure in place, to find the shortest weighted path between two given active vertices (which represent the inducers), one can define them as source s and sink t and apply one of several known algorithms for single-source shortest paths in a weighted graph (e.g., the Dijkstra and the Bellman-Ford algorithms; see Cormen, Stein, Rivest, & Leiserson, 2001). However, these algorithms are better suited to a serial computation, which conflicts with our goal of designing a biologically plausible model. Instead, we would like to find a way to compute the shortest path between the source and the sink in a parallel fashion, where eventually all the

[2] This assumption is not very realistic since it requires a solution to our curve completion problem for each two nearby vertices. However, before we replace it with a practical and computable approximation, making this assumption can help in understanding the proposed network algorithm as a whole.

vertices (i.e., neurons) that are not part of the shortest admissible path are "shut down," while those that are part of this path remain "active."

To achieve the latter goal, we first observe that the relaxation process employed by the Bellman-Ford algorithm can be trivially parallelized into a biologically plausible computation.[3] Let s be a source in a graph G, and suppose that each vertex $v \in V$ maintains one value that represents the length l_{sv} of the shortest path from s to v (and in the biological network may be represented by the firing rate of the corresponding cell, as discussed in section 3.1). Clearly, when the algorithm begins, l_{ss} is initialized to 0, while l_{sv} is set to infinity for all $v \neq s$.[4] Then, at each iteration, the following relaxation process is performed at all vertices independently and concurrently:

$$\text{if}(l_{sv} > l_{su} + w(u, v)) \quad \text{then} \quad l_{sv} \leftarrow l_{su} + w(u, v),$$

or in a more biologically plausible manner (recall the MIN operator from section 3.1)

$$l_{sv} \leftarrow \text{Min}\{l_{sv}, l_{su} + w(u, v)\}, \tag{3.1}$$

where u is a neighbor of v in G. Note that the neighbors can be scanned in arbitrary order and that the distributed process requires no network synchronization whatsoever.

By correctness and convergence of the Bellman-Ford algorithm (Cormen et al., 2001), it takes at most $|V|$ iterations (concurrently at each node) until all l_{sv} converge to the length $\mathcal{L}(s, v)$ of the shortest path from s to v (for all v in the graph). Obviously, at this convergence state, the value l_{st} at the sink (i.e., when $u = t$) would represent the length of the shortest admissible path from s to the sink t, and hence the length of the completed curve. However, this is still short of telling us which vertices participate in this path. To get this information, it would be desirable to have all vertices that belong to the minimal admissible path flagged in a particular way, after which all other vertices are shut down. To do so, we employ the same parallel Bellman-Ford procedure on a duplicate graph $G' = (V', E')$, where this time, we define vertex t as the source. Running concurrently with the execution on

[3] There are several distributed solutions for the single-source shortest-path problem, and for the Bellman-Ford algorithm in particular. Our suggested solution is a trivial parallelization of the serial Bellman-Ford algorithm that is fitted to the computational infrastructure of V1. At the same time, it is worth noting that the algorithm itself (Bellman-Ford) is an instance of dynamic programming (Cormen et al., 2001), which has long being considered a corresponding discrete version of the calculus of variations (Dreyfus, 1960; Bellman, 1954), which we use in our formal solution (Ben-Yosef & Ben-Shahar, 2010b, 2012).

[4] Biologically, "infinite length" may be represented by the length of a very long curve that is unlikely to be perceived by any given pair of inducers in the visual field.

graph G, after at most $|V|$ iterations, when both runs have converged, each two corresponding vertices $v \in V$ and $v' \in V'$ hold the length of the shortest admissible tangent bundle paths from s to v (i.e., l_{sv}) and from t' to v' (i.e., l'_{tv}), respectively. Obviously, summing these two values for each corresponding pair of vertices would provide the length of the shortest path between s and t that passes through v, an operation that can be done by a third neural layer (or graph) $G'' = (V'', E'')$ of the same structure, as soon as the first stage is completed. Formally, if the third layer computes

$$l''_{svt} \leftarrow l_{sv} + l'_{tv} \quad \forall v \in V'', \tag{3.2}$$

then by construction, the length SP of the shortest admissible path between s and t satisfies

$$SP = \min_{v}\{\, l''_{svt} \mid v \in V''\}, \tag{3.3}$$

and only those vertices in the third layer for which $l''_{svt} = SP$ are part of this optimal path. Hence, we finally employ a nonlinear nonminima suppression over l''_{svt} (see section 3.4) and pass the result to a new and final layer, where the only active cells are those that hold the shortest path. The activation pattern in the fourth layer represents the completed curve in the discretized version of the unit tangent bundle. The entire four-layer process is illustrated and exemplified in Figure 5.

Recall now that the network algorithm proposed above assumed that the weights of the edges in the graphs G and G' were preset to the length of the shortest admissible paths between the vertices they link. This therefore assumes that local solutions are available to the same problem that the entire algorithm aspires to solve globally, which amounts to a chicken-and-egg sort of problem. In practice, however, since any two neighboring vertices are very close to each other in $\mathbf{R}^2 \times \mathcal{S}^1$, a simple approximation is possible that eliminates this difficulty and facilitates the practical network solution of the curve completion problem.

Instead of using edge weights based on the true length of the shortest admissible path segments, we define the weight of the edge between the neighboring vertices $v_0 = [x_0, y_0, \theta_0]$ and $v_1 = [x_1, y_1, \theta_1]$ to be the sum of two terms: their distance in $T(I)$ (see equation A.2 in appendix A),

$$D(v_0, v_1) = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + \hbar^2(\theta_1 - \theta_0)^2}\ , \tag{3.4}$$

and a penalty term that penalizes deviations from admissibility. In discrete form, the admissibility constraint from equation 2.1 becomes

$$\frac{\sin\hat{\theta}}{\cos\hat{\theta}} = \frac{\Delta y}{\Delta x},$$

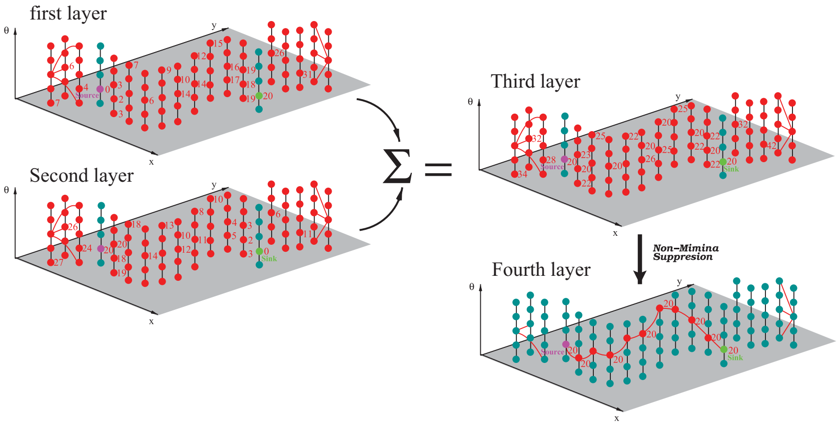

Figure 5: A biologically plausible network model for computing the shortest path in the tangent bundle. To make a parallel computation, our network consists of four layers, each of which takes the structure of the graph in Figure 4. At the first and second layers, we compute the shortest weighted path between each vertex in the layer to the source and the sink, respectively. The length of the shortest path to each neuron is computed via the iterative relaxation process described by equation 3.1, and is coded within the individual neuron (e.g., by its activity level). In the toy example shown here, red depicts activity (as opposed to cyan for baseline), and the numbers plotted next to some of the neurons depict its level (i.e., the length of their shortest path to the source or to the sink, assuming a sufficient number of iterations). The third layer carries a summation of the first and second layers, such that the activity of a vertex v in this layer reflects the length of the shortest path between the source and the sink that passes through v. Note that the minimum activity in the third layer (20 in this example) represents the length of the true shortest path between the source and the sink, and cells that exhibit this level of activity are part of that path. Hence, we now shut down all the vertices (neurons) in the third layer with activity larger than the minimal, an operation that can be done by a fourth layer of neurons that performs nonminima suppression over the results of the third. Active cells in the fourth layer constitute the shortest path in the graph between the inducing source and sink and their projection to the image (i.e., retinal) plane representing the perceived completed curve.

or

$$\Delta x \cdot \sin\hat{\theta} - \Delta y \cdot \cos\hat{\theta} = 0,$$

where $\Delta x = x_1 - x_0$, $\Delta y = y_1 - y_0$, and $\hat{\theta} = \frac{\theta_1+\theta_0}{2}$ is the average orientation of the connected vertices. Thus, a proper penalty term that grows with increasing deviation from admissibility would be

$$T(v_0, v_1) = |\Delta x \cdot \sin\hat{\theta} - \Delta y \cdot \cos\hat{\theta}|. \tag{3.5}$$

when both terms are combined (tangent bundle distance and admissibility penalty), the weight of an edge from v to u is thus defined as

$$\begin{aligned} w(v_0, v_1) &= D(v_0, v_1) + \eta T(v_0, v_1) \\ &= \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + \hbar^2(\theta_1 - \theta_0)^2} \\ &\quad + \eta |\Delta x \cdot \sin\hat{\theta} - \Delta y \cdot \cos\hat{\theta}|, \end{aligned} \tag{3.6}$$

where η is penalty weight or a regularization factor reminiscent of a Lagrange multiplier (see appendix B). As long as the distance between the two neighboring nodes v_0 and v_1 is small, equation 3.6 approximates well the shortest admissible path between them. At the same time, unlike the true shortest admissible path, equation 3.6 can be computed in a straightforward fashion, hence resolving our chicken-and-egg problem and facilitating the entire network computation. Running the four-layer network algorithm using these weights would prioritize admissible shortest paths, and hence would converge to the desired completion result up to the error introduced by the discretization.

To summarize the proposed solution, curve completion based on minimal length in the tangent bundle can be computed by a four-layer network in which each of the first two computes a field induced by one of the inducers, the third layer performs a simple summation over the first two, and a fourth layer performs nonlinear, nonminima suppression (or minima selection) over the results of the third. In what follows, we argue that this structure is well accommodated and implemented by the computational machinery found in the early parts of the visual cortex in a manner that supports the computational chain implied by neurophysiological findings (see section 2.3). Before we turn to that discussion, however, we first examine the time complexity of the network and the type of results it provides.

3.3 Time Complexity and Experimental Results. As discussed in section 3.2, the theoretical time complexity of our network algorithm is $O(|V|)$, proportional to the number of vertices (or neurons) in the graph (or network). A straightforward estimation of this number indicates that this asymptotic complexity is problematic in terms of biological plausibility. Indeed, the cortical surface area of an adult primate V1 is often approximated at 1300 mm^2 (Purves & LaMantia, 1990), with each hypercolumn estimated at 1 mm^2 in area (Hubel & Wiesel, 1977). Assuming that the estimated 1300 hypercolumns are organized in a square grid (say, approximately 36×36 in size), and adopting the reported V1 angular resolution of 10° (Hubel & Wiesel, 1977), we thus obtain a network size of $|V| = 1300 \times 36 = 46,800$ vertices. With a single iteration between two neighboring vertices clearly bounded from below by the cycle time of a neural action potential (about 3–5 ms for spike duration plus refractory period; Purves et al., 2004), we therefore obtain a theoretical lower bound on the convergence time on the order of minutes, far greater than the reported physiological findings (approximately 120 ms; see Guttman & Kellman, 2004; Ringach & Shapley, 1996; and section 2.3). Clearly, by this measure, the biological plausibility of the proposed algorithm is questionable.

But is the theoretical complexity also the practical one? The upper bound of $|V|$ iterations is effectively necessary for the most general curves (or graph paths) with an arbitrary number of twists, turns, and inflection points in their image plane projection. However, it is unreasonable to assume (and it was never reported) that perceptually completed curves consist of more than one inflection point. Hence, in practice, the number of network iterations until observed convergence may be expected to be significantly smaller than the theoretical limit. In fact, using an empirical validation, we have found that in practice, the necessary number of iterations is bounded by a constant. To do so, we have implemented our model according to the above estimated size of the network in V1. We have built a 40×40 array of hypercolumns, each consisting of 36 vertices to cover all orientations at 10° resolution. To match the range of the cortical horizontal connections, we set the radius of neighborhood to $r = 4$ hypercolumns (i.e., horizontal connections extending up to 4 mm parallel to the cortical surface; Gilbert & Wiesel, 1983; Rockland & Lund, 1982). Although $\hbar$ should be calibrated perceptually by psychophysical studies (a behavioral study in its own right, which is outside the scope of this computational work; see also Ben-Yosef & Ben-Shahar, 2012), and then η could be calibrated according to $\hbar$ to match the analytical results (see appendix B), at this stage, we have employed only pilot tests to set $\hbar = 13$ and $\eta = 3$, which seem to match well our perceptual completions over the given scale of a 40×40 size grid. We do emphasize that both of these $\hbar$ and η values were found to be stable over a large set of completion examples and well matched the analytical curves. Finally, to make sure that no synchronization limitation exists, we serialized each iteration by completely randomizing the order of the node relaxation computation (see equation 3.1) across the network.

Several experimental results of our four-layer algorithm on various inducing (source and sink) nodes are shown in Figure 6, where they are also compared to the analytical results. Figure 7 shows completion results by the network in several natural and synthetic completion scenarios. The inducers in these computations were measured and extracted manually and then were fed as two initial active neurons to the network. The network final shape (as in Figure 6) was then projected on top of the original image using the proper discretization. Figure 8 shows the temporal dynamics of the computed solution in one case as it evolves during the relaxation process of equation 3.1 and the operation of the four network layers. The results in all cases (and many others not shown) have converged to completed curves with no or at most one inflection point. Using the suggested spatial discretization and orientation quantization levels ($40 \times 40 \times 36$), the

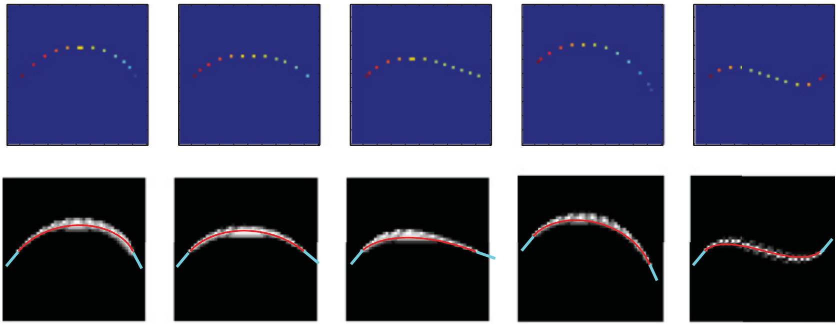

Figure 6: Experimental results of our network model for different settings of pairs of inducers (plotted here in the bottom row in cyan). The network was implemented as a 40 × 40 grid of columns, with each column consisting of 36 cells differing by orientations, starting at 0° and ending at 350°. Using simple pilot calibration, we set $\hbar = 13$ and $\eta = 3$, and the radius neighborhood was set to $r = 4$ units. (Top) The shortest weighted path as generated in the final fourth layer of our network computation. Colors represent the height (i.e., orientation). (Bottom) The shortest weighted path projected to the retinal field. To better capture the intuitive percept, we did not use a strict nonminima suppression here but suppressed all neurons in which $l''_{svt} > SP + \epsilon$, with $\epsilon = 0.1$. Compare this result to the red curves that show the output of the numerical procedure based on variational calculus from Ben-Yosef and Ben-Shahar (2012).

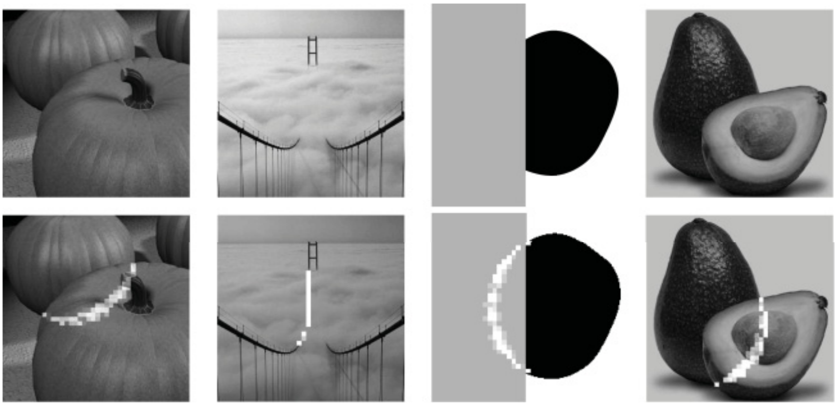

Figure 7: Experimental results of our neural computation on different completion scenarios. All network parameters set as in Figure 6. Inducers were chosen manually by measuring the network position and orientation at the points of occlusion. Active nodes in the fourth layer (up to ϵ, as in the bottom row of Figure 6) were projected to the image plane and drawn here as high-intensity pixels on top of the original image.

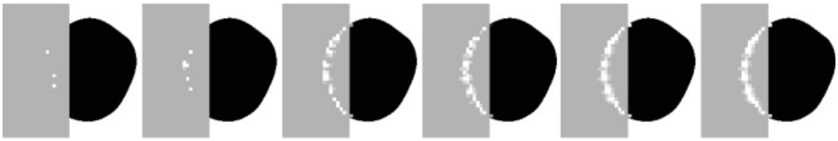

Figure 8: The temporal evolution of the network solution presented as snapshot through the first six computational iterations, where the inducer orientations are 20° and 170°. All parameters (network size, quantization, $\hbar$, η, and r) were set as in previous examples.

computation never required more than 10 iterations to converge, regardless of the image plane distance between the two inducers or their relative orientation. Assuming again that one iteration is an order of magnitude of the 5 ms spike cycle time (Purves et al., 2004), practical convergence in genuine parallel implementation now indeed fits the reported time of perceptual curve completion in V1 (Guttman & Kellman, 2004; Ringach & Shapley, 1996) as reported physiologically (see section 2.3).

Recall that all our experimental results are obtained by a network whose parameters are inspired and inferred from physiological data. In particular, network size was set around 40×40 and orientation quantization around 10°. Still, it may be important to understand how sensitive the performance may be to variations in these quantization parameters. Hence, we examined how the degree of quantization might affect both the shape of the resultant completed curves and the number of iterations to convergence. To do so, we selected a fixed pair of inducers and executed several runs of the network while keeping all parameters fixed except varying the quantization level of the spatial and angular axes to $(dQ \cdot 40) \times (dQ \cdot 40) \times (dQ \cdot 36)$ nodes, where $dQ \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$. Hence, the same network algorithm was tested on the same initial conditions with total network size varying from 7200 to 900,000 nodes. The results in Figure 9 show that the shape of the completed curve remains remarkably stable (up to the resolution of the result, of course), and in particular, that no systematic qualitative changes (such as excessive flattening or swelling) are introduced with increasing or decreasing quantization. The figure also shows the number of effective iterations it took the network to converge. Clearly, the numbers are far from growing as $O(dQ^3)$, as the theoretical $O(|V|)$ limit might suggest. Rather, it appears to grow even more slowly than $O(dQ)$. Since the actual network has this parameter fixed and since the number of iterations remains small even for very large quantization levels, we reaffirm our informal observation that effectively the perceptual result can be obtained in a constant amount of time.

3.4 Representation as a Cortical Circuit. Could the network computation presented above be embedded as a visual cortical circuit? And would

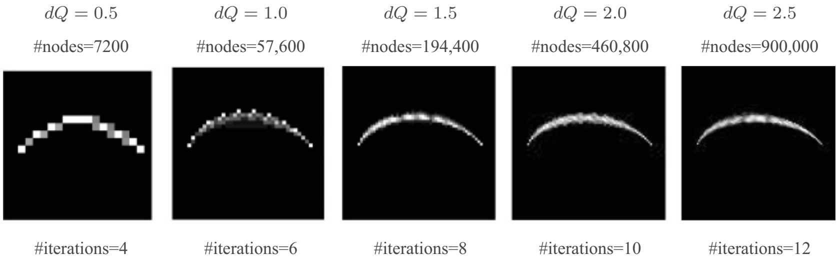

Figure 9: Experimental results of our network algorithm for different quantization levels and network sizes (left inducer 250°, right inducer 130°). Note how the effective number of iterations to convergence grows significantly more slowly than the size of the network (i.e., number of nodes) and by all practical measures can be considered constant.

it exhibit the functional features observed neurophysiologically (see section 2.3)? In this section, we answer both questions positively.

Observe first that the nodes that make the first two layers in our computational network can be thought of as orientation-selective cells in V1 that respond to completed (and possibly also to real) contour stimuli (see section 2.3). The edges between these nodes can be naturally thought of as horizontal connections, whose weights (see equation 3.6) are implemented by any of the mechanisms briefly discussed in section 3.1.

The representation of the third and fourth layers is slightly more complicated, but we argue that it is naturally settled with the two basic nonlinearities that were found in early visual cortex recordings and discussed in section 3.1. We first argue that the third layer consists of V2 neurons that perform the summation from equation 3.2 on the output of cells in layers 1 and 2. The necessary feedforward connections and the participation of V2 in curve completion are both discussed in section 2.3. Since (according to the computational model) only selected V2 neurons should propagate information to the final layer, we propose that a nonminima suppression process via interareal connections is done at this stage. Through feedforward connections, neurons in the third layer interact with neurons in higher cortical areas such as V4 or V5 (see Figure 10). These high-level neurons receive input from all neurons in the third layer, which together cover a large part of the visual field and, in particular, from the region in the visual field where the entire completion is formed (cf. the large receptive field and large spatial scale of high-level neurons and the large area in the visual field covered by their feedback terminals as reported by Angelucci et al., 2002). We propose that these high-level neurons in area V4 or V5 carry a MIN-like computation (Gawne & Martin, 2002) and compute a global minimum value of the network described in equation 3.3.

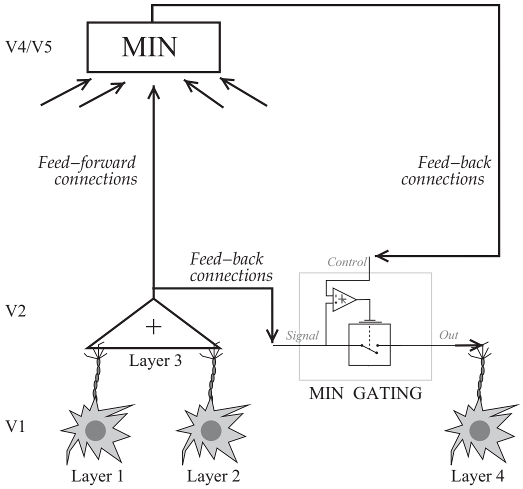

Figure 10: Neurophysiological circuit of our model. The first two layers consist of V1 neurons, which respond to completed (and possibly also to real) contours and whose operation is described in equation 3.1. The summation of the output of the first two networks is carried out in V2, and the result then propagates onward to a global minimum cell that could be located in higher visual areas such as V4 or V5. The global MIN is then projected downward as feedback to V1 and used to modulate the response of all the neurons that participate in the completion process via the nonlinear MIN-GATING operation. These modulated activities represent the final perceptual outcome.

This global MIN value computed by V4/V5 cells is then propagated back through feedback connections to low-level areas, where the selection process continues. Recall that at this stage, we care only about those layer 3 cells whose output is the minimal and equal to the global MIN value just computed. But this selection can be made by the MIN-GATING operation described in section 3.1. Hence, in addition to sending feedforward signals to high-level areas to compute the minimum value across the network (which represents the minimum-length path in the tangent bundle), the third-layer neurons also send their output via feedback connections to a MIN-GATING circuit that permits the propagation of only those signals that equal the global MIN value (see Figure 10). To do this, the MIN-GATING circuit receives the global MIN value as control through the feedback connections from V4/V5 mentioned above.

While the process above is done repeatedly from the onset of the completion process, only after a sufficient number of iterations (see section 3.3) does the global MIN converge to its final form, and active neurons in the fourth layer would be those that satisfy $w_{svt} = SP$.

To wrap up this section, it may be important to examine the proposed network computation once more in light of the fundamental principle of least action that governs its solution. After all, it is evident that the activation of layers 1 to 3 (in which almost all cells become active at some level) is energy demanding rather than energy conserving. In other words, it seems that much energy (i.e., activation) is required to obtain the eventual minimal set that represents the final answer. Indeed, we reemphasize that our theory does not claim that the computation itself consumes minimal energy. Such a theory would inevitably produce only trivial results (amounting to no computation at all). However, it asserts that the computational process flows into a minimal energetic steady state, which is then maintained as long as needed. In this context, it is also interesting to reflect on the link between our work and early ideas of the Gestalt movement, in particular the notion of tendency toward minimum energy. Gestaltists like Köhler (1920) hypothesized that the principles of perception, like those of mechanics, are the outcome of a "development in the direction of minimum energy" (p. 52), or what was encapsulated in the German word *Pragnänz*.

4 Discussion

Curve completion, despite its fundamental role in visual perception, has largely been studied in an intradisciplinary fashion. While psychologists and neurophysiologists have accumulated a large set of experimental facts in their own fields, computational shape completion theories have tended to ignore the majority of them. Inspired by Newell (1973) and Stevens (2000) (see also Palmer, 1999), one important goal in our work is to begin linking these ends more tightly so the process of curve completion is not merely modeled but (perhaps for the first time) becomes part of a coherent theory.

Our theory is founded on the similar representation of real and completed/illusory visual curves in the visual cortex. This similarity, which now has both perceptual and physiological support, leads to the idea that a completed curve is formed by a set of active orientation-selective neurons in the primary visual cortex, which are organized in a topographically continuous fashion and linked by horizontal connections. With the abstraction of V1 as the unit tangent bundle $\mathbf{R}^2 \times \mathcal{S}^1$, this set of active neurons becomes a continuous curve in this space that can be reconstructed according to certain principles that may govern certain activities in the visual cortex. In particular, here we study the shape of the completed curve that is formed by the principle of least action or minimum energy consumption. We further propose that this principle is implemented in the visual cortex by locally connected parallel networks whose operation fits critical physiological findings. We do note that our model, like most other computational models, does not attempt to cover every bit of neural behavior from system level to low-level biophysics of individual cells. It does aspire, however, to take a significant step to get there, taking the minimum activation in V1 as a guiding principle, parallel neural networks as the underlying computational configuration, and observed neural behavior as the sought-after function.

Indeed, some biophysical issues remain vague, and their further exploration might get us even closer to a coherent computational theory. Perhaps the most interesting one would be the use of memory since a dynamic programming technique such as Bellman-Ford requires a dynamic table (i.e., memory) to keep and maintain the solutions to subproblems. In our case, this "memory" is the stored lengths of shortest paths between neurons to the source and sink cells (denoted by l_{sv}, l'_{tv}). The representation of these lengths in individual cells is challenging and may be difficult to explain with current physiological evidence. Another intriguing issue is the one of the scale of computation, which in our model leaps from individual V1 and V2 RFs to global computation of minimum at V4/V5. One could conceive a more gradual process in a form of a pyramid-like computation across scales, where spatial evidence is collected locally at each scale up to the eventual global decision (in our case, for computing the shortest path in a weighted graph). While such a model may be possible, it raises other biophysical questions whose rigorous analysis and computational and biological implementation are left here for future research.

Although the compatibility with available physiological findings has been discussed throughout this letter, and the agreement with psychophysical and perceptual findings has already been described elsewhere (see appendix A for a summary of observations from Ben-Yosef & Ben-Shahar, 2012), it is important to examine additional implications that this theory entails, which could inform future work on the problem. First, it is easily acknowledged that the interesting aspects of curve completion, and the differences between the various shape models, are revealed for curvilinear completions. Still, to our best knowledge, all of the related electrophysiological studies have focused on linear completions—namely, on (modal or amodal) straight-line completions due to co-aligned inducers (von der Heydt et al., 1984; Redies et al., 1986; Grosof et al., 1993; Sheth et al., 1996; Lee & Nguyen, 2001). On the other hand, not unlike findings using psychophysics (Guttman & Kellman, 2004) and fMRI (Maertens et al., 2008), it is expected that single neurons in V1 or V2 would also respond to curvilinear illusory shapes. We therefore call on experimental electrophysiologists to expand the scope of their experiments accordingly, where one way to do so is to extend experiments such as Lee and Nguyen (2001) to include curvilinear Kanizsa shapes also. The results could then be used to prioritize certain shape models over others and for the first time match perceptual data with objective neural response at the single-cell level.

One result of our tangent bundle model is that completions with longer tangent bundle length yield (after convergence) larger activated areas in V1 since such completions are generated by more V1 neurons. One could therefore hypothesize that longer tangent bundle completions would exhibit larger activated regions in fMRI z-maps. Clearly the length of the completed curve (and hence the predicted size of the activated region) would increase if we keep their orientation but increase their distance. However, less trivial manipulations of the stimuli could entail similar predictions, for example, by fixing the inducers' location but increasing their angular difference. Verifying or refuting such predictions may require better spatial imaging resolution than is currently available but would be testable once this technology improves.

Our experimental validation shows that no completed curve consists of more than one inflection point, and although no rigorous proof has yet been found, we conjecture that this property is intrinsic to our tangent bundle completion theory. While to our best knowledge these experimental observation and theoretical conjecture have never been refuted perceptually or psychophysically, it would be interesting to explore this issue more closely. Similarly, although we did not observe any case where more than one perceived completed curve can be obtained for a same inducer-pair input, no proof has been proposed to the contrary. Hence, perceptual and psychophysical evidence of ambiguous or multiple completions would be very constructive in the context of validating curve completion theories.

The proposed neural computation presented in this letter associates a degree of biological plausibility to curve completion via least action in the tangent bundle and further supports it as a viable explanation for the corresponding perceptual process. Still, from a computational point of view, there could be other physical or biological completion principles that may be explored using the tangent bundle framework. One such appealing principle is completion based on the "least bending energy," and although it has been previously explored in the image plane (Horn, 1983; Mumford, 1994), it was never considered directly in the visual cortex or its unit tangent bundle abstraction. Although attempting the latter is challenging in terms of both the mathematics involved and the biological justification, preliminary analysis shows that it has some unique advantages over all existing models (including the one proposed here) in that it predicts an influence of inducer curvature. Since such effect was observed psychophysically (Takeichi, 1995; Singh & Fulvio, 2005), it provides a strong incentive to explore this direction as future computational work. The implementation of such a principle as a neural computation may require an extension of the tangent bundle network to a four-dimensional grid graph in which each vertex $v = [x, y, \theta, \kappa]$ represents an orientation-selective neuron that is also tuned to curvature (Dobbins, Zucker, & Cynader, 1987, 1989; Versavel, Orban, & Lagae, 1990; Ben-Shahar & Zucker, 2004). Interestingly, such networks have been utilized for edge enhancement in computer vision (August & Zucker, 2003), and it is intriguing to examine this direction for curve completion too.

It is worth mentioning that the involvement of high-level cortical areas in the completion process (see section 2.3) suggests that this interaction may not only be reactive but proactive as well. Indeed, it is not inconceivable that shape priors or other biases, although never tested psychophysically, have some effect on the result of the curve completion process. Such prior knowledge may be represented in higher visual areas such as lateral occipital complex in humans (Stanley & Rubin, 2003) or the inferior temporal cortex in monkeys (Sáry et al., 2008) and propagated down to lower-level computational circuits to affect the generated (perceptual) shape.

Finally, it is important to reiterate that our suggested network algorithm is designed to solve the curve completion problem between two inducers whose correspondence has already been established by a grouping process (recall section 1 and Figure 2A). In this sense, the presented network is closer in spirit to Ullman (1976), Kimia et al. (2003), and others, and it is different from previously suggested models in which both shape and grouping problems are attempted by a single process (Grossberg & Mingolla, 1985; Williams & Jacobs, 1997b; Citti & Sarti, 2006). Unfortunately, these few attempts can easily hallucinate false completions (in the perceptual sense) between inducer pairs that should not have been grouped in the first place, demonstrating how the relationship and mutual dependency between the grouping and shape problems in curve completion are still open questions. As already noted, the triggering conditions by which two inducers are chosen and grouped for a completion operation may also involve more higher-level mechanisms (Ullman, 1976), as may be suggested by contemporary theories and experimental findings (Mendola et al., 1999; Stanley & Rubin, 2003; Tse, 1999). While solving the grouping problem in the tangent bundle framework is clearly an interesting future direction (perhaps by more complex interactions between low- and high-level areas), doing so correctly and in a biologically plausible manner is by no means a trivial task and is important future work.

4.1 Beyond Vision: Cortical Completion Principles in Auditory Perception. To conclude our letter, it is useful to consider once more the underlying "least action" principle of our theory and reflect on whether its nonvisual nature could be employed in other contexts, for example, for the comparable filling-in phenomenon in the auditory system (Warren, 1970; Bergman, 1990; Sugita, 1997; Miller, Dibble, & Hauser, 2001; Petkov, O'Connor, & Sutter, 2007). Indeed, similar to the visual system, the auditory system often encounters only fragmented and incomplete information of conceptual "wholes," as would be the case when sounds from one animal interrupt vocal communication between other animals, when noise breaks in during a musical piece, or when poor communication fragments radio broadcasts. In auditory research literature, these interruptions also are called occlusions (Bergman, 1990; Kluender & Jenison, 1992), and the ability of the auditory system to fill in these gaps and organize the sensory information to keep a stable representation of the world has become known as auditory restoration (Warren, 1970), auditory induction (Petkov et al., 2007), and auditory amodal completion (Miller et al., 2001). Evidence of this perceptual phenomenon has been reported behaviorally for humans (Warren, 1970; Bergman, 1990; Kluender & Jenison, 1992) and monkeys (Miller et al., 2001), and physiologically for cats (Sugita, 1997) and monkeys (Petkov et al., 2007).

Experimentally, auditory completion is typically studied by inserting into or replacing parts of an auditory signal (e.g., a tone, phoneme, frequency glide, natural vocal signal) with acoustic noise and then testing the subject's response either behaviorally or neurophysiologically. In this way, it has been shown that the auditory system organizes information according to a set of basic principles equivalent to the Gestalt principles of visual perceptual organization (Bergman, 1990). Particularly, it has been shown that continuity plays a role in auditory completion as much as good continuation affects visual completion. For example, a classical demonstration of this sort due to Carlyon (2004) shows that auditory occlusion in the middle of the word meet (e.g., due to a loud clap) does not make subjects interpret the utterance as two words (*me* + *eat*) but rather perceive it as the original word (*meet*) extended behind the occlusion.

Could shape theories for visual curve completion like the one discussed in this letter be linked to auditory completion more tightly? An interesting experimental paradigm for testing this possibility would be the perception of occluded pure frequency modulated (FM) signals, also known as frequency trajectories (Kluender & Jenison, 1992, as illustrated in Figures 11B and 11C). Although these idealized stimulations are quite rare in natural vocalization, it is tempting to consider the perceived frequency shape of the occluded sound, as discussed for visual curves in this and other work in the visual curve completion literature. In particular, since the primary auditory cortex is known to tonotopically represent the auditory field and consist of frequency and frequency change selective neurons (Purves et al., 2004), we propose examining whether it is possible to describe this missing sound via a similar principle of minimum energy consumption, that is, through minimal patterns of frequency-selective active cells. While such directions require much experimental effort, they may provide a unified approach to sensory processing not only in vision but in other modalities as well.

Appendix A: A Short Theoretical Account of Curve Completion as Minimum Length in the Tangent Bundle

A.1 Theoretical and Numerical Analysis. Following the motivation, arguments, insights, and notations from section 2.2, here we formally define and analyze the curve completion problem via minimum length in the tangent bundle (MLTB).

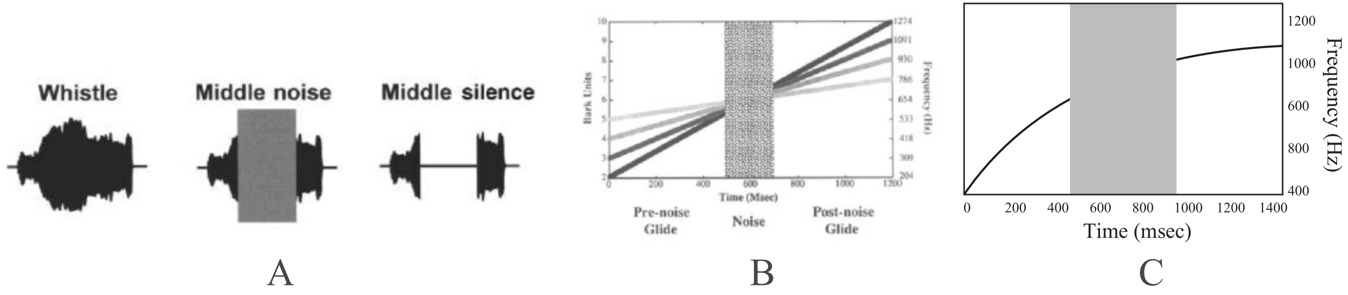

Figure 11: Phenomenology of auditory completion. (A) Reproduced from Miller et al. (2001), the middle part of the amplitude waveform of a monkey's whistle is occluded by a noise interval or replaced by a silent gap. It has been shown that the monkeys respond to the middle noise as if the full whistle was perceived, while they do not respond the same way to the middle silence. (B) Auditory amodal completion of frequency glides in humans can be used to test auditory grouping (i.e., which pair of the prenoise and postnoise slopes is perceived as the most continuous; reproduced from Kluender & Jenison, 1992). (C) Shape completion in the perception of occluded auditory signals. In this proposed experiment, would the completed sound in this occluded nonlinear frequency glider have a particular shape? Could this shape be described and predicted by a least-action principle in the auditory primary cortex? We believe that this and similar interesting questions could become natural future directions of our theory.

Let p_0 and p_1 be two given end points in $T(I)$ that represent two oriented inducers in the image plane I. If one were to seek the shortest admissible path in $T(I)$ between these two given end points, a proper objective function that employs the Euclidean metric would be

$$\mathcal{L} = \int_{p_0}^{p_1} \sqrt{\dot{\beta}(t)^2}dt = \int_{p_0}^{p_1} \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2 + \dot{\theta}(t)^2}dt, \tag{A.1}$$

subject to the admissibility constraint. However, a natural question that arises relates to the units and relative scale of the different dimensions in this space. Indeed, while x and y are measured in meters (or other length units), θ is measured in radians. Furthermore, the hypercolumnar organization of V1 suggests that the "cost" (or cortical distance) of moving one orientation unit is not necessarily similar to moving one spatial unit. Hence, to balance dimensions in the arc length integral and facilitate relative scale between the spatial and angular coordinates, a proportionality constant $\hbar$ in units of $\frac{meters}{radians}$ should be incorporated in equation A.1 (in a manner reminiscent of many physical proportionality constants such as the reduced Planck constant, which proportions the energy of a photon and the angular frequency of its associated electromagnetic wave). We thus generalize the distance measure between points in $T(I)$ and formulate our curve completion problem as follows:

Problem 2. Given two end points $p_0 = [x_0, y_0, \theta_0]$ and $p_1 = [x_1, y_1, \theta_1]$ in $T(I)$, find the curve $\beta(t) = [x(t), y(t), \theta(t)]$ that minimizes the functional

$$\mathcal{L}(\beta) = \int_{t_0}^{t_1} \sqrt{\dot{x}^2 + \dot{y}^2 + \hbar^2\dot{\theta}^2}dt \tag{A.2}$$

while satisfying the boundary conditions $\beta(t_0) = p_0$ and $\beta(t_1) = p_1$ and the admissibility constraint from equation 2.1.

Let $\alpha(s) = [x(s), y(s)]$ be an image curve given in arc length parametrization, whose corresponding lifted curve in $T(I)$ is

$$\beta(s) = [x(s), y(s), \theta(s)]. \tag{A.3}$$

Representing all admissible curves in $T(I)$ in this form, the functional $\mathcal{L}$ from equation A.2 becomes

$$\mathcal{L}(\beta) = \int_{0}^{l} \sqrt{\dot{x}(s)^2 + \dot{y}(s)^2 + \hbar^2\dot{\theta}(s)^2}ds\ , \tag{A.4}$$

where l is the total length of $\alpha(s)$ and the admissibility constraint, equation 2.1, can be written as

$$\cos\theta(s) = \dot{x}(s),$$
$$\sin\theta(s) = \dot{y}(s). \tag{A.5}$$

Given these observations, the following is the main theoretical result in this appendix (see Ben-Yosef & Ben-Shahar, 2012, for a proof and additional theoretical results):

Theorem 1. *Of all admissible curves in T(I), those that minimize the functional in equation A.4 belong to a two-parameter family* (c, ϕ)*, which is defined by the following differential equation:*

$$\left(\hbar\frac{d\theta}{ds}\right)^2 = \frac{c^2}{sin^2(\theta+\phi)} - 1\ . \tag{A.6}$$

The results in theorem 1 are still short of resolving the parameters c, ϕ of the specific curve from the family of equation A.6, which passes through the given tangent bundle boundary points p_0 and p_1. However, these parameters can be solved by applying a numerical procedure to solve ODE with boundary conditions. To begin, we notice that to facilitate the most general solution (i.e., which also handles inflectional curves), it is preferable to solve the second-order ODE that is obtained by differentiating equation A.6:

$$\hbar^2\frac{d^2\theta}{ds^2} = -\frac{c^2\cos(\theta+\phi)}{\sin^3(\theta+\phi)}\ . \tag{A.7}$$

This way we avoid determination of the sign of the square root when it is applied to equation A.6. Still, at first sight, this approach is problematic since it appears that the number of constraints in our problems is smaller than its degrees of freedom (or free parameters). Indeed, equation A.7 represents a family of planar curves in Whewell form (i.e., an equation that relates the tangential angle of the curve with its arc length), which induces an image parametric curve via the following integrations:

$$x(s) = x_0 + \int_0^s \cos\theta(\tilde{s})d\tilde{s},$$
$$y(s) = y_0 + \int_0^s \sin\theta(\tilde{s})d\tilde{s}. \tag{A.8}$$

Thus, a single and unique curve from our family of solutions is determined by 7 degrees of freedom: θ_0, $\dot{\theta}_0$ (or put differently, the curvature κ_0 at p_0), ϕ, and c are needed to resolve a unique $\theta(s)$ function via equation A.7, and x_0, y_0 and l are then needed to determine the curve's coordinate functions $x(s)$ and $y(s)$ from the first to the second inducer via equation A.8. At the same time, the curve completion problem provides only six constraints expressed by the two given inducers:

$$x(0) = x_0 \quad y(0) = y_0 \quad \theta(0) = \theta_0,$$
$$x(l) = x_1 \quad y(l) = y_1 \quad \theta(l) = \theta_1.$$

Fortunately, this initial observation does not imply that our problem is underdetermined (and therefore lacks unique solutions) since it turns out that c can be expressed in terms of κ_0 and θ_0. To do so we evaluate equation A.6 at θ_0

$$\left(\hbar \left.\frac{d\theta}{ds}\right|_{\theta_0}\right)^2 = \frac{c^2}{\sin^2(\theta_0 + \phi)} - 1,$$

which results in the following identity:

$$c^2 = (\hbar^2 {\kappa_0}^2 + 1) \cdot \sin^2(\theta_0 + \phi). \tag{A.9}$$

Substituting equation A.9 in equation A.7 we obtain

$$\ddot{\theta} = \frac{-({\kappa_0}^2 + \frac{1}{\hbar^2}) \cdot \sin^2(\theta_0 + \phi)\cos(\theta + \phi)}{\sin^3(\theta + \phi)}, \tag{A.10}$$

in which c no longer participates.

Following these algebraic manipulations, we assert that our curve completion problem can be answered by solving equation A.10 and then use the resolved parameters to construct the completed curve with equation A.8. One standard numerical technique for solving such ODE is based on nonlinear optimization that seeks the values of the equation parameters that satisfy the given boundary conditions. In our case, this entails the following general algorithm:

1. Make an initial guess regarding the values of the parameters κ_0, ϕ, and l.
2. Construct a curve of length l starting from $p_0 = [x_0, y_0, \theta_0]$ in a way that obeys equation A.10.
3. Evaluate the correctness of the parameters by assessing the error between the obtained end point of the constructed curve (i.e., the point $[x(l), y(l), \theta(l)]$) and the desired end point ($p_1 = [x_1, y_1, \theta_1]$).
4. Use the error $E(\kappa_0, \phi, l)$ between these two tangent bundle points to update the parameters before iterating back to step 2.

More specifically, for each iteration i with a given starting point p_0 and parameter values κ_0 and ϕ, we first solve the differential equation, equation A.10, via Euler's method (though more sophisticated methods could be used too, of course). Initializing arc length $s_0 = 0$ at the beginning of each iteration we compute

$$
\begin{aligned}
s_{n+1} &\overset{\triangle}{=} s_n + h,\\
\kappa_{n+1} &\overset{\triangle}{=} \kappa(s_{n+1}) = \kappa(s_n + h) \approx \kappa(s_n) + h \cdot \ddot{\theta}(s_n)\\
&= \kappa_n + h \cdot \frac{-({\kappa_0}^2 + \frac{1}{\hbar^2}) \cdot \sin^2(\theta_0 + \phi)\cos(\theta_n + \phi)}{\sin^3(\theta_n + \phi)},\\
\theta_{n+1} &\overset{\triangle}{=} \theta(s_{n+1}) = \theta(s_n + h) \approx \theta(s_n) + h \cdot \kappa(s_n)\\
&= \theta_n + h \cdot \kappa_n,\\
y_{n+1} &\overset{\triangle}{=} y(s_{n+1}) \approx y(s_n) + h \cdot \dot{y}(s_n)\\
&= y_n + h \cdot \sin\theta_n,\\
x_{n+1} &\overset{\triangle}{=} x(s_{n+1}) \approx x(s_n) + h \cdot \dot{x}(s_n)\\
&= x_n + h \cdot \cos\theta_n,
\end{aligned}
$$

where h is a preselected step size and the error is of order $O(h)$. The curve $\beta(s_i) = [x(s_i), y(s_i), \theta(s_i)]$ computed by this step is then evaluated at $s_n = l$ (i.e., at step $n = l/h$) to obtain the point $[x_{end}, y_{end}, \theta_{end}] = [x(l), y(l), \theta(l)]$ and the error $E(\kappa_0, \phi, l)$ associated with the current value of the parameters is computed by

$$E(\kappa_0, \phi, l) = \|[x_1, y_1, \theta_1] - [x_{end}, y_{end}, \theta_{end}]\|.$$

The new values for κ_0, ϕ, and l are then computed by gradient descent on $E(\kappa_0, \phi, l)$. A demonstration of a minimum curve (and its image projection) that is generated by this procedure is shown in Figure 12. See also Ben-Yosef and Ben-Shahar (2012) for many additional theoretical results and analyses related to the problem, including an analytical solution to the completed curves in terms of elliptic integrals.

A.2 Dependency on Scale and Other Visual Properties. So far we have illustrated how the problem of curve completion can be formulated and solved in the space that abstracts the early visual cortical regions, where this perceptual process is likely to occur. Since the theory, and the single principle of minimum action that guides this solution, are nonperceptual, it is important to understand what perceptual properties they entail and how these predictions correspond to existing perceptual findings and the geometrical axioms reviewed in section 2.1.

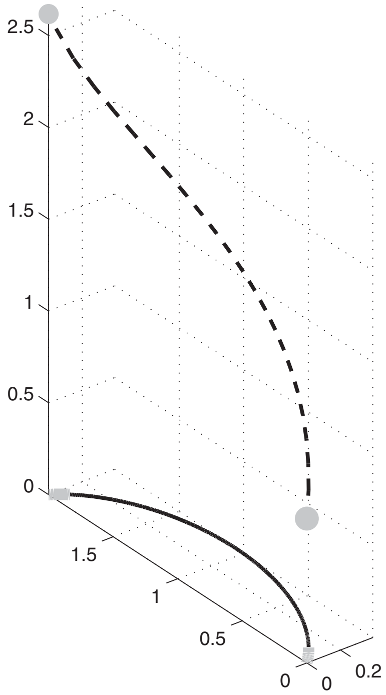

Figure 12: The curve of minimum length in the tangent bundle as computed by our numerical procedure. The shortest admissible path in $T(I)$ between points (marked in gray) $p_0 = [0, 0, 45^\circ]$ and $p_1 = [0, 2, 150^\circ]$ for $\hbar = 1$ is shown in the dashed line, and its projection to I is plotted in solid (blue).

As suggested at the beginning of this appendix, the value of the $\hbar$ constant could have a significant influence over the shape of curves of minimum length in the tangent bundle, as it implicitly controls the relative contribution of total length in I versus total curvature in I during the minimization process, or put differently, the relative scale ratio (i.e., the proportion between units of measurement) of the length and orientation axes in the unit tangent bundle.[5] In this context, the behavior at the limits of $\hbar$ provides important qualitative insights regarding its effect. On the one hand, if $\hbar$ is very small, the minimization process becomes similar to minimization of length in I (subject to boundary conditions), and we therefore expect the

[5] Note that since $\int_{t_0}^{t_1} \sqrt{\dot{x}^2 + \dot{y}^2} dt$ amounts to total length in the image plane while $\int_{t_0}^{t_1} \sqrt{\dot{\theta}^2} dt$ represents total curvature in the image plane, $\hbar$ in equation A.2 can be interpreted as balancing between these two terms.

resultant curve to straighten (or "flatten"). On the other hand, when $\hbar$ is very large, the minimization process is dominated by the minimization of the orientation derivative (again, subject to boundary conditions), a condition that resembles (qualitatively) the classical elastica and converges to a corresponding shape. (See Ben-Yosef & Ben-Shahar, 2012, for formal proofs of these properties.)

In addition to scaling issues, several properties of our model can be pointed out regarding the six axioms of curve completion mentioned in section 2.1. First, since our solution is not linked to any specific frame, it is trivially isotropic. Note that since the rotated minimum length curve also satisfies equation A.6 (for $\tilde{c} = c$ and $\tilde{\phi} = \phi + \rho$, where ρ is the angle of rotation), the solution is invariant under rotations. Second, since the solution minimizes total arc length in $T(I)$, it must be extensible in that space and hence in the image plane also. Third, since the completed curves can be described by a differential equation A.6, they clearly satisfy the axiom of smoothness.

Obviously, the analysis of scale that we discussed indicates that our theory generates scale- variant solutions, or, put differently, it does not satisfy the axiom of scale invariance. Another axiom where our model departs from prior solutions is the axiom of roundedness, since it is easy to confirm that the case of constant curvature ($\frac{d\theta}{ds} = const$) does *not* satisfy equation A.6. At first sight, these two properties could undermine the utility of our model, but given the refutation of both scale invariance and roundedness at the perceptual and psychophysical level (Guttman & Kellman, 2004; Gerbino & Fantoni, 2006; see also the elaborated discussion in Ben-Yosef & Ben-Shahar, 2012), we consider these properties an important advantage of our theory rather than a limitation. That these properties were derived as emergent properties rather than imposed as axioms is yet another benefit of our approach as a whole.

Appendix B: The Neural Network Computation in the Continuum

We show here the relationship between the shortest path in the discrete network described in section 3.2 and the shortest admissible path in continuous tangent bundle $\mathbf{R}^2 \times \mathcal{S}^1$. Let $\beta = \{p_0, p_1, p_2, \ldots, p_n\}$ be a set of vertices in the graph G (see section 3.2) that constitute the shortest (weighted) path in G between the vertices p_0 and p_n. For each vertex p_k in the 3D grid G, define $p_k = [x_k, y_k, \theta_k]$ as its (discretized) coordinates in the grid, and let $\Delta t = 1/n$ such that

$$\begin{aligned} t_k &\overset{\triangle}{=} \sum_{i=1}^{k} \Delta t, \\ \Delta t &= \Delta t_k \overset{\triangle}{=} t_k - t_{k-1}. \end{aligned} \tag{B.1}$$

Since t_k is a monotonically increasing function of k, we now can replace the index k with t_k as follows:

$$\begin{aligned} x_k &= x(t_k), \\ y_k &= y(t_k), \\ \theta_k &= \theta(t_k). \end{aligned} \tag{B.2}$$

Recall that the length (i.e., weight) of the set (or path) β is

$$\begin{aligned} L(\beta) &= \sum_{k=1}^{n} w(p_{k-1}, p_k) \\ &= \sum_{k=1}^{n} \Big[\sqrt{(x(t_k) - x(t_{k-1}))^2 + (y(t_k) - y(t_{k-1}))^2 + \hbar^2(\theta(t_k) - \theta(t_{k-1}))^2} \\ &\quad + \eta|(x(t_k) - x(t_{k-1})) \cdot \sin\hat{\theta} - (y(t_k) - y(t_{k-1})) \cdot \cos\hat{\theta}|\Big]. \end{aligned} \tag{B.3}$$

Multiplying equation B.3 by $\frac{\Delta t_k}{\Delta t_k}$, we get

$$\begin{aligned} L(\beta) &= \sum_{k=1}^{n} \Delta t_k \\ &\times \left[\sqrt{\left(\frac{x(t_k) - x(t_{k-1})}{t_k - t_{k-1}}\right)^2 + \left(\frac{y(t_k) - y(t_{k-1})}{t_k - t_{k-1}}\right)^2 + \hbar^2\left(\frac{\theta(t_k) - \theta(t_{k-1})}{t_k - t_{k-1}}\right)^2} \right. \\ &\left. + \eta\left|\frac{x(t_k) - x(t_{k-1})}{t_k - t_{k-1}} \cdot \sin\hat{\theta} - \frac{y(t_k) - y(t_{k-1})}{t_k - t_{k-1}} \cdot \cos\hat{\theta}\right|\right]. \end{aligned} \tag{B.4}$$

When increasing grid resolution within the same boundaries, one obtains (in the limit) $|V| \to \infty$ and $n \to \infty$, and hence $dt \overset{\triangle}{=} \Delta t_k \to 0$ and $\hat{\theta} \to \theta$. Consequently, assuming unbounded neighborhood radius r, the length of β becomes

$$\begin{aligned} L(\beta) &= \int_{p_0}^{p_n} dt \left[\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \hbar^2\left(\frac{d\theta}{dt}\right)^2} \right. \\ &\quad \left. + \eta\left|\frac{dx}{dt} \cdot \sin\theta - \frac{dy}{dt} \cdot \cos\theta\right|\right], \end{aligned} \tag{B.5}$$

and t becomes a parameter of integration. Since β is chosen to minimize L, β amounts to the shortest admissible curve in $\mathbf{R}^2 \times \mathcal{S}^1$ that is found by employing the Euler-Lagrange equation on equation B.5, where η serves as a Lagrange multiplier.[6] In constrained variational problems like equation B.5, the Lagrange multiplier can be a function of the input boundary values. In our case, this suggests that η may be dependent on the input inducers. However, in practice, η was found to be stable and bounded in a narrow range around $\eta = 3$ (given $\hbar = 13$) regardless of the input. This calibration was done by performing a coarse search over a reasonable domain (e.g., $[0, 20]$) to find η that gives the best match between the network output and the analytical results from Ben-Yosef and Ben-Shahar (2012) over a set of randomly selected inducer pairs. Several of these inducer pairs and their corresponding completion results are shown in Figure 6.

Acknowledgments

This work was supported in part by the National Institute for Psychobiology in Israel (grant no. 9-2012/2013) founded by the Charles E. Smith Family, the Israel Science Foundation (ISF grant no. 259/12), and the European Commission in the 7th Framework Programme (CROPS GA no. 246252). We also thank the Frankel Fund, the Paul Ivanier Center for Robotics Research, and the Zlotowski Center for Neuroscience at Ben-Gurion University for their generous support.

References

Angelucci, A., Levitt, J., Walton, E., Hup, J., Bullier, J., & Lund, J. (2002). Circuits for local and global signals integration in primary visual cortex. *Journal of Neuroscience*, *22*(19), 8633–8646.

August, J., & Zucker, S. (2003). Sketches with curvature: The curve indicator random field and Markov processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(4), 387–400.

Bellman, R. (1954). Dynamic programming and a new formalism in the calculus of variations. *Proceedings of the National Academy of Sciences of the USA*, *40*, 231–235.

Ben-Shahar, O., & Zucker, S. (2003). The perceptual organization of texture flows: A contextual inference approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(4), 401–417.

Ben-Shahar, O., & Zucker, S. (2004). Geometrical computations explain projection patterns of long range horizontal connections in visual cortex. *Neural Computation*, *16*(3), 445–476.

[6] We note that this representation of the functional of equation B.5 may not be the easiest way to solve the variational problem. A better approach could rely on the reparametrization suggested in Ben-Yosef and Ben-Shahar (2012).

Ben-Yosef, G., & Ben-Shahar, O. (2008). Curvature-based perceptual singularities and texture saliency with early vision mechanisms. *Journal of the Optical Society of America A*, *25*, 1974–1993.

Ben-Yosef, G., & Ben-Shahar, O. (2010a). A biologically-inspired theory for non-axiomatic parametric curve completion. In *Proceedings of the Asian Conference on Computer Vision*. Berlin: Springer.

Ben-Yosef, G., & Ben-Shahar, O. (2010b). Minimum length in the tangent bundle as a model for curve completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE.

Ben-Yosef, G., & Ben-Shahar, O. (2012). A tangent bundle theory for visual curve completion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*, 1263–1280.

Bergman, A. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Borg-Graham, L., Monier, C., & Frégnac, Y. (1998). Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature*, *393*, 369–373.

Bosking, W., Zhang, Y., Schofield, B., & Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in the tree shrew striate cortex. *Journal of Neuroscience*, *17*(6), 2112–2127.

Bruckstein, A., & Netravali, A. (1990). On minimal energy trajectories. *Computer Vision, Graphics and Image Processing*, *49*(3), 283–296.

Bushnell, B., Harding, P., Yoshito, K., & Pasupathy, A. (2011). Partial occlusion modulates contour-based shape encoding in primate area V4. *Journal of Neuroscience*, *31*(11), 4012–4024.

Carlyon, P. (2004). How the brain separates sounds. *Trends in Cognitive Sciences*, *8*(10), 465–471.

Citti, G., & Sarti, A. (2006). A cortical based model of perceptual completion in the roto-translation space. *Journal of Mathematical Imaging and Vision*, *24*(3), 307–326.

Cormen, H., Stein, C., Rivest, R. L., & Leiserson, C. E. (2001). *Introduction to algorithms*. New York: McGraw-Hill.

Dobbins, A., Zucker, S., & Cynader, M. (1987). Endstopped neurons in the visual cortex as a substrate for calculating curvature. *Nature*, *329*(6138), 438–441.

Dobbins, A., Zucker, S., & Cynader, M. (1989). Endstopping and curvature. *Vision Research*, *29*(10), 1371–1387.

Dreyfus, S. (1960). Dynamic programming & the calculus of variations. *Journal of Mathematical Analysis and Applications*, *1*, 228–239.

Ffytche, D., & Zeki, S. (1996). Brain activity related to the perception of illusory contours. *Neuroimage*, *3*, 104–108.

Finn, I., & Ferster, D. (2007). Computational diversity in complex cells of cat primary visual cortex. *Journal of Neuroscience*, *27*(36), 9638–9648.

Frégnac, Y., Monier, C., Chavane, F., Baudot, P., & Graham, L. (2003). Shunting inhibition, a silent step in visual cortical computation. *Journal of Physiology*, *97*, 441–451.

Fulvio, J., Singh, M., & Maloney, L. (2008). Precision and consistency of contour interpolation. *Vision Research*, *48*, 831–849.

Gawne, T., & Martin, J. (2002). Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *Journal of Neurophysiology*, *88*, 1128–1135.

Gerbino, W., & Fantoni, C. (2006). Visual interpolation in not scale invariant. *Vision Research*, *46*, 3142–3159.

Gilbert, C., & Wiesel, T. (1983). Clustered intrinsic connections in cat visual cortex. *Journal of Neuroscience*, *3*(5), 1116–1133.

Gregory, R. (1972). Cognitive contours. *Nature*, *238*, 51–52.

Grosof, D., Shapley, R., & Hawken, M. (1993). Macaque V1 neurons can signal 'illusory' contours. *Nature*, *365*, 550–552.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations. *Perception and Psychophisics*, *38*(2), 141–171.

Guttman, S., & Kellman, P. (2004). Contour interpolation revealed by a dot localization paradigm. *Vision Research*, *44*, 1799–1815.

Hirsch, J., DeLaPaz, R., Relkin, N., Victor, J., Kim, K., Li, T., et al. (1995). Illusory contours activate specific regions in human visual cortex: Evidence from functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the USA*, *92*, 6469–6473.

Hoffman, W. (1989). The visual cortex is a contact bundle. *Applied Mathematics and Computation*, *32*, 137–167.

Horn, B. (1983). The curve of least energy. *ACM Transactions on Mathematical Software*, *9*(4), 441–460.

Hubel, D., & Wiesel, T. (1977). Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London, Series B*, *198*, 1–59.

Kanizsa, G. (1979). *Organization in vision: Essays on gestalt perception*. New York: Praeger.

Kellman, P., & Shipley, T. (1991). A theory of visual interpolation in object perception. *Cognitive Psychology*, *23*, 141–221.

Kimia, B., Frankel, I., & Popescu, A. (2003). Euler spiral for shape completion. *International Journal of Computer Vision*, *54*(1–3), 159–182.

Kluender, R., & Jenison, R. (1992). Effects of glide slope, noise intensity, and noise duration on the extrapolation of FM glides through noise. *Perception and Psychophisics*, *51*(3), 231–238.

Koch, C., Poggio, T., & Torre, V. (1983). Nonlinear interactions in a dendritic tree: Localization, timing, and role in information processing. *Proceedings of the National Academy of Sciences of the USA*, *80*, 2799–2802.

Köhler, W. (1920). Physical gestalten. In W. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 17–53). New York: Routledge & Kegan Paul.

Kruggel, F., Herrmann, C. S., Wiggins, C. J., & von Cramon, D. Y. (2001). Hemodynamic and electroencephalographic responses to illusory figures: Recording of the evoked potentials during functional MRI. *Neuroimage*, *14*, 1327–1336.

Lampl, I., Ferster, D., Poggio, T., & Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the max operator in complex cells of the cat primary visual cortex. *Journal of Neurophysiology*, *92*, 2704–2713.

Lawson, R., & Gulick, W. (1967). Stereopsis and anomalous contours. *Vision Research*, *7*, 271–297.

Lee, T., & Nguyen, M. (2001). Dynamics of subjective contours formation in the early visual cortex. *Proceedings of the National Academy of Sciences of the USA*, *98*(4), 1907–1911.

Maertens, M., Pollman, S., Hanke, M., Mildner, T., & Moller, H. (2008). Retinotopic activation in response to subjective contours in primary visual cortex. *Frontiers in Human Neuroscience*, *2*(2).

Mendola, J., Dale, A., Fischi, B., Liu, A., & Tootell, B. (1999). The representation of illusory and real contours in human cortical visual areas revealed by functional magnetic resonance imaging. *Journal of Neuroscience*, *19*(19), 8560–8572.

Miller, G., Dibble, E., & Hauser, M. (2001). Amodal completion of acoustic signals by a nonhuman primate. *Nature Neuroscience*, *4*(8), 783–784.

Mumford, D. (1994). Elastica in computer vision. In B. Chandrajit (Ed.), *Algebraic geometry and its applications*. Berlin: Springer-Verlag.

Murray, M., Wylie, G., Higgins, B., Javitt, D., Schroeder, C., & Foxe, J. (2002). The spatiotemporal dynamics of illusory contour processing: Combined high-density electrical mapping, source analysis, and functional magnetic resonance imaging. *Journal of Neuroscience*, *22*(12), 5055–5073.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. Chase (Ed.), *Visual information processing* (pp. 283–308). Orlando, FL: Academic Press.

Palmer, S. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Petitot, J. (2003). Neurogeometry of V1 and Kanizsa contours. *Axiomathes*, *13*(3–4), 347–363.

Petkov, C., O'Connor, K., & Sutter, M. (2007). Encoding of illusory continuity in primary auditory cortex. *Neuron*, *54*, 153–165.

Purves, D., Augustine, G., Fitzpatrick, D., Hall, W., Lamantia, A., McNamara, J., et al. (Eds.). (2004). *Neuroscience*. Sunderland, MA: Sinauer.

Purves, D., & LaMantia, A. (1990). Numbers of "blobs" in the primary visual cortex of neonatal and adult monkeys. *Proceedings of the National Academy of Sciences of the USA*, *87*, 5764–5767.

Redies, C., Crook, J., & Creutzfeldt, O. (1986). Neuronal responses to borders with and without luminance gradients in cat visual cortex and dorsal lateral geniculate nucleus. *Experimental Brain Research*, *61*, 469–481.

Ringach, D., & Shapley, R. (1996). Spatial and temporal properties of illusory contours and amodal boundary completion. *Vision Research*, *36*(19), 3037–3050.

Rockland, K., & Lund, J. (1982). Widespread periodic intrinsic connections in the tree shrew visual corte. *Science*, *215*(19), 1532–1534.

Rutkowski, W. (1979). Shape completion. *Computer Vision, Graphics and Image Processing*, *9*, 89–101.

Sáry, G., Koteles, K., Kaposvári, P., Lenti, L., Csifcsák, G., Frankó, E., et al. (2008). The representation of Kanizsa illusory contours in the monkey inferior temporal cortex. *European Journal of Neuroscience*, *28*(10), 2137–2146.

Schumann, F. (1904). Beitrage zur analyse der gesichtswhrnehmungen. *Zeitschrift für Psychologie*, *33*, 161–185.

Seghier, M., Dojat, M., Delon-Martin, C., Rubin, C., Warnking, J., Segebarth, C., et al. (2000). Moving illusory contours activate primary visual cortex: An FMRI study. *Cerebral Cortex*, *10*(7), 663–670.

Sharon, E., Brandt, A., & Basri, R. (2000). Completion energies and scale. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(10), 1117–1131.

Sheth, B., Sharma, S., Rao, S., & Sur, M. (1996). Orientation maps of subjective contours in visual cortex. *Science*, *274*, 2110–2115.

Singh, M., & Fulvio, J. (2005). Visual extrapolation of contour geometry. *Proceedings of the National Academy of Sciences of the USA*, *102*, 939–944.

Stanley, D., & Rubin, N. (2003). FMRI activation in response to illusory contours and salient regions in the human lateral occipital complex. *Neuron*, *37*, 323–331.

Stevens, C. (2000). Models are common; good theories are scarce. *Nature Neuroscience*, *3*, 1177.

Sugita, Y. (1997). Neuronal correlates of auditory induction in the cat cortex. *Neuroreport*, *8*, 1155–1159.

Takeichi, H. (1995). The effect of curvature on visual interpolation. *Perception*, *24*, 1011–1020.

Torre, V., & Poggio, T. (1978). A synaptic mechanism possibly underlying directional selectivity to motion. *Proceedings of the Royal Society of London, Series B*, *202*, 409–416.

Tse, P. (1999). Volume completion. *Cognitive Psychology*, *39*, 37–68.

Ullman, S. (1976). Filling in the gaps: The shape of subjective contours and a model for their creation. *Biological Cybernetics*, *25*, 1–6.

Versavel, M., Orban, G., & Lagae, L. (1990). Responses of visual cortical neurons to curved stimuli and chevrons. *Vision Research*, *30*(2), 235–248.

von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, *224*, 1260–1262.

Warren, R. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393.

Weiss, I. (1988). 3D shape representation by contours. *Computer Vision, Graphics and Image Processing*, *41*, 80–100.

Williams, L., & Jacobs, D. (1997a). Local parallel computation of stochastic completion fields. *Neural Computation*, *9*(4), 859–881.

Williams, L., & Jacobs, D. (1997b). Stochastic completion fields: A neural model of illusory contour shape and salience. *Neural Computation*, *9*(4), 837–858.

Yu, A., Gisse, M., & Poggio, T. (2002). Biophysiologically plausible implementations of the maximum operation. *Neural Computation*, *14*, 2857–2881.

Received August 18, 2011; accepted June 14, 2012.