Higher-Dimensional Computational Models of Perceptual Grouping and Silhouette Analysis and Representation

by
Nathaniel R. Twarog
S.B., Massachusetts Institute of Technology (2007)

Submitted to the Department of Brain and Cognitive Sciences
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Cognitive Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2012

Signature of Author:

Department of Brain and Cognitive Sciences
August 28, 2012

Certified by:

Edward H. Adelson, PhD
John and Dorothy Wilson Professor of Visual Sciences
Thesis Supervisor

Accepted by:

Matt Wilson, PhD
Sherman Fairchild Professor of Neuroscience
Director of the Graduate Program

Higher-Dimensional Computational Models of Perceptual Grouping and Silhouette Analysis and Representation

by
Nathaniel R. Twarog
S.B., Massachusetts Institute of Technology (2007)

Submitted to the Department of Brain and Cognitive Sciences
on August 28, 2012 in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Cognitive Science

ABSTRACT

In the following thesis, I describe the investigation of two problems related to the organization and structural analysis of visual information: perceptual grouping and silhouette analysis and representation. For the problem of perceptual grouping, an intuitive model framework was developed which operates on raw images and locates relevant groupings utilizing a higher dimensional space that contains not only the two spatial dimensions of the image but one or more dimension corresponding to relevant image features such as luminance, hue, or orientation. A psychophysical experiment was run to measure how human visual observers perform perceptual grouping across a variety of spatial scales and luminance differences. These results were compared with the predictions of our grouping model, and the model was able to capture much of the grouping behavior of the human subjects. A second experiment was run in which the perception of groups was disrupted by the presence of noise or shifts in brightness. Though the experiments showed only small effects resulting from these disruptions on the behavior of human subjects, the model was still able to successfully capture much of the image-to-image variability.

For the question of silhouette representation and analysis, I suggest that human silhouette representation may be inextricably tied to 3D interpretation of 2D shapes. To support this, I propose a novel algorithm for 2D silhouette inflation called Puffball, which closely matches human intuition for a variety of simple shapes and can be run on almost any input. Using this algorithm, a new model of human part segmentation was derived using 2D-to-3D inflation; this model was evaluated against human-generated part segmentations and two competing part segmentation algorithms. Across a variety of different analyses, Puffball part segmentation performed as well or better than its competitors, suggesting a potential role for 2D-to-3D inflation in the segmentation of silhouette parts. Finally, I suggest several avenues of research which may further illuminate the role of inflation in the human representation and analysis of 2D and 3D shape.

Thesis Supervisor: Edward H. Adelson

Title: John and Dorothy Wilson Professor of Visual Sciences

Acknowledgements

I would like to thank the National Eye Institute of the National Institutes of Health, whose Special Training Program have provided with me with the funding and freedom to pursue my research for much of my graduate career.

I would like to thank the tireless staff of the MIT Brain and Cognitive Sciences department, who have been a constant source of support. I would in particular to like to express my gratitude to the remarkable Denise Heintze, who has been looking after me and keeping me on track ever since she became my freshman advisor in my first year at MIT, back in the fall of 2003.

I would like to thank all the members of my thesis and advisory committee: Edward Gibson, Wilson Geisler, and Whitman Richards, each of whom took time out of their schedule to evaluate my work, offer me their guidance, challenge me to dig deeper, and help me to keep moving forward.

I am of course immensely grateful to my two research advisors, Edward Adelson and Ruth Rosenholtz, who have given me the freedom, the support and the confidence to pursue some of the most fascinating questions I have ever encountered. I could not have hoped for two more knowledgeable, more talented, or more inspiring guides in my exploration of the visual world.

I would also like to thank my parents, Bruce and Barbara Twarog, who, for as long as I can remember (and likely even longer for that) have inspired me to learn, grow, explore, and understand the world around me. They raised me, they taught me, the pushed me to be better, and they have never faltered in their belief in me. Without them, I would never have gotten close to where I am now, and I owe them more than I can express.

And finally, I thank my fiancé, Eleanor Pritchard. Without her support, her care, her affection, her inspiration, her advice, her humor, her cooking, and her love, not a single point in a single graph in this paper would ever have been plotted.

Contents

List of Figures

List of Tables

Introduction

Consider the image in Figure 1A. When you look at this image you likely perceive an array of dots, arranged in six-dot columns, which themselves can be grouped into a large thirty-dot array. What you probablly do not see is dots arranged in five-dot rows; it is possible to perceive these rows, but only with some effort. This is a demonstration of the Gestalt principle of grouping by proximity. If you consider the image in Figure 1B, however, you likely perceive a different organization. Here you most easily see five-dot rows, which themselves can be grouped into perhaps one or two larger arrays. The locations of the dots are identical to those in Figure 1A, but the brightnesses have altered the perceived organization. This is a demonstration of the Gestalt principle of grouping by similarity.

It is important to note that none of these organizations is explicitly present in the image; as arrays of pixels, neither of these images carries any explicit organization whatsoever. Nor is the interpretation of these images semantic; we have no conception of these dots as real world object or entities, and our perception of their organization tells us nothing about their identity or meaning. Yet the perception of structure and grouping is intuitive and inescapable.

Now consider the two silhouettes in Figures 2A and 2B. You almost certainly perceive these shapes as having a complex structural organization: there are clearly identifiable parts, and each of those parts has shape properties which can be remembered and compared with other parts. You also likely recognize a similarity or kinship between the two silhouettes, one which is not shared with the third silhouette in Figure 2C. As with the grouped dots, there is no explicit information present in the images of these silhouettes that suggests or

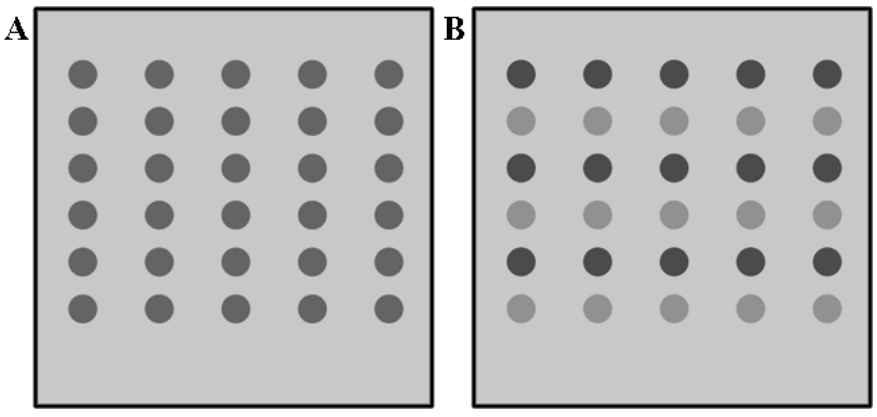

Figure 1: Arrays of dots demonstrating the Gestalt principles of grouping by proximity and grouping by similarity.

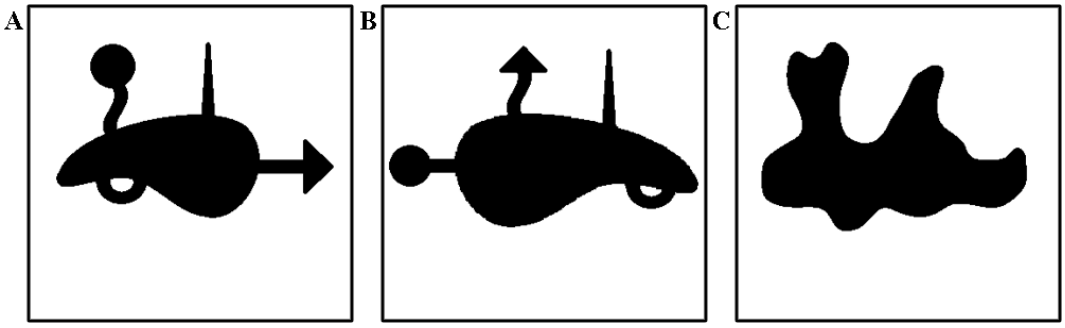

Figure 2: Several silhouettes illustrating shape parts analysis and similarity.

necessitates these judgments. One could segment these silhouettes into parts in any arbitrary way, and each would be just as correct. There is no simple image measure which can capture the similarity that we perceive between Figure 2A and Figure 2B; for example, the raw pixel difference between the two images is very large, and smaller between A and C than between A and B. Nor are these judgments semantically derived: your perception of the silhouettes cannot be governed by their identities as objects because they have no identities as objects. Yet the parts analyses and similarity judgments made by humans with shapes such as these are both strong and consistent.

Though these examples seem somewhat trivial - rarely in your life will your survival or success depend on perceiving rows of dots, and you will likely never encounter the silhouettes in Figure 2 outside of this pape - they are indications of mechanisms which play an essential role in how we perceive and organize the world. To identify the objects and materials in the world, we must first separate them from one another. This requires a powerful system for identifying, without top-down semantic knowledge, what parts of a visual scene are generated by the same item or process; the Gestalt dot displays reveal this system at work. And few cues tell us more about the identity, function, structure, and behavior of an object or entity in the world than its shape; because the information presented to the eyes is two-dimensional, how our visual system processes, organizes, and represents two-dimensional shape - and relates it to three-dimensional shape - is fundamental to a successful understanding of the world we interact with.

Both of these phenomena lie in the broad and poorly understood realm between low-level processing and high-level semantic interpretation that is often referred to as mid-level vision. Both deal with the structural and geometric interpretation of the visual world, above and beyond what is explicitly present in the input. And both, I shall try to demonstrate, fall into a class of visual problems which require a careful balance of scientific and engineering considerations. One cannot simply treat these problems as black boxes, trying different inputs, observing the outputs, inferring the dependence between the two; the underlying computations are too complex. We cannot understand the computational structure without some idea of what computational structures we are looking for; if our existing toolbox of computational tools is insufficient, we must seek new tools. Thus the scientist's goal of seeking understanding cannot be divorced from the engineer's goal of seeking newer and more intuitive solutions to real world problems.

However engineering without science is also insufficient; there are many possible solutions

to any problem, and most will offer little to no scientific insight when lined up next to the human visual system. Our new models therefore must also be grounded not only in the mechanics of computation, but in the conceptual framework of perception and perceptual science. The behavior of our models must be relatable to measurable human behavior, and they must be intuitive enough and adaptive enough that we can gain further insight and understanding. In the following thesis, I will describe two such models; one of perceptual grouping, the other of silhouette analysis and representation. Neither is the first model to approach these problems, and neither will be the last word in their respective domains; but both offer a new and unique bridge between the too-often separated worlds of human vision and computer vision, opening the door to new insights that would not have been possible without them.

Part I

Gestalt Grouping

Chapter 1

Motivation

When the visual system is presented with a scene, its goal is to extract from that scene semantic information about the outside world, including what items and entities are present and how they are spatially arranged relative to one another. It is implausible that the visual system would accomplish this by learning the relationship between complete scenes and full images; there are too many possible images for the visual system to experience and it is difficult to believe that it would see the same one twice. Therefore, the visual system must have a way of breaking the visual world into pieces and indentifying the individual components.

One way to accomplish this is to combine the process of locating the pieces and identifying the objects present into a single step, by searching for known object images in the visual scene. This template-matching approach can be very effective, and is widely used in computer vision; but it still suffers from several limitations. First, it would be highly inefficient to search for every object category in every visual scene. Second, many objects, even some that we see quite often, are too complex and variable to be expressed by one or a small number of templates. Also, many scene components or real world entities do not have a consistent appearance, including water, the sky, and other amorphous entities. Finally, such an approach would have no way of dealing with a novel object. If the visual system is to robustly comprehend the information presented to it, it must organize the visual world before it understands it.

It is therefore necessary that the visual system be equipped with mechanisms for piecing apart the visual world without full knowledge of the real-world entities that are present. It is these mechanisms that the Gestalt theorists investigated when they proposed the Gestalt laws of grouping, patterns and regularities which allow the visual system to infer that different parts of the visual world are generated by the same underlying process. However, the precise computational nature of these mechanisms remains an open question. In this part of my thesis, I propose a computational framework which simply and intuitively captures much of the strength and variety of these organizational principles. I then describe how this framework can be extended and evaluated in comparison with human behavior, and describe a pair of psycophysical experiments which measure human subjects' ability to identify groups under various conditions of proximity, similarity, and noise. I show the results of our model framework capture much of the variation in human behavior in these experiments, and finally describe several ways in which the model framework might be extended or improved.

Chapter 2

Previous Work

One cannot discuss the history of the study of perceptual organization without discussing the Gestalt theorists around the turn of the century. The Gestalt school of psychology arose, in large part, in response to Wilhelm Wundt and the Structuralist school, which attempted to codify the elements of thought and behavior in much the same manner as chemists, analyzing each element as the composition of smaller elements, down to the fundamental particles of mind. The Gestalt school, on the other hand, felt that one could not understand the mind by breaking down the elements of thought into smaller components. They argued that the most important aspect of the mind was not the particles, but the arrangements and forms in which they occurred; Christian von Ehrenfels, one of the earliest proponents of the Gestalt approach, introduced the idea of a gestalt (German for "shape" or "form") and gave the example of a musical melody, defined not by the identity of the individual notes but rather by their placement and timing relative to one another (Ehrenfels, 1937). Later theorists, most notably Koffka, Köhler, and Wertheimer furthered the concept of the gestalt as the fundamental tool of perception (Koffka, 1922; Wertheimer, 1923; Köhler, 1929).

The ideas of the Gestaltists were nowhere more influential than in the study of vision. The creation or detection of gestalts in the visual world was said to be the result of several perceptual "laws" meant to enforce the Gestalt principle of pragnanz, a German term best translated as "pithiness," and usually understood as implying simplicity, regularity, and elegance. The proposed "laws" included the law of proximity, the law of similarity, and the law of good continuation, as well as other laws such as common motion, symmetry, parallelism, and closure.

The term "law" here requires qualification because, unlike the laws of physics and chemistry by which they were inspired, the Gestalt laws were grossly limited in their specificity and predictive power. Though based on extensive observation, and at least anecdotally correct, the Gestalt theory offered little insight into how these principles relate to one another, when one applies and another does not, how competing principles are to be resolved, and what exceptions may exist. In addition, little or no quantitative specification of the principles was given, and the Gestalt theorists made few suggestions as to how these principles were implemented neurologically or computationally. However, given the compelling (albeit non-quantitative) evidence of the existence of these principles in some form, it was inevitable that later perceptual scientists would begin to fill in these gaps.

Much of this later work has focused on the conditions under which grouping of perceptual

elements does or does not occur. In the study of contour integration, initial studies focused on the ability of subjects to detect straight lines in fields of randomized distractors (Smits et al., 1985; Beck et al., 1989; Moulden, 1994); but the more flexible path paradigm developed by Field et al. (1993) showed that humans are able to perceive completed contours with considerable change in orientation. Work by Geisler et al. (2001) showed a close relationship between the conditions under which contour integration occurs and the image statistics of nearby oriented elements in the natural world, while work by Elder and Zucker (1993) highlighted the apparent importance of closure in contour completion.

Grouping by proximity and similarity have also received considerable attention. A number of studies have investigated the strength or relative strength of different grouping contexts, either by having subjects adjust the parameters of a display until two competing organizations are evidently in equilibrium (Rush, 1937; Hochberg and Silverstein, 1956; Hochberg and Hardy, 1960; Oyama et al., 1999) or by having subjects report which grouping they perceive in a variety of potentially ambiguous grouping displays (Oyama, 1961; Callaghan, 1989; Kubovy and Wagemans, 1995; Quinlan and Wilton, 1998; Claessens and Wagemans, 2005). The results of these studies ranged from simple rank-ordering of grouping principle strengths to sophisticated probabilistic models (for an in depth discussion of many of the various results and conclusions, see Kubovy and van den Berg (2008)).

The grouping experiments described above largely make use of controlled and highly abstract stimuli containing clear and discrete elements (dots, line segments, Gabor patches, or simple shapes); this is a sensible approach as it allows for careful parametric control of the conditions of the experiment and avoids the confounding top-down influences that would arise with real-world objects and natural images. However, an unfortunate side effect is that the several mathematical and/or computational models proposed to explain the results in grouping by proximity and similarity (Kubovy and Wagemans, 1995; Kubovy et al., 1998; Kubovy and van den Berg, 2008) all implicitly or explicitly presuppose the existence of these discrete elements to evaluate their coherence. But this representation on which the models depend - visual information neatly parceled into contained, finitely describable elements - is precisely the form of visual information we are trying to reach when we perform perceptual organization. In a sense, in order to operate, the models must assume that the problem of organizing the visual input is largely already solved. This approach, unfortunately, cannot be extended to the more general visual domain. In contour integration, integration of the computational and mathematical has been somewhat more progressive, with several computational models actually operating on raw image data (Lowe, 1985, 1989; Gigus and Malik, 1991); nevertheless, the majority of models still make use of discrete contour elements and mathematically defined association fields (Grossberg and Mingolla, 1985; Ullman and Sha"ashua, 1988; Parent and Zucker, 1989; Kellman and Shipley, 1991; Field et al., 1993; Elder and Zucker, 1996; Jacobs, 1996; Yen and Finkel, 1998; Elder and Goldberg, 2002).

Consider Figure 2.1A. This is still a relatively simple scene, and the elements can be easily identified; but what are their parameters? What is the proximity in this scene? Figure 2.1B is even more challenging. How do we describe the segments of this image? What is the distance between them? Are they similar in shape? Any model which depends on abstract parametric representations of an image has very limited utility beyond the narrow constraints of the experiment on which it is built.

Another class of models comes out of the field of computer vision. These models tackle the

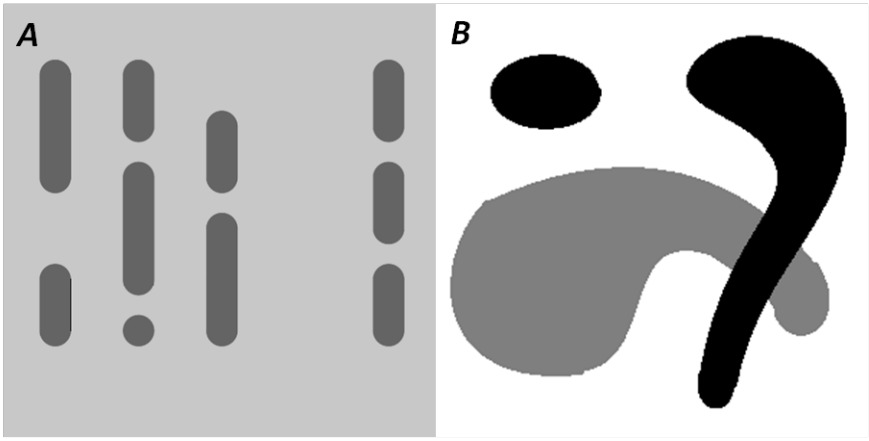

Figure 2.1: (A) A simple scene that cannot be represented by a small number of parameters. (B) An even more complex but still easily understandable scene.

problem of image organization directly, operating on raw image data; they utilize methods like normalized cuts (Shi and Malik, 2000; Malik et al., 2001) and mean-shift (Comanciu and Meer, 2002; Paris and Durand, 2007). Unfortunately, though some work has been done to ground these models in human segmentation behavior (Martin et al., 2001), this work has focused only on continuous segments. Little work has been done evaluating the ability of these models to group elements across breaks or occlusions; nor has any work been done to compare the behavior of the models with human behavior on the large variety of classic Gestalt stimuli which form the foundation of human grouping research.

In addition, while these computer vision models are mathematically well grounded and operate on a wide variety of inputs, they are often nonintuitive and difficult to manipulate. The normalized cuts model is dependent on very large and nonintuitive linear algebra operations, making it difficult to predict how manipulations of the model will affect its output. And both normalized cut and mean-shift have a counterintuitive tendency to over-segment, particularly in large flat image regions, such as the sky (Figure 2.2).

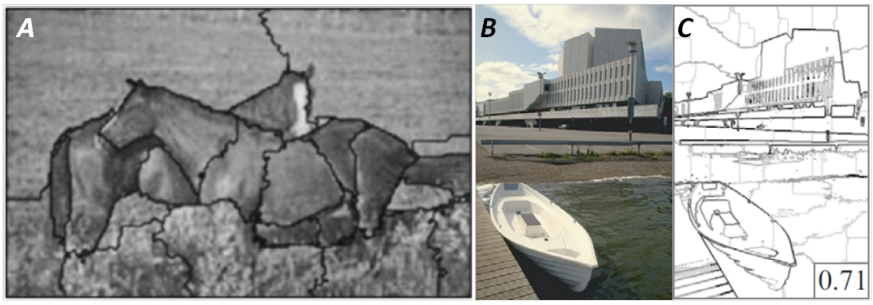

Figure 2.2: (A) An image segmented by the Normalized Cuts algorithm (Malik et al., 2001). Note that the field in the background and the main body of the horse have been oversegmented somewhat arbitrarily. (B) An input image for the mean shift segmentation algorithm (Paris and Durand, 2007). (C) The results of mean shift on that image. Again, large flat regions have been arbitrarily oversegmented.

What is needed therefore is an approach that combines the strengths of both these classes of model. A successful and informative model of perceptual grouping should be intuitive to manipulate, and grounded in the simple Gestalt principles that form the basis of the perceptual grouping literature. Efforts should be taken to relate that model to psychophysical results in a way that offers insight into how to refine and improve the model. But the model must also be computational, rather than mathematical; it should operate on images rather than descriptions, and should be versatile enough to handle a wide range of inputs, including those not specifically designed for psychophysics experiments. My objective is to describe such a model here.

Chapter 3

The New Idea: From Grouping to Clustering

Consider the image in Figure 3.1. The most common percept when viewing this figure is that of dots arranged in columns, all part of a larger grid. It is possible to perceive the dots organized as rows rather than columns, but only with some effort. It is more natural to perceive the dots grouping vertically than horizontally; this is consistent with the classical Gestalt principle of grouping by proximity.

Suppose we wish to develop a computational process which will yield the appropriate organization of the dots in this image; that is, an algorithm or function which will take this image as input, and produce an output which identifies five vertically oriented groups or segments corresponding to the five columns of dots. If we wish our process to be of any use beyond this simple toy example, the process should not be specific to this stimulus, or even this class of stimuli. In short, the process should not seek or identify dots, grids, or columns, as these entities may not have any meaning when organizing other visual inputs.

Because this signal is an image, a natural toolbox to draw from when developing our algorithm is that of image processing. One of the most basic operations in image processing

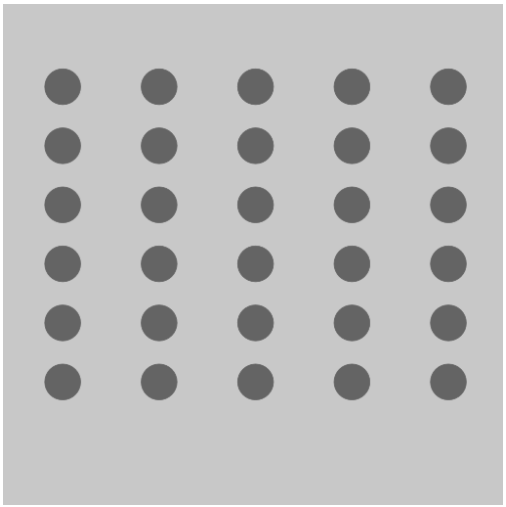

Figure 3.1: A simple Gestalt array. The most natural percept is dots arranged in columns.

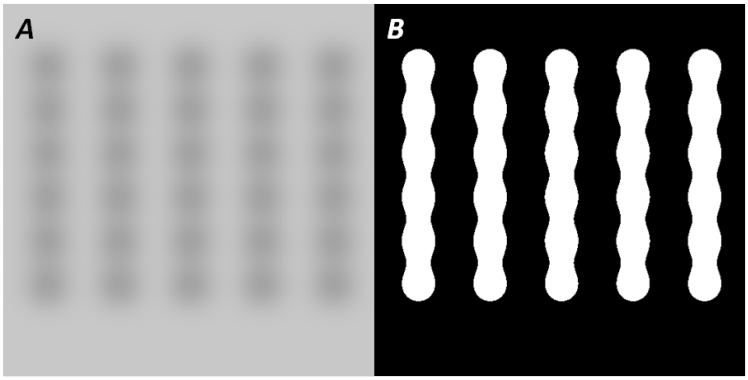

Figure 3.2: (A) Blurring the image with a Gaussian filter causes nearby dots to blend into one another. (B) Passing (A) through a threshold yields 5 columnar segments.

is filtering; and if we filter the above image with a basic blurring filter such as a Gaussian (of sufficient size), we see that the dots in the columns blur into one another (see Figure 3.2A). Selecting all parts of the image below some threshold will pick out five segments overlapping the five columns that can be perceived in the image (Figure 3.2B). This process of filtering an image with a varying smoothing kernel like the Gaussian to locate image structure is closely related to the scale-space approach to image analysis (Witkin, 1983; Koenderink, 1984).

But how would we choose this threshold? Choosing the threshold incorrectly will fail to identify the appropriate grouping in this image, and different images will require different thresholds. A better solution would be one which requires no choice of threshold; we can again take our cue from the scale-space literature and peform edge-detection at the appropriate scale to partition the image and more robustly analyze its structure (Marr and Hildreth, 1980; Babaud et al., 1986; Perona and Malik, 1990). If we filter the image not with a Gaussian filter, but with a difference-of-Gaussians filter (Figure 3.3A), partitioning the image at zero-crossings identifies those regions of the image which are either darker or lighter than the areas around them. Identifying the areas of negative response in Figure 3.3A yields same five columns located before, without the need for a specific threshold (Figure 3.3B).

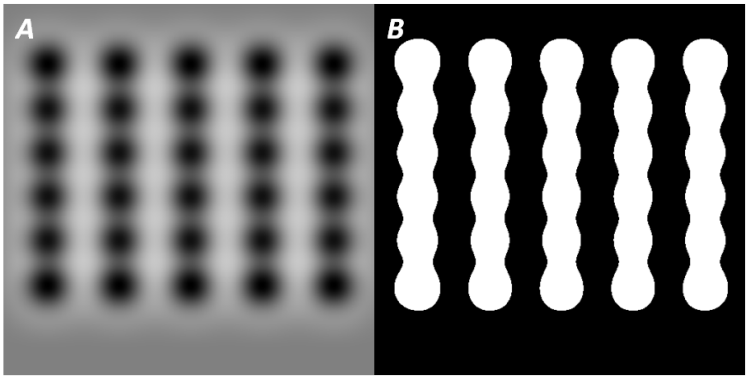

Figure 3.3: (A) Filtering with a Difference-of-Gaussians yields regions of positive and negative response. (B) The zero-crossings carve out image pieces without the need for an adaptive threshold.

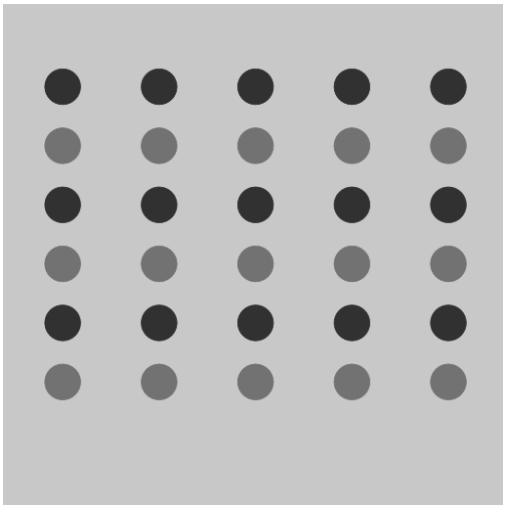

Figure 3.4: A more complex Gestalt dot array. The introduction of different luminance values makes the perception of rows much easier, and the perception of columns more difficult.

While this approach can work quite well for images with only two luminance values, introducing a third luminance reveals even deeper limitations. Consider now the image in Figure 3.4. The locations of the dots in this image are identical to those in Figure 3.1, but the luminance values of alternating rows are now noticeably different. In this image, it becomes far easier to perceive the dots as grouping by rows; indeed the perception of rows now dominates the perception of columns. In the classical Gestalt framework, the principle of grouping by similarity (in this case luminance similarity) has overridden grouping by proximity. However, if we filter the image with a difference-of-Gaussians as before, we find that looking at areas of negative value does not give this result (Figure 3.5). In fact, it gives a weaker version of the column interpretation we saw for Figure 3.1.

The problem is that filtering the image, regardless of the filter type, blends and merges nearby pixel values, washing out and often obscuring the complexity and structure present in the original signal. The solution is to filter not the image, but a higher dimensional

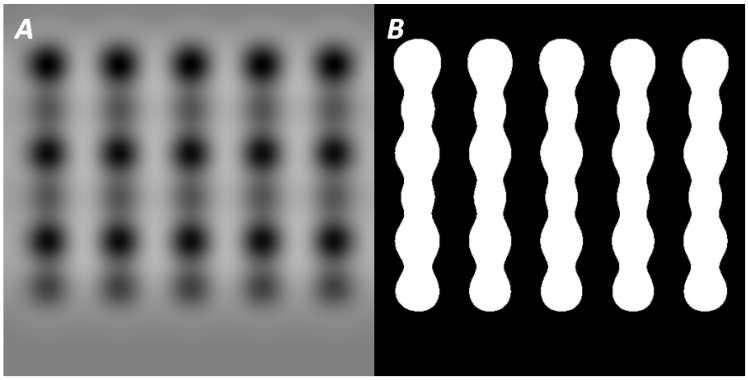

Figure 3.5: (a) Filtering with a Difference-of-Gaussians does not yield an intuitive results. (b) The zero-crossing still identify columnar segments.

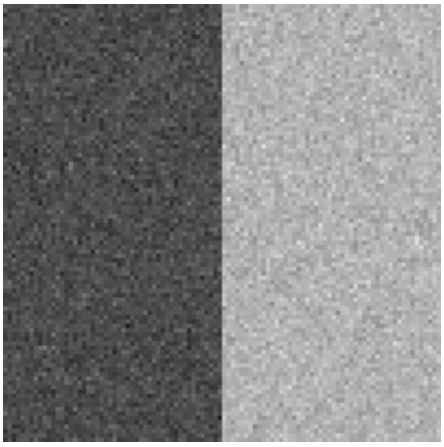

Figure 3.6: An image with a clearly discernible organization.

representation of the image in which pixels with different feature values differences which signal the structure of the world which generated the image do not interfere with one another.

This approach was first described in Rosenholtz et al. (2009). The new representation of the image exists in a higher-dimensional space which has not just the two spatial dimensions of the original image, but one or more dimensions corresponding to the relevant feature in a given image. In the above examples, the feature would be some measure of luminance (e.g. L^* of the CIELab colorspace). Given an image $I(x, y)$ which maps locations in the image domain to luminance values, we define a new function J on x-y-L^* space:

$$J(x, y, L^*) = \delta(L^* - I(x, y)) \tag{3.1}$$

If one views an image a continuous function, J will look like a surface in x-y-L^* space; but if we view an image as a discrete function, the resulting function J can be seen as a three-dimensional scatterplot, in which pixels at nearby locations with similar luminance will map to points that are near one another in a three-dimensional Euclidean space. Thus the problem of grouping pixels becomes a problem of grouping nearby points; in other words, a grouping problem becomes a clustering problem.

Take the image in Figure 3.6. This image contains two regions of Gaussian noise; though the distributions of pixel values in these two regions do overlap slightly, the division between them is quite clear. If we map this image into x-y-L^* space as described above, the function J will contain two clouds of points, centered on different luminance levels (Figure 3.7A). There are many ways to perform this clustering, but one method that works very well is to filter this three-dimensional space with a three-dimensional difference of Gaussians:

$$J_{\sigma_s,\sigma_L} = J * (G_{\sigma_s,\sigma_L} - G_{1.5\sigma_s,1.5\sigma_L}) \tag{3.2}$$

where

$$G_{\sigma_s,\sigma_L}(x, y, L^*) = \frac{1}{(2n)^{3/2}{\sigma_s}^2\sigma_L} \exp\left(-\frac{x^2 + y^2}{2{\sigma_s}^2} - \frac{{L^*}^2}{2{\sigma_L}^2}\right) \tag{3.3}$$

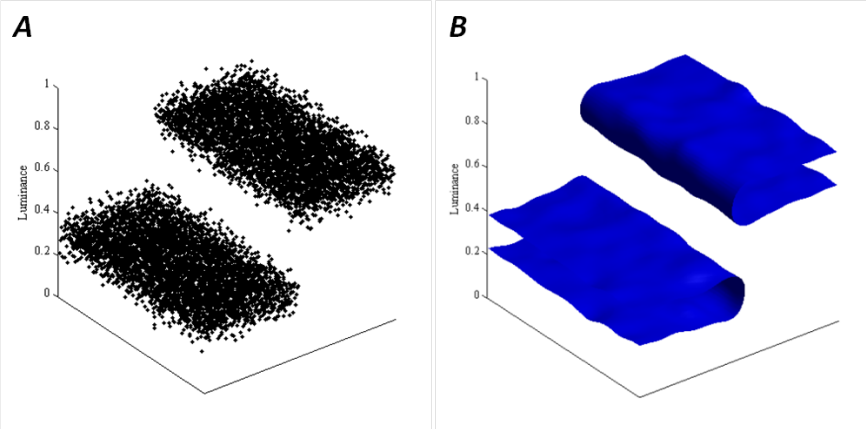

Figure 3.7: (A) A representation of the image in Figure 3.6 in x-y-L^* space. Note that the two regions of the image map to two clouds or clusters in x-y-L^* space. (B) The zero-crossings of the function J_{σ_s,σ_L} which results from filtering J (Figure 3.7A) with a difference-of-Gaussians filter. These zero-crossing clearly mark the boundaries of the two clusters that can be seen in (A).

The zero-crossings of the resulting function J_{σ_s,σ_L} mark the boundaries of regions of x-y-L^* space with positive response (Figure 3.7B). These regions correspond to the identified clusters, which themselves correspond to the predicted pixel groups.

Of course, luminance is not the only feature that one can measure in an image; and grouping based on luminance will not always give the right interpretation. Consider the image of intersecting arcs in Figure 3.8A. If we were to apply our grouping model as described above to this image, it would easily separate the black curves in the foreground from the white background, but would identify the two arcs as single group (Figure 3.8B), which is not at all how we perceive them.

At the point of intersection, the feature that separates the two arcs is not their luminance, but their orientation. Suppose then that instead we begin by calculating the strongest orientation at each point in this image. There are many ways to do this; we use a technique from Landy and Bergen (1991) that utilizes steerable filters (Freeman and Adelson, 1991). The result is shown in Figure 3.8C. We can now map every pixel of our image to a point in x-y-θ space, where θ is a circular dimension ranging from 0 to π representing orientation. Of course, we do not wish points with absent or imperceptible oriented energy to influence the perceived organization; so when we map a pixel to a point in x-y-θ space we weight the pixel with the strength of the orientation at that point (Figure 3.8D). When we view the representation of the image in this space, we can see that the points corresponding to the pixels of the two arcs are now cleanly separated from one another; once again, mapping our image to the appropriate x-y-feature space has turned a difficult grouping problem into a very simple clustering problem.

If we filter the x-y-θ space as before, oriented segments and elements that are near one another and similar in angle will group together; in certain circumstances this performs an effective contour integration result (e.g. Figure 3.8A), but in general this approach is far closer to a model of segmentation of simple oriented textures. For example, if we map

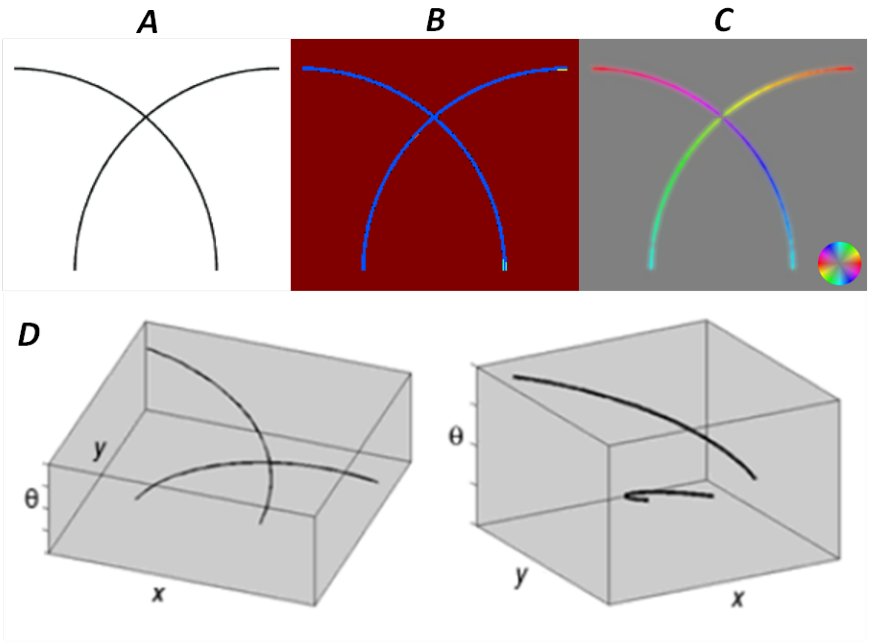

Figure 3.8: (A) Two intersecting arcs. (B) Segmenting them according to luminance similarity treats them as a single connected entity. (C) A representation of the oriented energy in (D). (d) The oriented energy of (a) mapped in to x-y-θ space.

the image in Figure 3.9A into x-y-θ space, and filter the x-y-θ space representation with a difference of Gaussians as we did with x-y-L^* space, the image is cleanly separated into two texture regions (Figure 3.9B).

If we wish to model contour integration more generally, we must refine our approach slightly. Consider the oriented segments in Figure 3.10A. It is clear that according to good continuation, element a should group with element b, as they are aligned and similar in orientation; and element a should not group with element c because their orientations are quite different. But neither should element a group with element d; for, though they have

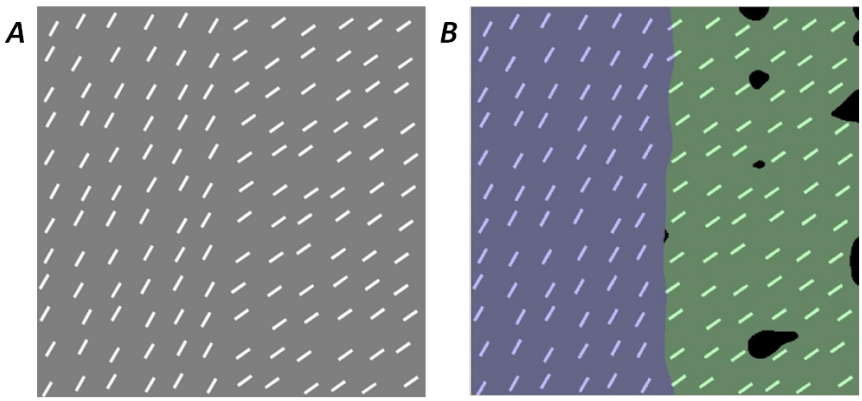

Figure 3.9: (A) An image with two distinct oriented texture. (B) The groups identified by our orientation grouping model.

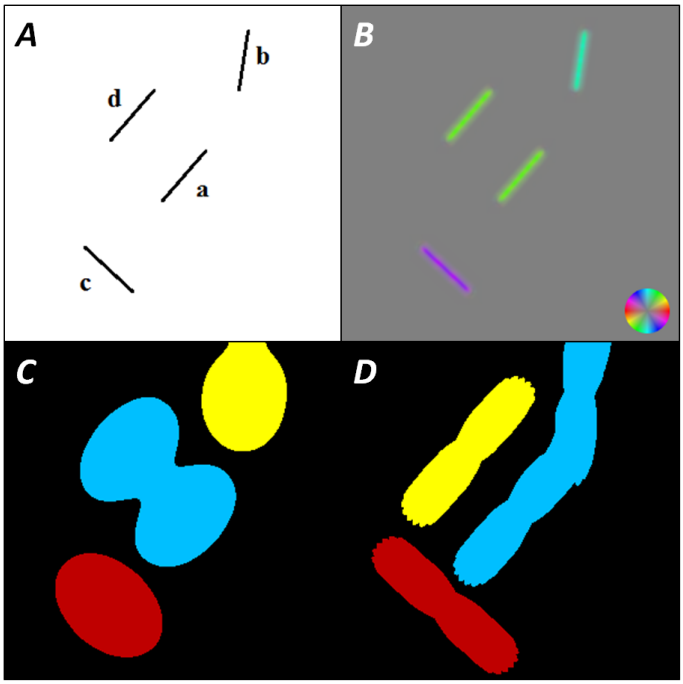

Figure 3.10: (A) Several oriented segments. The law of continuation stipulates that a should join with b. (B) Oriented energy in (A). (C) If one blurs x-y-θ space isotropically, nearby parallel contours blur into one another. (D) Blurring anisotropically yieds the appropriate continuation. Becuase we are blurring with a difference of gaussians filter, the reponse is strongest at the ends of the oriented elements, resulting in slightlly lower response near the center of the elements.

identical orientation, they are not aligned, and thus are unlikely to have been generated by the same contour. However, if we run the orientation grouping algorithm described above, the two parallel segments group quite easily (Figure 3.10C).

To achieve contour grouping rather than texture grouping, we must alter the way we filter x-y-θ space. Until now, we have chosen our difference of Gaussians filter so that each of the two Gaussians blurs isotropically in the x-y plane; distance, not direction, was what mattered to grouping. But this is not the case for contours; so, to achieve a more effective contour grouping we instead filter the space anisotropically. Specifically, each slice of x-y-θ space is filtered with an anisotropic difference-of-Gaussians each oriented along that slices corresponding orientation. Thus, points in a particular slice of x-y-θ space will be more likely to blur together if they are both similar and aligned in orientation. Using this technique, we see that the elements of Figure 3.10A group together much more intuitively (Figure 3.10D). This model implements an implicit association field between nearby and similar oriented elements, much like models described by previous work on contour integration and good continuation (Parent and Zucker, 1989; Field et al., 1993; Yen and Finkel, 1998; Geisler et al., 2001); our model differs from these approaches, however, in that it is implemented in the language and framework of filtering and image processing. This added flexibility allows it to be applied to any possible image input, and process continuous contours just as easily – or even more easily – than isolated contour elements.

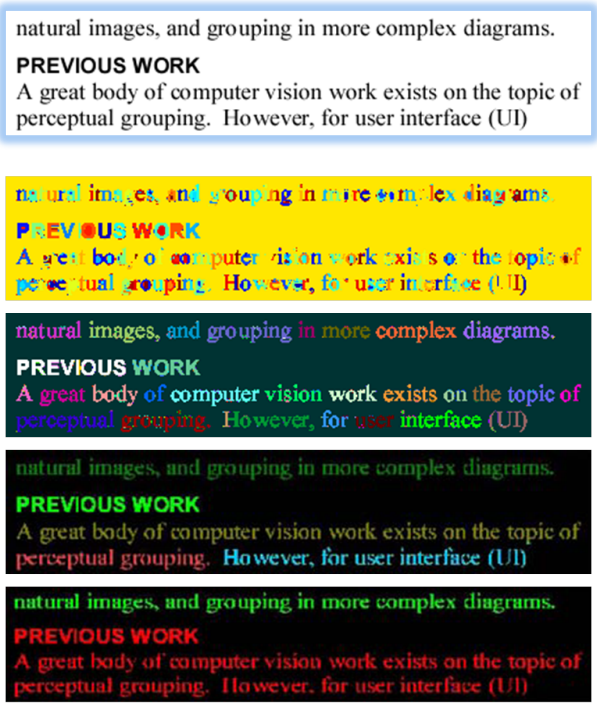

Figure 3.11: The influence of the spatial blurring parameter.

While this model has the power to recreate a variety of classical Gestalt phenomena, and the added versatility afforded by operating on raw images rather than highly constrained scene descriptions, perhaps its greatest strength is its intuitiveness and simplicity. The wide range of possible model outputs stem from two parameters, σ_s and σ_L, and the effect of these two parameters is highly transparent. Take the black and white text image in Figure 3.11. When we look at this picture we do not see just one grouping; we perceive a hierarchy of organization, ranging from the contiguous segments of the individual letters, to the tightly arranged words, to lines and sentences, and finally the sections and paragraphs of the overall text. This rich multiscale organization is beautifully mirrored in the output of the luminance grouping model across a range of values for σ_s. This example illustrates the highly intuitive behavior of σ_s: as σ_s increases, so does the scale of the resulting groups.

Figure 3.12 similarly illustrates the effect of varying the luminance blurring parameter σ_L. Given the patchwork quilt image shown, differing values of σ_L give very different organizations of the image. The lowest value identifies the faintly discernible individual squares making up the smallest scale of the patch. Increasing σ_L then groups together the smaller squares, identifying the larger-scale patches of similar luminance. Finally, at the broadest setting of σ_L gives a segmentation in which all tiles have grouped except the very salient lower right patch, which has a very different luminance from all its neighbors. Thus the function of σ_L is quite clear: as it increases, the segmentation becomes more insensitive to luminance variations and the internal luminance variability of the resulting groups is higher.

Though only the implementations of luminance and orientation are described here, any low-dimensional feature that can be measured throughout an image can be fit into this framework. One could implement grouping by one or dimensions of color (e.g. hue, saturation), a

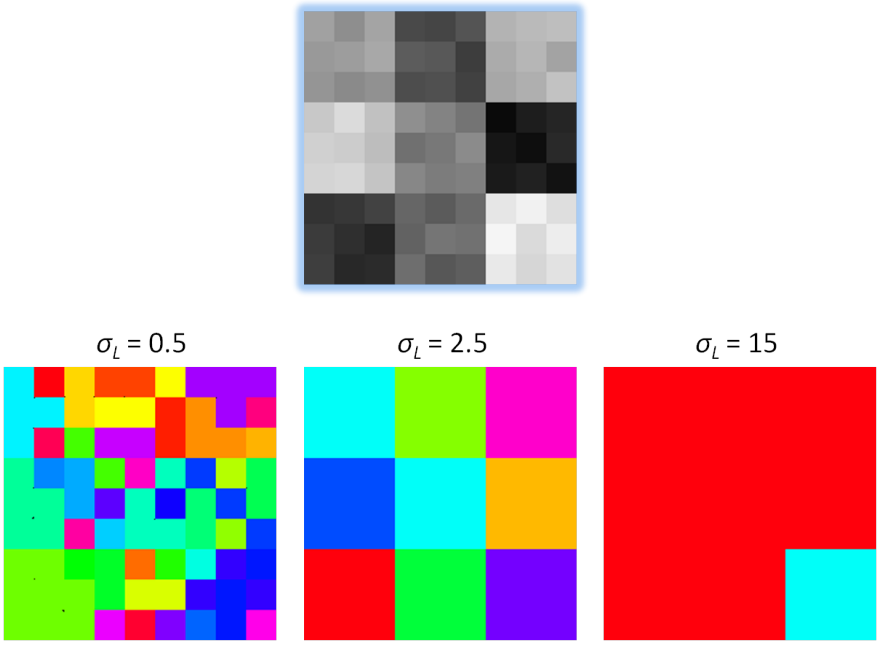

Figure 3.12: The influence of the luminance blurring parameter.

low-dimensional representation of local texture, or contrast energy; the power of the model is that it can convert a wide variety of different dimension, features and properties traditionally each approached with their own classes of models into a single, easy-to-understand computational framework.

Of course, the model is not without its limitations. While the clustering approach that we employ filter with difference of Gaussians and use zero-crossings to identify regions of high-density performs very well in most cases, it can sometimes undersegment. For example, given the image in Figure 3.13, the model would likely identify all pixels as belonging to one

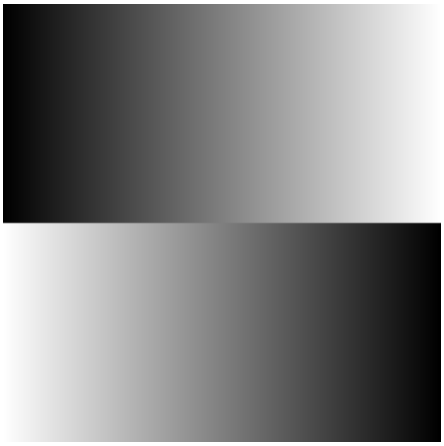

Figure 3.13: Two adjacent gradients, which the visual system easily separates, will be grouped together by our model because the very small point of similarity in the middle of the image.

group, due to the very small point at the edge where the luminances are equal. There are also some clear examples of undersegmentation in the finest grained segmentation of the quilt image in Figure 3.12.

In addition, the model is computationally very demanding to implement, as it requires filtering a high-dimensional space. This limitation is the primary reason we have not implemented grouping with features represented by more than one dimension: though there is no theoretical obstacle, processing such a large data structure becomes prohibitively slow and demands a great deal of space. However, while such calculations are very inefficient in modern digital computers, they would not necessarily be inefficient in a highly-distributed, highly parallel information processing system like the human brain. Indeed, as several researchers have observed, the connections present in V1 and V2 closely mirror the circular orientation space described in our model of orientation and contour grouping (Bosking et al., 1997; Yen and Finkel, 1998; Ernst et al., 2012). And given that cells in v1 V2 are known to fire at the locations of illusory or completed contours (von der Heydt et al., 1984; Grosof et al., 1993), it is not unreasonable to propose that a calculation like the one described in our model might be utilized by V1 and V2 for the integration and completion of contours.

Chapter 4

Putting It to the Test

4.1 Preliminary Results

Figure 4.1 shows the results of our grouping model on two Gestalt dot arrays, specifically those from Figures 3.1 and 3.4. At smaller spatial scale, the dots of Figure 4.1A, which are of the same luminance, are grouped together into columns, giving the perceptual organization most people would report when viewing this image. At a broader spatial scale, the dots in the same row also group together; the resulting segmentation identifies all 30 dots as a single segment, as well as the lighter background. Figure 4.1D, on the other hand, exhibits a rather different behavior. At a smaller spatial scale (and sufficiently small luminance scale), adjacent dots in a column do not group together, because they sit at different levels in x-y-L^* space. Thus each dot is identified as a complete segment. At a broader spatial scale, adjacent dots in rows now group together; because dots in the same column remain separate, the resulting segmentation identifies the rows of dots as segments. So, with the right parameter settings, the model can recreate the intuitive structure of these images.

Figure 4.2A shows a random field of oriented elements from Geisler et al. (2001); Figure 4.2B depicts the sets of elements which are grouped by a pair-wise local grouping function derived from the co-occurrence statistics of contours in natural images. These groupings represent the contour integration selections of an ideal observer based on the learned statistics of the natural world, and were shown by Geisler et al. to agree well with the contours perceived by human observers. The results of our contour integration function, using only the image in Figure 4.2A as input, are shown in Figure 4.2C; all contour groupings which covered more than one contour element are shown. Not only does the model successfully locate the largest and most salient contour, it also closely mirrors the predictions of the model from Geisler et al. in grouping the remaining elements. Most importantly, it does all this with no implicit or explicit representation of individual contour elements.

In addition to these classical psychophysics stimuli, we also tested our model on several figures described and analyzed by information visualization expert Edward Tufte (1983). One example is shown in Figure 4.3A; this graphic depicts cancer rates among white females in counties across the United States, where darker values indicate higher rates of cancer.. According to Tufte, when viewing this figure, observers will note the large number of high rate counties in the Northeast, along with isolated high-rate pockets in northern Minnesota and

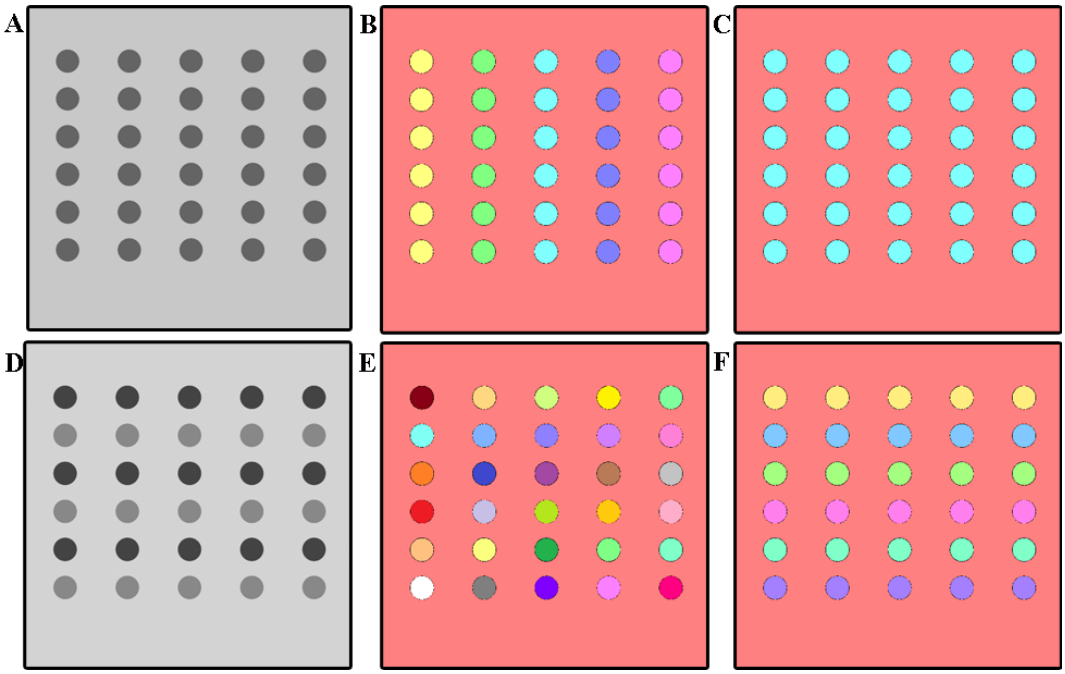

Figure 4.1: Grouping model results on two classical Gestalt dot arrays. Figures (B) and (E) were generated with a σ_s of 20 pixels and a σ_L of 4. Figures (C) and (F) were generated with a σ_s of 30 and a σ_L of 4.

Figure 4.2: Comparison of our grouping model with model of Geisler et al. (2001). (A) A field of oriented elements with a single salient contour. (B) The groups of segments predicted by cooccurence statistics of contours in natural images. (C) The contour groups located by our contour integration model.

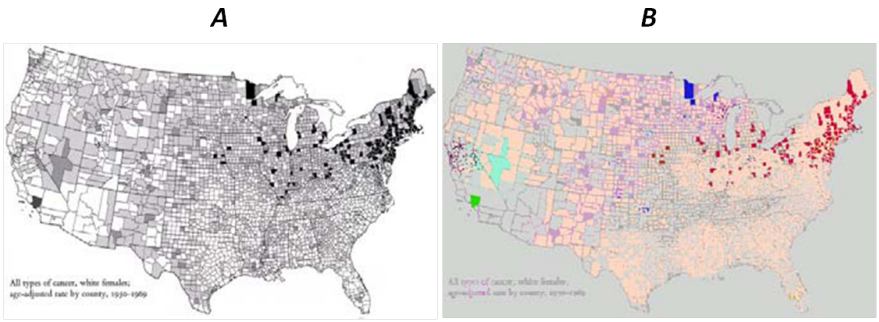

Figure 4.3: (A) An information graphic from Tufte (1983). (B) Our model's analysis of image (A).

southern California. Figure 4.3B shows the results of our luminance similarity and proximity grouping algorithm on this figure; in addition to broad low-rate swaths throughout the country, the model similarly identifies the large cluster of high rate counties in the Northeast, and the two high-rate pockets in northern Minnesota and southern California. This result is quite encouraging: Tufte has pointed out that when presented with this information, the human visual system unconsciously organizes the information in such a way that the cancer clusters pop out. Any model of human perceptual organization should be able to replicate this, without prior information, as ours does.

Figure 4.4 depicts another Tufte demonstration using variations of a plot by Pauling. In the first plot (Figure 4.4A), several families of points can be easily perceived as lying

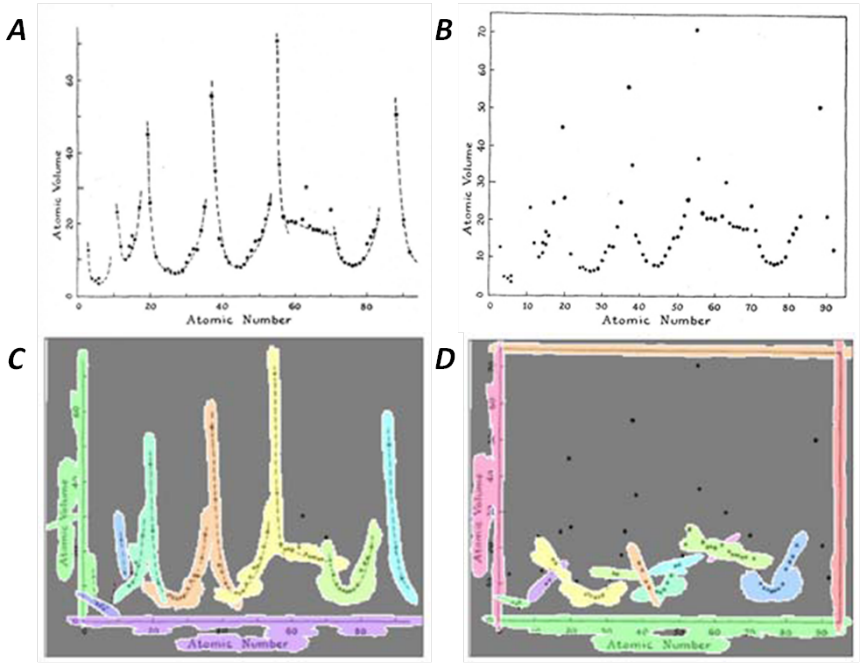

Figure 4.4: (A) A plot by Pauling. (B) Removing the dotted lines makes the image harder to parse. (C,D) Our model agrees with this intuition.

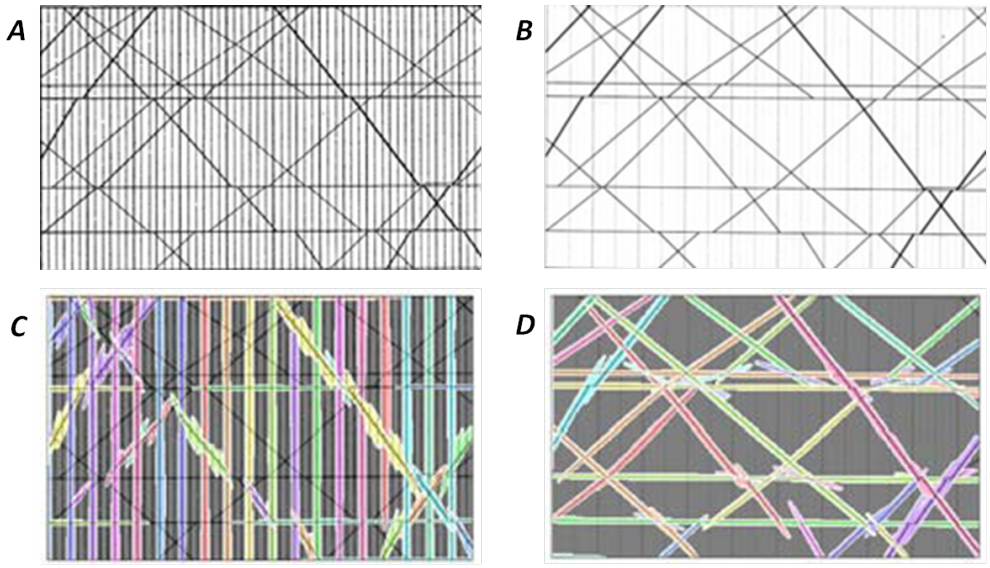

Figure 4.5: (A) Marey's train schedule. (B) Decreasing the contrast of the vertical lines makes the figure easier to understand. (C,D) Again, our model confirms this finding.

along a curved contour, depicted by the dotted lines running through them; in the second (Figure 4.4B), the dotted lines have been removed, and as Tufte points out, it becomes much more difficult to perceive the underlying structure of the plotted points. Figures 4.4C and 4.4D depict the results of our contour integration model on these same figures. The model of Figure 4.4A identifies several groups of points lying along the same contour, and in particular does a good job connecting the sparser dots in the upper part of the plot with the denser dots near the bottom. However, without the dotted lines, the model, just like a human observer, has a much more difficult time organizing the plot; the dots at the bottom are grouped haphazardly, and the sparser dots in the upper part of the figure are left out all together.

A third example from Tufte is shown in Figure 4.5. Figure 4.5A depicts a section of a train schedule by Marey, showing trains running between Paris and Lyon. In this plot, the vertical axis is location between the two cities, and the horizontal axis is time; vertical lines demarcate passing intervals of time (hours), while the diagonal lines indicate individual trains; occasionally a train will stop and wait at a location, which appears as a small horizontal offset in the trains diagonal. In describing this figure, Tufte noted that the dark, high-contrast vertical lines indicating the passing hours make the diagonal lines representing the trains more difficult to parse; he showed that if the contrast of the vertical lines was decreased (Figure 4.5B), the same information could be conveyed without disrupting the perception of the continuous paths of the trains. Once again, the contour integration model confirms these insights; when the first schedule image is passed through the contour integration model, the perception of the diagonals is highly disrupted and often prevented altogether (Figure 4.5C). When the contrast of the verticals is reduced, however, the model is able to identify and parse many of the trains paths, integrating them even across the horizontal offsets corresponding to short stops (Figure 4.5D).

Unfortunately, while these examples are illustrative, they are largely qualitative in nature,

especially the Tufte examples. Though it is encouraging that our model replicates the effects described by Tufte, it is impossible to say *how well* our model is mirroring human behavior in these cases. What is required is a more robust, quantitative method for evaluating the output of our model; in particular, we need a method for comparing our output with more controlled, quantitiative, experimental data.

4.2 From Data to Decision

We are now armed with an intuitive, flexible, and robust model of perceptual grouping which can recreate many classic Gestalt phenomena. To achieve these results, we must hand tune the blurring parameters of the higher-dimensional difference of Gaussians, but this is by no means a major limitation: almost any set of parameters used will give a reasonable, if sometimes uninformative, grouping, and a model with tunable parameters will in many settings be more valuable to a user than one which chooses parameters automatically.

Nevertheless, we know that the human visual system, if it employs a system similar to the one described above, must also employ a mechanism for selecting the appropriate set or sets of parameters. We should therefore seek an intelligent way to filter the output of the overall grouping framework and develop a parameter-free representation of the model results on a given image.

The naïve approach is to simply select a wide array of possible parameter settings and calculate the output of the grouping model for every combination of parameters. Though the visual system likely employs a more adaptive and hence more efficient system, running a wide range of parameter settings will give us the richest possible starting point for our analysis. The problem is that running the model on such a wide range of values yields a very large set of outputs. For the remainder of our discussion of this grouping model, we will assume that the model is grouping by proximity and luminance similarity, and the output of the grouping model for an image I and particular set of parameters σ_s and σ_L is a segmentation of the image pixels, $S(I, \sigma_s, \sigma_L)$; that is, pixels are grouped so that every pixel is in exactly one group. Generating such a large number of segmentations produces an output which is much larger and much more complex than the original input; hardly the direction we want to move in.

What is needed therefore is a way to convert this large set of segmentations to a smaller, better behaved piece of information which can be analyzed numerically and represented in a manageable space so that inferences and predictions can be made with it. The solution we propose is to introduce a set of appropriate hypotheses for explaining the image data; this approach is similar to a task that a human subject might encounter in a psychophysics experiment. For example, suppose a subject is shown an image like that in Figure 4.6A; the subject is told that the image contains a region which is separate from its background, and asked where that grouping might be found: above and left of center, above and right of center, below and left of center, or below and right of center. The subject must evaluate how well these four hypotheses about the structure of the image agree with his or her percept of that image.

In this case, the correct hypothesis is that the region is above and left of center; one way of representing this hypothetical image structure is a simple segmentation of the image into

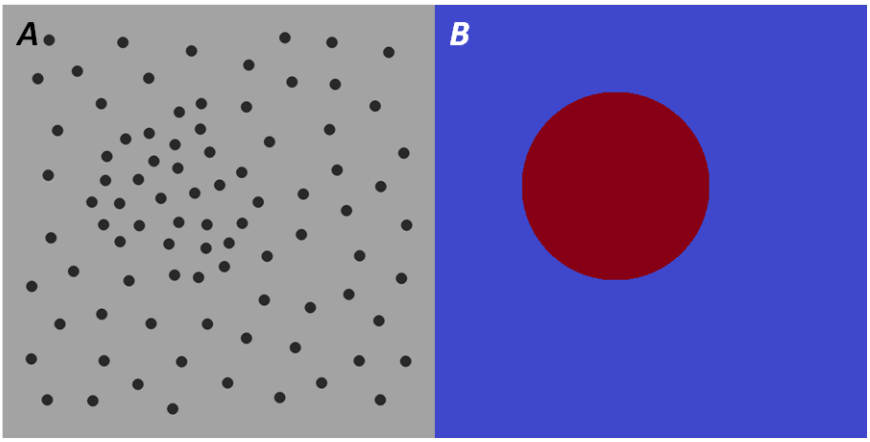

Figure 4.6: (A) A random arrangement of dots with two discernible regions. (B) A simple segmentation representing a hypothesis about the structure of (A).

two regions: the region above and left of center, and the background; this segmentation is shown in Figure 4.6B.

There is in general no way to directly compare an image with a segmentation; but our model takes an image as input and yields a segmentation of the pixels of that image as output. We can thus calculate the degree to which a hypothesis agrees with the perceptual structure of the image by comparing the segmentation representing that hypothesis with the segmentation output by the model. Because we do not know a priori what set of model parameters will yield an informative segmentation, we try a wide range and calculate each segmentation's similarity to the hypothesis segmentation.

Measuring distance between segmentations is by no means a straightforward task; there are many metrics to choose from and each has its own strengths and advantages. For the present study, the Variation of Information metric proposed by Marina Meilă (2007) was used. For two segmentations (or partitions), S_1 and S_2, the variation of information is defined as:

$$VI(S_1, S_2) = H(S_1) + H(S_2) - 2I(S_1, S_2) \tag{4.1}$$

where H and I are entropy and mutual information, respectively. They are defined as:

$$H(S_1) = \sum_i P(s_{1i}) \log P(s_{1i}) \tag{4.2}$$

$$I(S_1, S_2) = \sum_{i,j} P(s_{1i} \cap s_{2j}) \log \frac{P(s_{1i} \cap s_{2j})}{P(s_{1i})P(s_{2j})} \tag{4.3}$$

where $P(s)$ is the probability of a segment s, that is, the area of s divided by the total area of the image; and the s_{1i} and s_{2j} are the individual segments of S_1 and S_2 respectively. Many other metrics have been suggested as measures of differences between set partitions (see Meilă (2007) for a full review), but few are conceptually suited to image segments; the Variation of Information, on the other hand, is built around the sizes and intersections of

segments. Two measures of disagreement between segmentations were proposed by Martin et al. (2001), global consistency error and local consistency error:

$$LCE(S_1, S_2) = \frac{1}{A} \sum_p \min \left(E(S_1, S_2, p), E(S_2, S_1, p) \right) \tag{4.4}$$

$$GCE(S_1, S_2) = \frac{1}{A} \min \left(\sum_p E(S_1, S_2, p), \sum_p E(S_2, S_1, p) \right) \tag{4.5}$$

where A is the area of the silhouette, and E is defined as:

$$E(S_1, S_2, p) = \frac{|R(S_1, p) \setminus R(S_2, p)|}{|R(S_1, p)|}$$

Unfortunately, these measures are designed to give zero error to refinements (that is, if S_1 is a refinement of S_2, then $LCE(S_1, S_2)$ and $GCE(S_1, S_2)$ will both be zero), which will not work for our purposes.

There is also something intuitively appealing about this metric. Entropy and mutual information are easy concepts to understand, and computationally simple to implement. Researchers in both computer vision and human vision are familiar with these concepts; and as we are trying to determine how well a particular hypothesis (represented by our hypothesis segmentation) explains the perceptual structure of an image (represented by the models output segmentation), using a measure based on the mutual information between those two segmentations seems entirely appropriate.

So, let us suppose we are given the image and hypothesis shown in Figure 4.6A. To measure the degree to which this image agrees with this hypothesis, we run our grouping model on the image at a wide range of blurring parameters; this yields a large number of segmentations of the image pixels. Several segmentations are depicted in Figure 4.7. For each of these segmentations, we then measure the Variation of Information between that segmentation (output by the grouping model) and the hypothesis segmentation (Figure 4.6B). This yields an array of values which is illustrated in Figure 4.8. Figure 4.8 also shows the same array

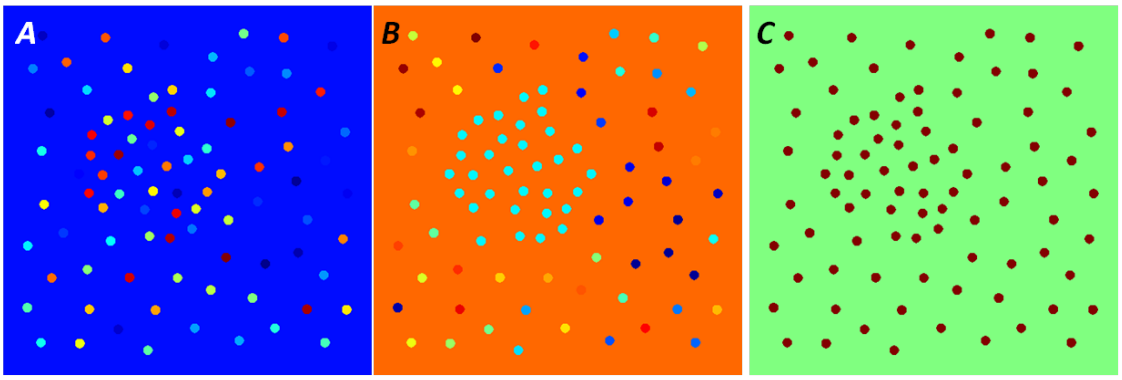

Figure 4.7: (A) Segmentation of image 4.6 with a σ_s of 5 pixels and a σ_L of 4. (B) Segmentation with a σ_s of 10 pixels and a σ_L of 4. (C) Segmentation with a σ_s of 20 pixels and a σ_L of 4.

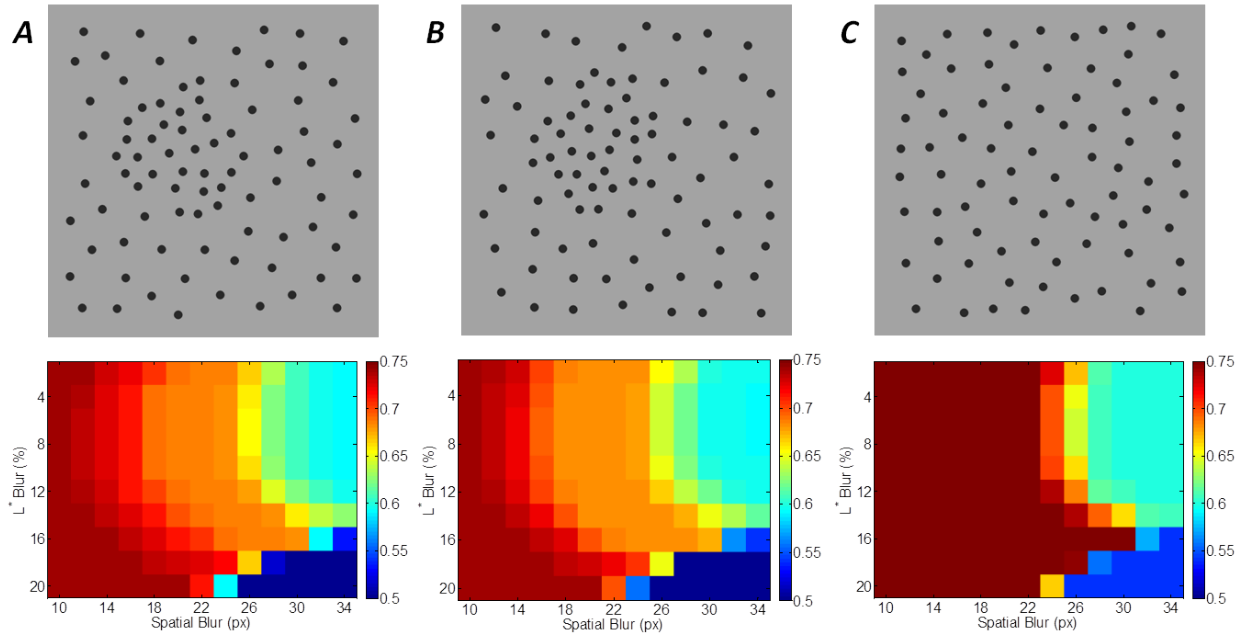

Figure 4.8: Results of grouping model and hypothesis comparison on several images.

of values for two other images: one with a very similar image structure to Figure 4.6A, and one without this image structure.

These figures contain a great deal of information, but they can be fairly intuitively understood. For all three images, when the spatial blurring parameter is very small, the image is segmented into the background and each individual dot; the locations of these individual dots is information that is not present in the hypothesis segmentation, so the VI at these parameter settings is relatively high. Conversely, when the spatial blurring parameter and luminance blurring parameter are both large, all the pixels in the image group together; there is almost no information in this model output segmentation, so the VI is quite low. But when the spatial blurring parameter is right in the middle (around 20 pixels) and the luminance blurring parameter is low enough that the foreground and background are separate, the single segment corresponding to the cluster of dots in the model output segmentation aligns very well with the hypothesis segmentation, increasing the mutual information between the two and lowering the VI relative to a similar image with no such cluster.

In short, this process - calculating an array of segmentations at different scales and comparing with a hypothesis segmentation - yields an array of values which is closely linked to perceptual structure in images. Images with similar perceptual structures will yield similar arrays, and images with different perceptual structures will yield different arrays if one chooses an appropriate hypothesis. With this method in hand, we can begin to analyze the performance of our model with respect to actual human behavior.

4.3 Experiment 1: The Effects of Proximity and Luminance

4.3.1 Methods

Subjects performed a 2-alternative forced choice task. In each trial, they were presented with two images containing random fields of dots; they were told that "one image contains a region such the dots inside the region are different from the dots outside the region" while the other image "contains no organization at all." Subjects saw each image for 200 ms, with a 400 ms ISI. After viewing both images, subjects were asked which image they believed contained the region.

Subjects viewed 10 practice trials in which the region in question was clearly discernible as a result of proximity or luminance. They then viewed 300 test trials in 6 blocks of 50 trials.

4.3.2 Subjects

The experiment was run on 19 subjects from the Boston area. There were 13 male subjects and 6 female subjects. Ages ranged from 19 to 57, with a median age of 33.

4.3.3 Stimuli

In each pair, one image, referred to as the target image, was generated with an off-center circular region such that the scene parameters of the dots inside the region were different from those outside the region. The other image, referred to as the distractor image, contained dots generated with uniform proximity such the number of dots was on average the same as the first image. The luminances of the dots in the first image were then randomly permuted and assigned to the dots in the second image. This generated two images with approximately the same number of dots and same overall distribution of luminances.

The specific procedure for placing dots is described in Appendix A. In any region of the image the proximity in that region was represented by distance parameter d; dots were placed such that any dot placed in an area with distance parameter d must be at least d pixels away from any other dot. Luminance values correspond to the L^* value of the L^*ab colorspace, and range from 0 to 100, where 100 is the maximum brightness of the experimental monitor.

On any given trial, the differences between the interior of the region and the exterior of the region in the target image fell into one of four categories:

1. Proximity Alone (PA): Dots inside the region were generated with a different proximity parameter than the dots outside the region. The minimum distance inside the region and outside the region were varied so that the total number of dots was similar across all trials. The ratio of the inner region minimum distance to the outer region minimum distance (which shall be referred to as proximity ratio) varied from 0.59 to 1.43.

   In all PA trials, the luminance (L^*) of all dots was 40, while the luminance of the background was 80.

2. Luminance Alone (LA): Dots inside the region had a different luminance from dots outside the region. The difference in luminance between the interior and exterior dots ranged from 6.67 to 26.67.

   In 40% of LA trials, the luminance of all dots was darker than the background; specifically, the luminance of the background was 80 while the luminances of interior and exterior dots were equidistant from 40. In another 40% of LA trials, the luminance of all dots was lighter than the background; specifically, the luminance of the background was 20 while the luminances of interior and exterior dots were equidistant from 60. In the final 20% of LA trials, the background luminance was set to 50 and the luminances of the interior and exterior dots were equidistant from 50.

   In all LA trials, the proximity parameter of the dots was the same throughout the image.

3. Proximity and Luminance in Concert (PLC): Dots inside the region were generated with a different proximity parameter and had a different luminance from those outside the region. The proximity ratio ranged from .75 to 1.10, and the difference in luminance was either 6.67 or 13.33.

   In two thirds of PLC trials, the luminance of all dots was darker than the background; specifically, the luminance of the background was 80 while the luminances of interior and exterior dots were equidistant from 40. In the other third of PLC trials, the luminance of all dots was lighter than the background; specifically, the luminance of the background was 20 while the luminances of interior and exterior dots were equidistant from 60.

4. Proximity with Luminance Distracting (PLD): Dots inside the region were generated with a different proximity parameter than dots outside the region, while luminances of all dots both inside and outside the region were assigned one of two different values with equal probability. The proximity ratio ranged from .75 to 1.43, and the luminance difference ranged from 6.67 to 20.

   In one third of PLD trials, the luminance of all dots was darker than the background; specifically, the luminance of the background was 80 while two luminance values were equidistant from 40. In another third of PLD trials, the luminance of all dots was lighter than the background; specifically, the luminance of the background was 20 while the two luminances value were equidistant from 60. In the final third of PLD trials, the background luminance was set to 50 and the two luminance values were equidistant from 50.

Examples of image pairs from the four categories of trial are shown in Figure 4.9.

One important note: though the dot images used in this experiment were randomly generated, and the order of trials in the experiment was randomized for each subject, each subject saw the same 300 images. Though this increases the noise in the overall result - human behavior for a given set of scene parameters will be more strongly dependent on the nature of individual images, as each of those individual images will appear multiple times - it allows us to measure not only the performance of human subjects on certain scene

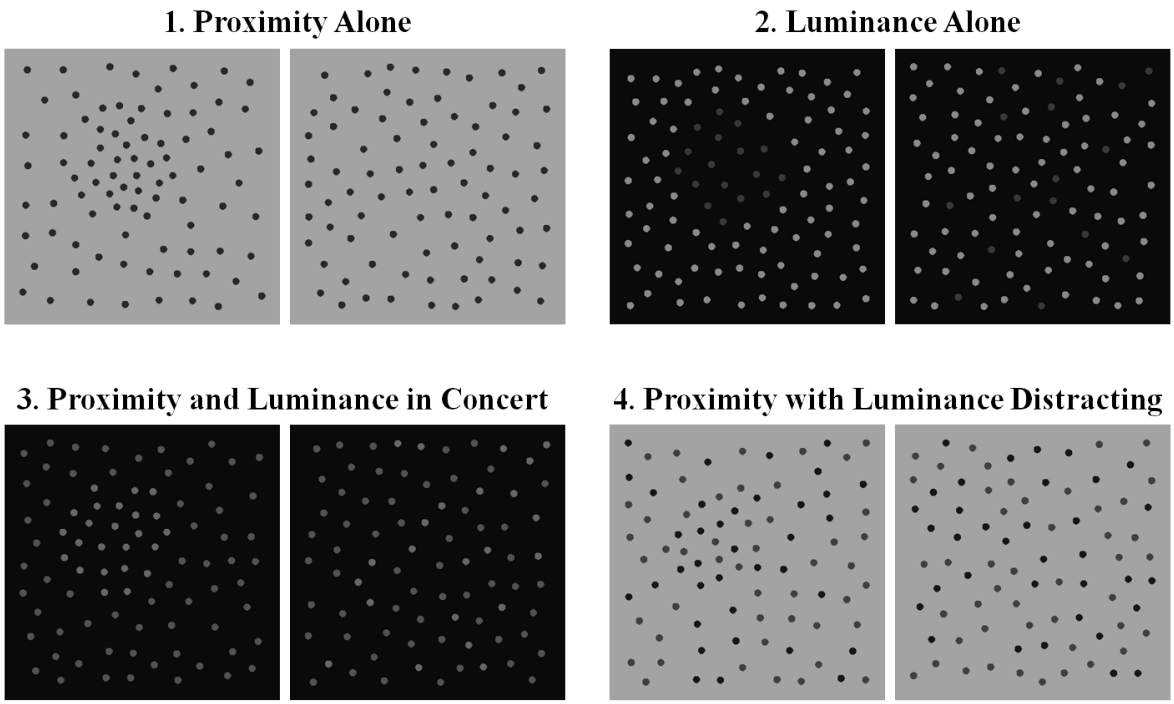

Figure 4.9: Sample image pairs from Experiment 1. In each pair, the target image is on the left and the distractor image is on the right. The target images have been generated with a region just above and left of center.

parameters, but also on specific images themselves. We can thus determine whether our model explains any variation of human performance that is not dependent on the underlying scene parameters, but on the particular properties of an individual image.

4.3.4 Experimental Results

Experimental results for proximity alone trials are shown in Figure 4.10. In all figures, unless otherwise noted, *subject preference* refers the proportion of trials in which subjects picked the target image in an image pair rather than the distractor image; a subject preference of 0.5 represents chance. The results on the PA trials are very much what one would expect: when the proximity ratio is low (that is, when the dots inside the region are much denser than the dots outside) subject preference is very high. As the proximity ratio approaches 1 and the densities inside the region and outside the region approach equality, subject preference approaches chance. And finally, when the proximity ratio is high, with a lower density inside the region than outside, the subject preference rises again, though not as high as for the conditions with a low proximity ratio.

The results for luminance alone trials are shown in Figure 4.11; again, the results match well with intuition. As the difference in luminance between the dots inside the region and the dots outside the region increases, the subject preference also increases, with very small luminance differences yield performance close to chance. One surprising result: when the dots inside and outside the region have opposite polarity (that is, when the dots inside the

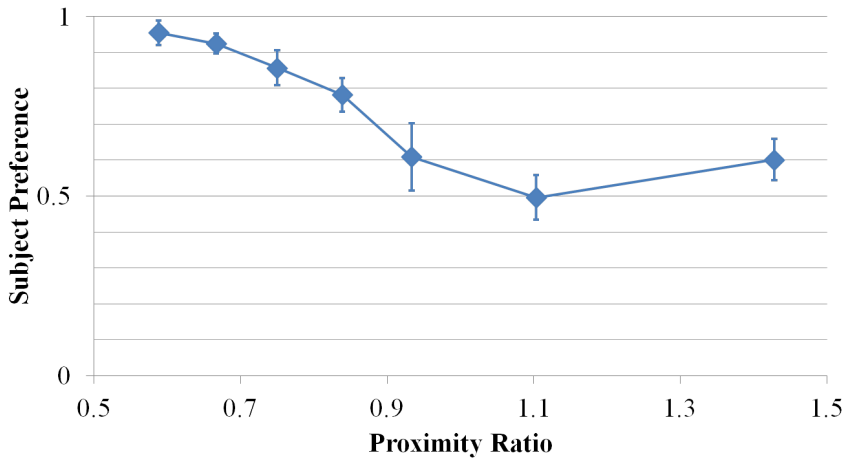

Figure 4.10: Results from Experiment 1 on proximity alone trials. Proximity ratio is equal to d_{in}/d_{out} where d_{in} is the minimum distance between dots inside the region, and d_out is the minimum distance between dots outside the region; subject preference is the proportion of trials in which subjects believed the target image contained a region different from its background.

region are darker than the background and the dots inside the region are lighter than the background, or vice versa) the subject preference is very high. Subjects are very good at detecting groups defined by opposite polarity, far better than they are at detecting group defined by equal luminance difference but the same polarity.

The results for proximity and luminance in concert (PLC) trials are shown in Figures 4.12 and 4.13. In these plots, the blue curves represent the subject preferences on images with no luminance difference between the dots inside the region and the dots outside the region. The red curves represent trials in which a small luminance difference of 6.7 was introduced between the interior and exterior dots, and the green curve represents an additional set of

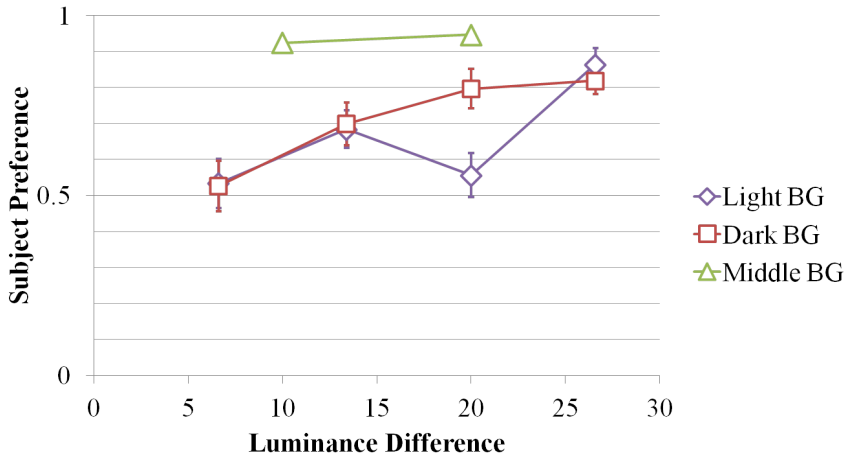

Figure 4.11: Results from Experiment 1 on luminance alone trials.

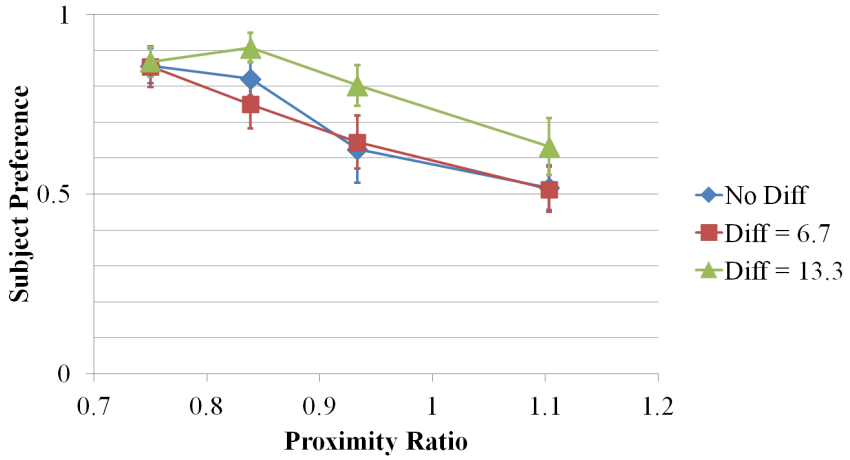

Figure 4.12: Results from Experiment 1 on PLC trials with light backgrounds.

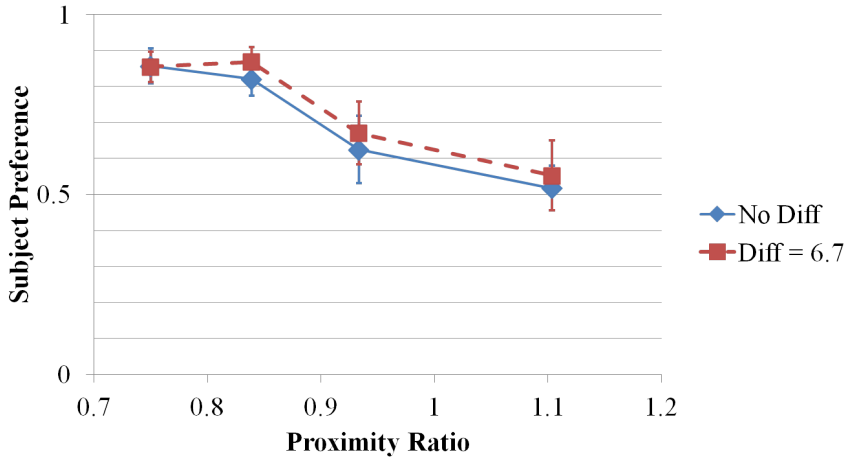

Figure 4.13: Results from Experiment 1 on PLC trials with dark backgrounds.

trials in which a larger difference of 13.3 was introduced. The results have been separated into trials in which all dots are darker than the background and trials in which all dots are lighter than the background. Though the results seem to agree with intuition - groupings defined by proximity and luminance together are easier to detect than groupings defined by proximity alone, particularly for the larger luminance difference - the effects are by no means strong; many differences lie within the range of standard error. The results are suggestive, but it would be difficult to make any strong inferences about the effect, and it is unlikely that our model will successfully mirror it

Finally, the results of the proximity with luminance distracting (PLD) trials are shown in Figure 4.14, 4.15 and 4.16. Again, blue curves represent trials in which all dots have the same luminance; the red curves represent trials in which dots are randomly assigned one of two luminances that differ by a small value (6.7 or 10), and green curves represent trials in which the dot luminance differ by a larger value (20). The results have been separated into trials where all dots are darker than the background, trials where all dots are lighter than

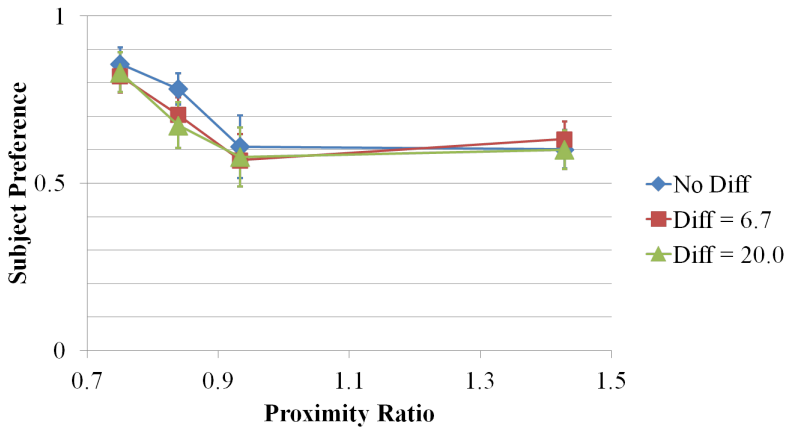

Figure 4.14: Results from Experiment 1 on PLD trials with light backgrounds.

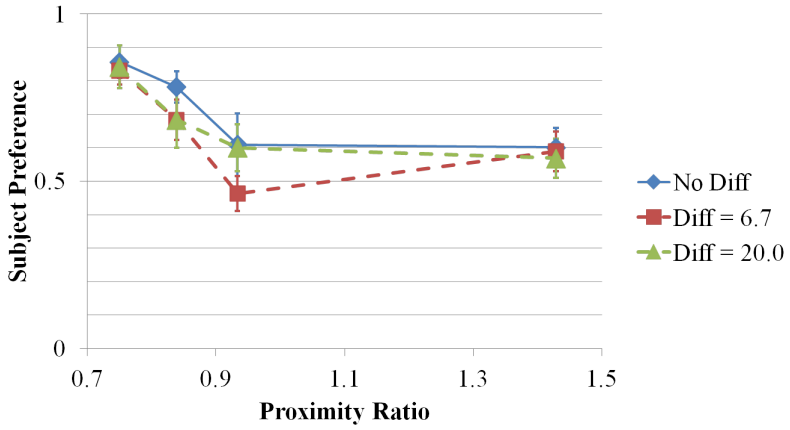

Figure 4.15: Results from Experiment 1 on PLD trials with dark backgrounds.

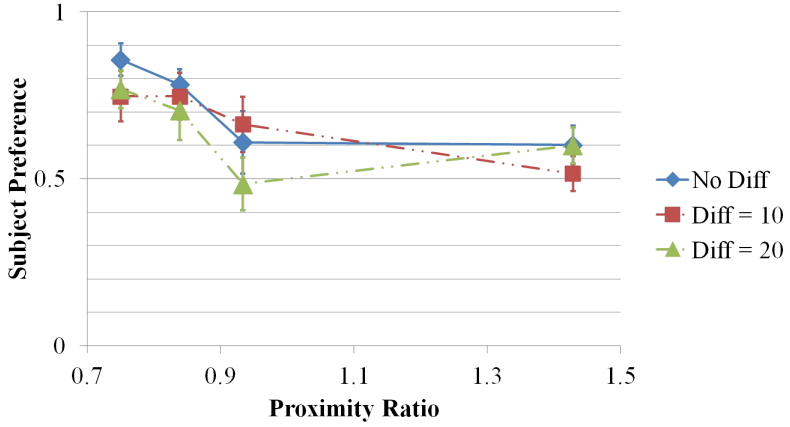

Figure 4.16: Results from Experiment 1 on PLD trials with central backgrounds (between the two dot luminances).

the background, and trails where the dot luminances lie on either side of the background luminance. The results here are similar to those in the PLC trials; they seem to agree with the intuition that the presence of varying luminance makes proximity groupings more difficult to effectively separate from their backgrounds, but the effects again are quite small, and not particularly reliable. Interestingly, the effect seems not to be present for very high proximity ratios, but this makes intuitive sense. The images with high proximity ratios have regions defined by a sparseness of dots; detecting sparseness would be largely unaffected by the variability of the few dots present in the array. Still, it is something that a model of perceptual grouping should consider.

4.3.5 Model Results

To evaluate the performance of our model in comparison to the human behavior in the above experiment, we began by generating a large library of image pairs generated using the same range of scene parameters as were used in the experiment. Both images in each of these image pairs were run through the grouping model over a large range of model parameter settings, and the resulting segmentations were compared with four structural hypotheses like the one in Figure 4.6B. Thus every image pair in the library was converted to a full response array; these arrays were then mapped using principal components analysis into a much more manageable space (20-25 dimensions). In this space, images generated by the same scene parameters and same hypothesis map to tight clusters of points, and the set of all images generated by one hypothesis forms an intricate distribution of points that can easily be modeled as a mixture of Gaussian distributions generated by each of the individual scene parameter settings.

So, with a library of image pairs generated by the same hypothesis and their resulting response arrays, we can estimate the likelihood of a new image pair given that same hypothesis by calculating the new response array, and determining its likelihood in the response array distribution. Doing this for all eight region locations (four possible region locations in both images in the image pair) yields the likelihood of a new image pair given each of the eight structural hypotheses. Thanks to Bayes' Rule, if we assume that all eight hypotheses are equally likely, then the probability of a hypothesis H given an image pair A is proportional to the probability of A given H. Thus calculating the estimated relative likelihood of an image pair for all eight hypotheses gives us the final calculation of the estimated probability that the region is in the first or second image.

We'll begin with only those images in which grouping is defined by proximity. These are the images in which the model performs the most consistently, making them an ideal testing ground for our model evaluation. The images in the experiment were generated at 7 different proximity ratios; using a sample distribution of images generated with those same 7 parameter settings, Figure 4.17 shows the average likelihood index of images generated with those parameter settings. Likelihood index is a quantity defined as:

$$LI = \log \frac{P(I_1)}{P(I_2)} \tag{4.6}$$

where $P(I_1)$ is the estimated probability that the region is in the first image, and $P(I_2)$ is the estimated probability that the region is in the second image. In general, the estimated

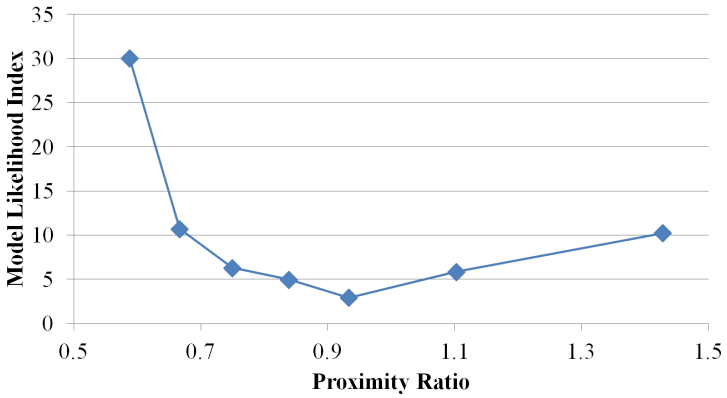

Figure 4.17: Model prediction for proximity alone trials.

probabilities are very close to 0 or 1, so the likelihood index gives a richer sense of the relative confidence of the model about different images.

We can see here that the confidence of the model regarding image pairs generated by the different proximity parameter settings closely mirrors the qualitative pattern of preference shown in the results of the experiment. Indeed, the model, like human subjects, is less confident in identifying regions of lower density, thought the asymmetry in the model's confidence is not as pronounced. Figure 4.18 shows the relationship between the subject preferences on individual image pairs and the models likelihood index for those pairs. The figure also includes the reflection of the subject preferences and likelihood indices, to show the subject preferences and model predictions for image pairs in which the target is the second image; therefore, in this plot, *subject preference* refers to the proportion of trials in which subjects selected the *first* image in the image pair, rather than the second. Though

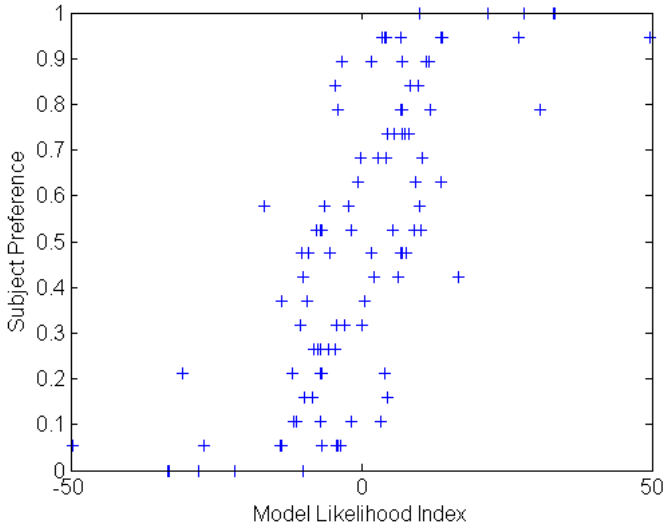

Figure 4.18: Model prediction for individual image pairs in proximity alone trials. Model data and subject data are correlated with $r = 0.7177$.

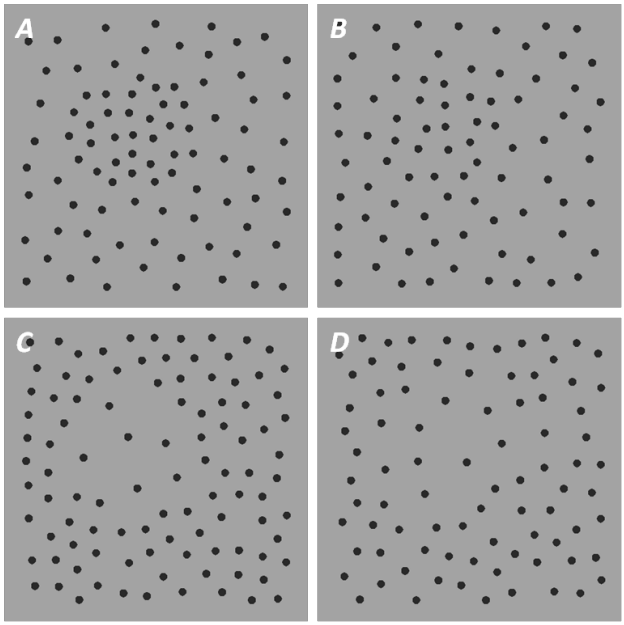

Figure 4.19: (A) Image with proximity ratio of 2/3 judged by the model to be the most easily grouped. (B) Image with same generating parameters judgedto be the least easily grouped. (C) Image with proximity ratio of 1.4 judged to be the most easily grouped. (D) Image with the same generating parameters judged to be the least easily grouped.

including these reflected points doubles the data, humans were presented with the image pairs in both orders, so this plot gives a fuller sense of the behavior of humans and the model across all possible image pairs.

While the results here are not as clean as we might like, there is nevertheless a clearly discernible upward trend, suggesting the model is able to capture some of what drives human grouping judgments. The correlation between the likelihood index and subject preference for the points shown here is 0.7177, a highly statisically significant correlation ($p < 0.001$); if we do not mirror the data points but only consider image pairs where the target image came first, the correlation is weaker at 0.3872, but still statistically significant ($p < 0.01$).

One might now ask: does the model capture the variability in human grouping judgment driven by individual images? Or is predictive power strictly tied to the underlying scene parameters, and unaffected by the variability from image to image? To address this, consider Figure 4.19. Figures 4.19A and 4.19B were both generated with a proximity ratio of 2/3, but despite the similar underlying parameters, the two images came out very differently. In Figure 4.19A the spacing of the dots inside the region is quite regular, and the transition from high density to low density is easily perceived. In Figure 4.19B, on the other hand, the dots have randomly landed further apart, and the transition from the higher density inside the region to the lower density outside the region is much more difficult to perceive. Of all images generated with a proximity ratio of 2/3, Figure 4.19A was the image the model was most confident contained a group; while Figure 4.19B was the image the model was least confident contained a group. The same is true of Figures 4.19C and 4.19D, except that these two were generated with a proximity ratio of 1.4. Once again, the image judged more

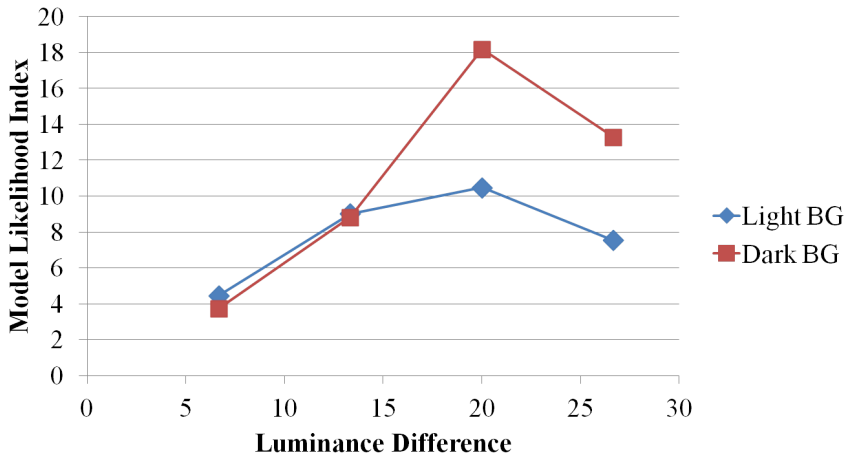

Figure 4.20: Model prediction for luminance alone trials.

confidently by the model is easier to parse, with a clearly discernible region of low density and a sharp transition; conversely, in the image judged least confidently by the model, the region of low density is poorly localized, and the transition is almost imperceptible. Clearly the model is capturing the grouping strength of images beyond the broad influence of the underlying scene parameters.

Figure 4.20 shows the predictions of the model when operating only on images in which groupings are defined by luminance; once again the model successfully recreates the qualitative pattern of discriminability, though the shapes are not precisely the same, and the model is far more confident about images with darker backgrounds. Figure 4.21 shows the relationship between the model prediction and the subject preference for individual images.

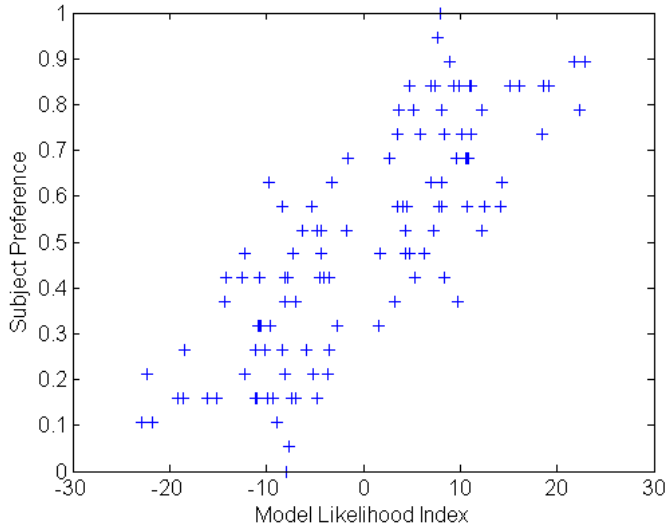

Figure 4.21: Model prediction for individual image pairs in luminance alone trials. Data are correlated with $r = 0.798$.

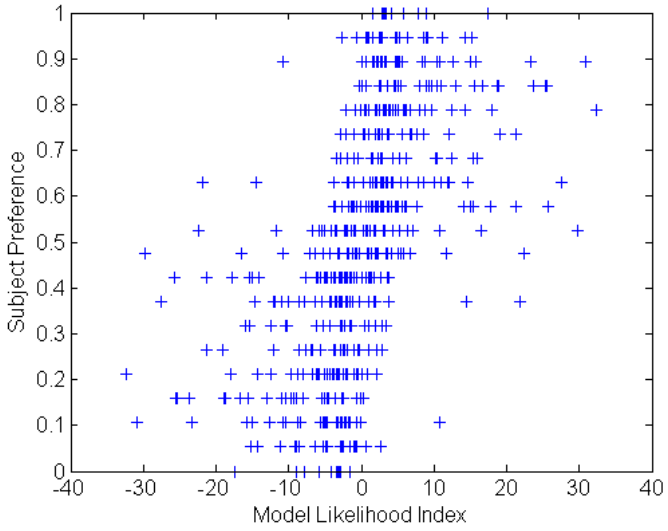

Figure 4.22: Model prediction for individual image pairs in all trials calculated together. Data are correlated with $r = 0.479$.

The relationship here is even stronger than with the PA trials, with a correlation of 0.798, statistically significant with $p < 0.001$, driven once again by the models ability to respond to individual variations between images; the correlation for unmirrored data is 0.4571, also significant ($p < 0.01$).

Finally, Figure 4.22 plots the predictions of the model based on the complete family of generating parameters. Clearly, the model has a bit more difficulty closely following human behavior in this case. Nevertheless, a clear relationship between the model output and the subject preference can be perceived, and manifests itself as a correlation of 0.479, which is statistically signficant ($p < 0.001$); unfortunately, the correlation for the unmirrored data is only 0.1392, which is much more weakly statistically significant ($p < 0.05$).

These results are encouraging, but many gaps remain. One vexing point in particular is the influence of the image background; human observers seem to be able to discount it almost entirely and group around it quite easily, but it can have very disruptive effect on the model's performance. In addition, it is clear from the results of the experiment - particularly the luminance alone trials - that the further the luminances of the dots are from the background luminance, the less salient their differences become. One possible method for modeling this is to find a way to subtract out the image background, and pass the resulting signal through a compressive non-linearity, which enhances differences near the background, and suppresses differences which are further from the background. Such a point-nonlinearity could easily have an analogue in the visual cortex; and the result of this non-linearity could easily be passed into our model, just as luminance was here.

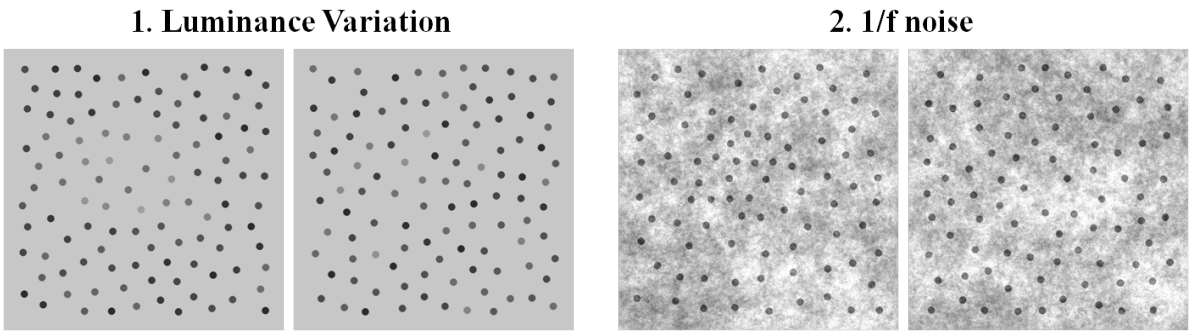

Figure 4.23: Example image pairs under the two types of disruption used in Experiment 2.

4.4 Experiment 2: The Influence of Noise

4.4.1 Methods

Methods used were identical to those used in Experiment 1.

4.4.2 Subjects

The experiment was run using 11 subject from the Boston area. There were 7 male subjects and 4 female subjects. Ages ranged from 21 to 56, with a median age of 40. An additional 4 subjects were excluded for failing to give consistent responses on test trials.

4.4.3 Stimuli

300 random dot image pairs were generated as in Experiment 1, using 4 scene parameter settings. Two of these settings generated images as in PA trials; one setting with dots darker than their background and one with dots lighter than their background. The other two settings generated images as in LA trials; again, one setting with dots darker than their background and one with dots lighter than their background. The magnitude of the proximity and luminance differences in these images was chosen based on the results of experiment 1 such that baseline performance on similar image pairs was above chance but below ceiling.

Of these 300 image pairs, 76 were left unaltered. The remaining 224 image pairs were passed through one of two disruptions:

1. Luminance Variation: (96 image pairs) Under this disruption, the luminance of each dot was perturbed from its original value; these perturbations were drawn from a normal distribution. The strength of this disruption - measured by the standard deviation of the perturbation distribution - varied from 2.5 to 10.

2. $1/f$ Noise: (128 image pairs) Under this disruption, randomly generated pink noise, or $1/f$ noise, was added to both images in the image pair. The strength of this disruption - measured by the standard deviation of the overall noise - also ranged from 2.5 to 10.

Example image pairs after both types of disruption are shown in Figure 4.23.

4.4.4 Experimental Results

The results of Experiment 2 are shown in Figures 4.24 through 4.27. Overall, perhaps the most striking result is how stable the preferences are; in only one of the conditions - luminance groups on a dark background disrupted by 1/f noise - is there any clear drop off in performance as the noise increases. Despite disruptions of considerable noise, human grouping preference seems surprisingly robust, especially to variations in individual dot luminances.

Of course, there may be some other factors at work here; take the image in Figure 4.28. The disruption in this image has made the boundary between the interior region, defined by brighter luminance, and the exterior region, defined by darker luminance, almost imperceptible. But nonetheless, a savvy subject could likely pick up that the average luminance in the upper right region is higher than that of the remainder of the image. Thus, while

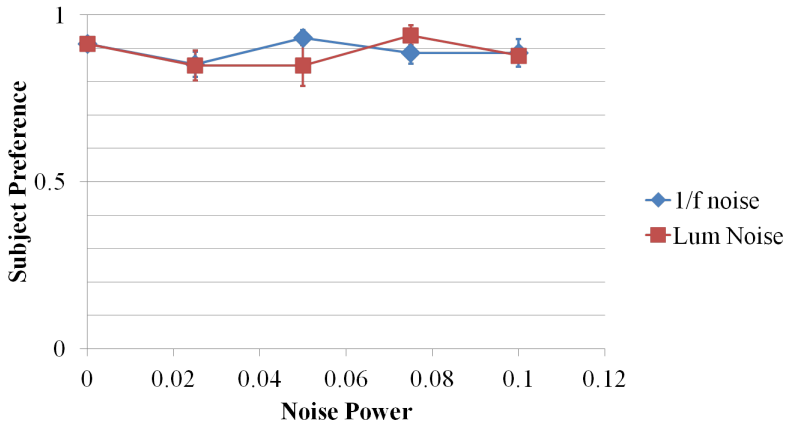

Figure 4.24: Results from Experiment 2; influence of noise on detection of proximity-defined groups on light backgrounds.

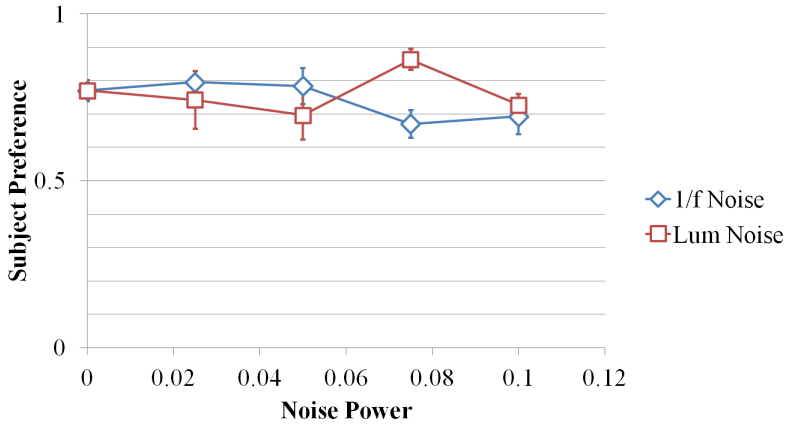

Figure 4.25: Results from Experiment 2; influence of noise on detection of proximity-defined groups on dark backgrounds.

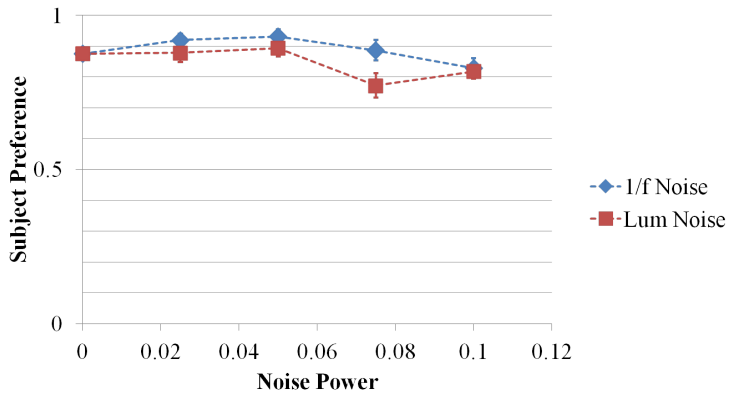

Figure 4.26: Results from Experiment 2; influence of noise on detection of proximity-defined groups on light backgrounds.

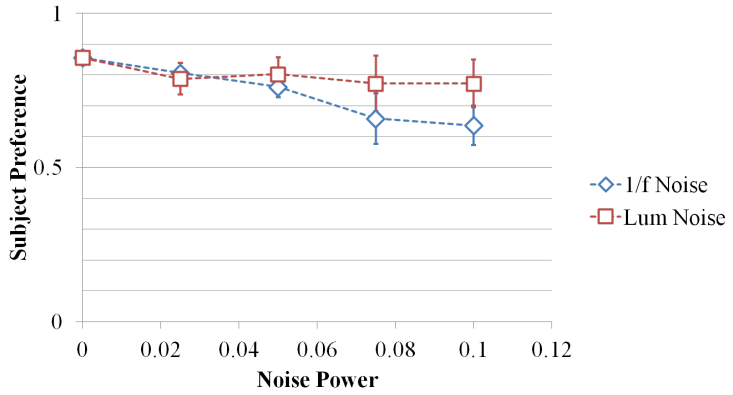

Figure 4.27: Results from Experiment 2; influence of noise on detection of proximity-defined groups on dark backgrounds.

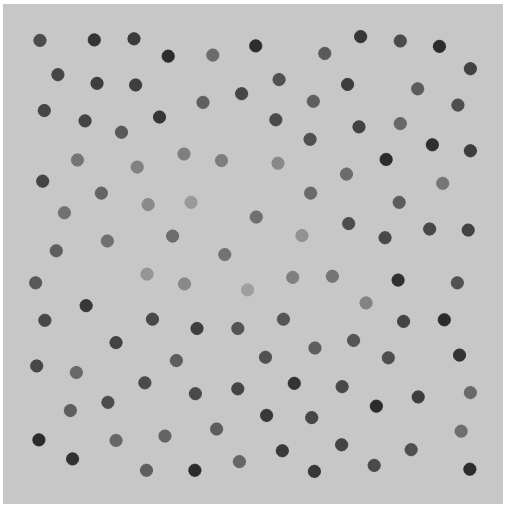

Figure 4.28: An image with large variance added to the dot luminances. A group or region is no longer visible, but its presence may still be inferred from local statistics.

the subject could not be said to perform any segmentation or grouping, he or she might still be able to perform the task with relatively high accuracy. This blurred line between true grouping and simple local statistical measurement is an important factor to consider in future experiments.

4.4.5 Model Results

Based on the findings of Experiment 1, for these images, the model was fed not the raw luminance of the original image, but the background subtracted and compressed signal which enhances differences closer to the background. For images which did not have a constant background, a simple local median filter larger than twice the size of one dot was used to represent the "local" background. The predictions of the model for the various scene and noise parameters in proximity grouping trials are shown in Figures 4.29 and 4.30.

The model seems to have had considerably more trouble with these images; the overall likelihood indices are much lower than many of the images in previous conditions. But like the subjects, the model seemed largely unaffected by the magnitude of noise; indeed, as the

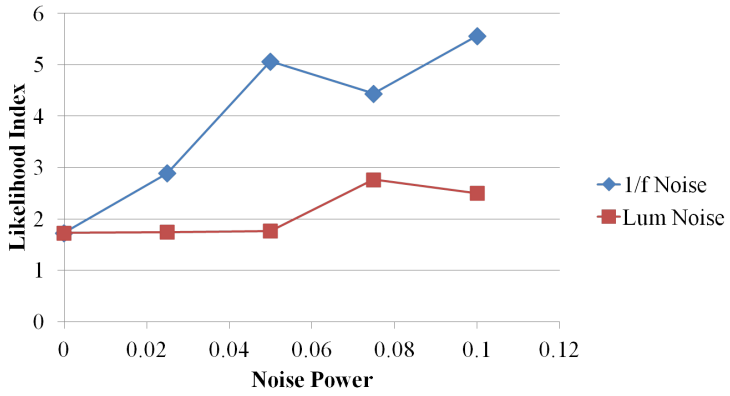

Figure 4.29: Model predictions of the influence of noise on detection of proximity-defined groups on light backgrounds.

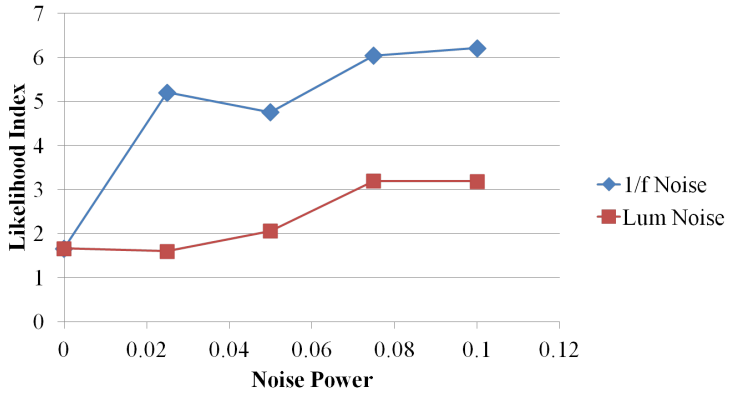

Figure 4.30: Model predictions of the influence of noise on detection of proximity-defined groups on dark backgrounds.

1/f noise magnitude increased, the model actually performed better. It is not clear why this is the case; it is possible that the 1/f noise in the background made the background segment less coherent, and reduced interfering effects that the background pixels had on the foreground pixels in x-y-L^* space. The plots of the individual image predictions and corresponding subject preferences is shown in Figure 4.31; again, the performance was considerably lower. This is perhaps driven by the fact that subject performance on these images was so consistently high. There are relatively few images whose preference rates lie at or near chance; thus the model has little variablilty to predict.

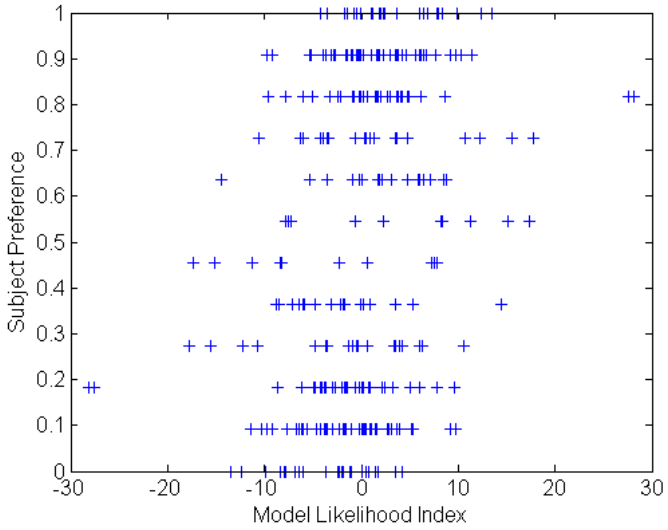

Figure 4.31: Model prediction for individual image pairs in all noise trials calculated together. Data are correlated with $r = 0.3157$.

Chapter 5

Future Work

5.1 Experimental Variations

The experimental protocol described in section 4.3 far from exhausts its potential in the experimental results shown here. One possible variation on the protocol is a task in which subjects are shown one image rather than two. In each trial, the subject would be asked where in the image a group is present; e.g. the group might be in the left half or the right half of the image.

In such an experiment, the majority of images would contain only one grouping; but the advantage of the single-image protocol rather than the serial two-image protocol is that in some fraction of images, one can have two groupings present in the same image; by asking the subject which side or region of the image contained a group, one can probe which of the two groupings is more visible. This allows one to evaluate the relative strength or influence of different factors, such as proximity vs. luminance similarity. But unlike earlier studies analyzing the interaction and competition of different grouping factors, the use of randomly generated grouping images rather than organized grids of elements creates a richer, broader set of grouping stimuli and forces a grouping model to operate in a more generic, image-like domain.

As described in the appendix, the method by which the random dot arrays are constructed is a highly constrained random process; one result of this is that, in most cases, different scene parameters generate completely non-overlapping ranges of images. Thus, an ideal observer with complete knowledge of the image-generation process would be able to distinguish images generated by different scene parameters with 100% accuracy. A more informative approach would generate the images more probabilistically, such that any given image has non-zero probability of being generated by several distinct sets of scene parameters; this would allow statistically analysis of the ideal observer's performance on the grouping detection task, which could serve as a point of comparison for human and model performance. However, as we saw in the results of experiment 2, introducing noise into the scene generation process can also blur the distinction between region grouping and simple measurement of local statistics; so any such changes to the dot array image-generation process would need to be done carefully.

5.2 Variation of the Model

The two primary parameters investigated in the above experiments, σ_s and σ_L, encompass much of the intuitive behavior of the model; but these are not the only parameter choices present in the grouping model. For example, the model performs its filtering with a difference-of-Gaussians; this DoG filter consists of a multi-dimensional Gaussian of integral 1, and subtracted from it another multi-dimensional Gaussian of integral 1, but with standard deviation equal to 1.5 times that of the first Gaussian. This family of filtering kernels has performed well enough for us thus far, but it is possible that different kernels would improve performance even further.

For example, one could increase the ratio of standard deviations to 2 or more; this would broaden the area of space across which the strong presence of one feature suppresses other nearby, similar features. One could also alter the filter to have an integral other than 0; letting the positive Gaussian have a larger integral than the negative Gaussian would cause the filter to give a positive response in large, flat regions of high density. This would improve the model's ability to detect large flat regions and reduce the number of unclassified pixels; it might also, however, increase the models tendency to undersegment.

Though changing these aspects of the DoG filter would likely have a measurable impact on the behavior of the model, we have thus far spent little time experimenting with these changes. Without a means to quantitatively evaluate the overall performance of the model, there has been no way to state that one form of the model is better than any other; so we have continued to use the simplest form of the model. Given a robust method for evaluating model performance, such as the one described in section 4.2, choice of filtering kernel should be one of the first questions investigated.

An even wider variety of filtering options are available in the domain of contour integration. Our current approach utilizes a relatively simple anisotropic difference of Gaussians (the parameters of which could also be varied and investigated); until now, we have made use of the anisotropic Gaussian primarily for reasons of computability. But psychophysical work on contour integration by Field et al. (1993) and investigation of natural image statistics by Geisler et al. (2001) suggest a much more complex kernel may be appropriate, one that integrates considerations of alignment and co-circularity. Of course, even without such complications, we have seen that our model is able to recreate some of the success of these more complex formulations (Figure 4.2); and recent work by Watt et al. (2008) has shown that contour integration behavior ascribed to complex association models can be explained by much simpler models using anisotropic linking along the direction of local orientation. Therefore it is possible that introducing a more complex filtering kernel will improve performance only slightly.

One limitation of our model is that it is too accepting of abrupt changes in contour curvature. Psychophysics show that humans can link adjacent contour elements even with a large change in orientation, but contours which zigzag back and forth between orientations are almost undetectable (Feldman, 1997; Sigman et al., 2001; Ledgeway et al., 2005); this suggests that subjects can accept a relatively large curvature in a contour, but that the contour must be relatively constant. Indeed, several models of contour integration make explicit use of local curvature measurements to determine strength or likelihood of grouping (Parent

and Zucker, 1989; Gigus and Malik, 1991), though it seems that enforcing co-circularity is not necessarily the optimal choice ore the one made by humans (Singh and Fulvio, 2005; **?**). One possible solution would be to add an additional dimension to the contour space, locating contours in x-y-θ-κ space, where κ represents the local curvature. This would require a significantly more intricate blurring process, as the relationship between location, orientation, and curvature is very non-linear, but it would enforce more constant curvature contour integration; it would also reduce the prevalence of branching, in which two contours merge into one. In such instances, the stronger curvature - usually the curvature with lower magnitude - would suppress the other curvature, causing it to remain unlinked.

5.3 Classic Gestalt Dot Arrays

One of our original intentions when developing this model was to generate a computational approach which had the generic applicability and versatility of a model which operated on raw image data, but the intuitive simplicity and adaptability necessary to be compared with human performance even simple abstract tasks. And while we have made great progress towards satisfying those goals with the experiments described in section 4.3, a great deal of phenomena in the Gestalt grouping literature remains unaddressed. One of the most prevalent topics of investigation is the classic Gestalt array, such as the dot arrays depicted in Figures 3.1 and 3.4 (reprinted here in Figures 5.1A and 5.1B); how might our grouping model and hypothesis based evaluation process be applied here?

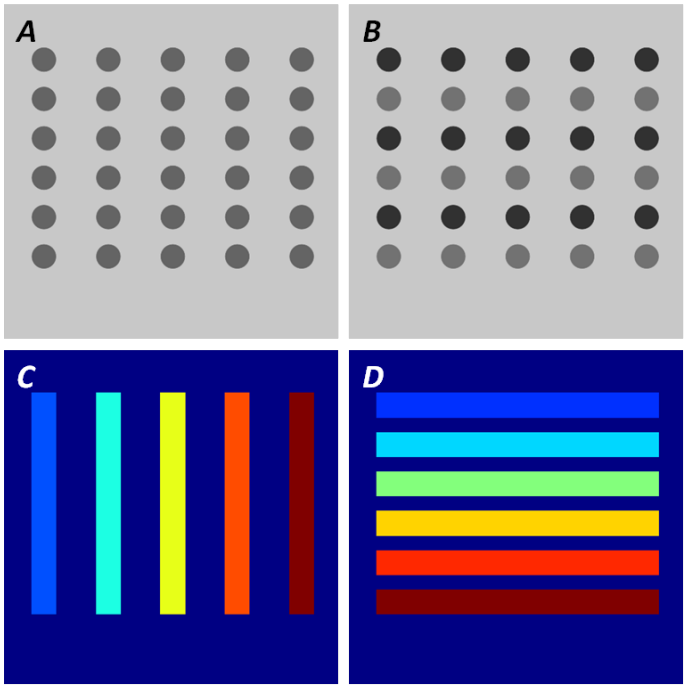

Figure 5.1: (A,B) Simple Gestalt dot arrays. Both can conceivably be organized as rows or columns. (C,D) Two hypotheses about the images in (A) and (B), represented as segmentations.

Each of these images has the same two possible mid-level interpretations; segmentations representing these two hypotheses are shown in Figures 5.1C and 5.1D. (Of course, both images can also be segmented into individual dots, grouped such that all dots form one segment, or grouped entirely into one full-image segment. But these interpretations are perceptually equally valid for both images.) For Figure 5.1A, hypothesis 5.1C seems more appropriate; conversely, hypothesis 5.1D seems a better explanation of Figure 5.1B.

Given these two hypotheses, how do we determine the model's prediction about these two images? One solution is to generate a population of images consistent with each of the two hypotheses. For example, an image consistent with hypothesis 5.1C would have dots arranged in 5 columns; the number of dots in each column might vary; some columns would contain dots of constant luminance, while in other the luminance would vary. The proportions and ranges of the parameters of each column would form an implicit prior on the set of possible images represented by hypothesis 5.1C. Each of these images would then be passed through the grouping model at a range of parameters, and each of the resulting segments would be compared with the hypothesis; thus one could calculate the distribution of these response arrays for images consistent with hypothesis 5.1c. The same procedure would be repeated for hypothesis 5.1D. With these two response array distributions in hand, a new image, such as Figure 5.1A or Figure 5.1B, would also be run through the grouping model; the resulting range of segmentations would be compared with both segmentations, yielding two full comparison arrays. The likelihood of each would be calculated, and whichever hypothesis yielded the highest likelihood would be deemed the correct interpretation of this image.

This process is closely related to that described in Section 4.2. In both cases, we are generating large sets of images generated by similar underlying structures. In Section 4.2, those images had similar luminance and proximity parameters, but random placement of individual dots; in this process, images are generated by the same structural hypotheis and with regular dot placement, but with randomized dot numbers and luminances. Also in both cases, we are representing a distribution of images with the distribution of hypothesis based-repsonse arrays, and the likelihood of a novel image is judged by the likelihood of its response array in that distribution.

Of course, there are many choices going into the extent and variety of images generated by the two hypotheses. These choices, however, would provide valuable insight: for example, if the model makes a prediction consistent with human behavior only when dot luminances in the same row or column vary in a particular range, that implies that human observers are making similar assumptions about the processes that generated the dot arrays. The grouping model, paired with simple hypotheses and a variety of possible image distributions, allow one to more directly probe the assumptions humans make when making inferences about the structure the visual world.

Part II

Silhouette Analysis and Representation

Chapter 6

Motivation

One of the greatest challenges the visual system must overcome is that the world we inhabit is three-dimensional, while the information presented to the visual system is two-dimensional. This difficulty manifests itself in many ways, but it is particularly troublesome in the analysis and representation of shape. Consider the images in Figure 6.1. These two horses are different in several ways - point of view, stance, background - but these differences are seemingly superficial. However, if we look at the silhouettes of these two horses (Figures 6.1C and 6.1D), we can see that these superficial three-dimensional differences introduce complex and variable changes in the resulting two-dimensional projections. Recognizing the similarity between these two silhouettes is very difficult, and cannot be reduced to a simple geometric comparison.

Nevertheless, silhouettes play an important role in the visual systems understanding of shape. Silhouettes are relatively easy to extract from a visual scene, and humans can often easily identify an object from its silhouette alone; in addition, some silhouettes, even very

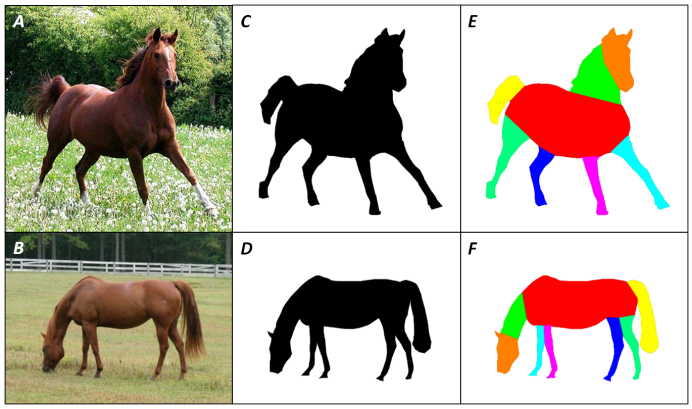

Figure 6.1: (a,b) Two images of a horse. Though the horse is in a different position, different background and viewed from a different angle, we can still clearly identify both images as horses. (c,d) Though we can easily idenitfy these silhouettes as horses, it is clear that no simple geometric relation will illuminate the kinship between these two silhouettes. (e,f) The part structure of the silhouettes allows us to find this similarity; corresponding parts have similar shapes, and the arrangement of parts is equivalent in both silhouettes.

simple ones, can elicit a strong perception of three-dimensionality (**?**). Also, as we saw in Figure 2, humans can makes sophisticated judgments of complexity and similarity about completely unfamiliar shapes; judgments which appear to be closely linked to the apparent part-structure of the given shapes.

It should come as no surprise that parts would play an important role in our understanding of silhouettes. Consider Figures 6.1E and 6.1F. Though the silhouettes of the two horses are, by most metrics, very different, their part structure is highly-similar. In addition, each of the parts, with the possible exception of the neck, has a very similar appearance in both silhouettes. Thus, if parts of a silhouette could be consistently and robustly extracted, recognition and identification of silhouettes would be greatly simplified. However, as with Gestalt grouping in the previous section, though the importance of part segmentation has long been recognized, a robust and versatile model of the process has remained elusive. In this section of my thesis, I define a simple, intuitive inflation technique called Puffball; I then describe how Puffball can be applied to the problem of silhouette part analysis in such a way that it avoids many of the pitfalls of many previous techniques. I describe several experimental and mathematical analyses which demonstrate that Puffball part segmentation performs as well as or better than existing part segmentation techniques, despite a much simpler implementation; and finally, I describe how Puffball might be applied to other silhouette analysis tasks, and what the success of Puffball part segmentation might suggest about the representation of silhouettes in the human visual system.

Chapter 7

Previous Work

Proposed by Harry Blum in 1967, the medial axis transform, or MAT, was one of the first approaches to silhouette representation inspired by the human visual representation of shape (Blum, 1967). (For clarity, in this proposal, when I say silhouette, I refer to a two-dimensional region of arbitrary topology and complexity; hence, any binary mask may be a silhouette.) More variable than the abstract classifications of topology, but more robust than raw geometry, Blum envisioned the MAT as a fundamental building block of human visual shape processing. The structure of the medial axis is robust to translations, rotations, and many non-rigid distortions; in addition, the branches of the medial axis often mirror the perceptual part structure of silhouettes (Figure 7.1). However, despite its power, the medial axis suffers from several severe limitations as well. Medial axes often contain extraneous branches or additional structural complexity, which do not reflect perceptually salient aspects of the shape; also, the structure of the medial axis is very sensitive to the path of the silhouette's contour, with large branches often created or destroyed by slight

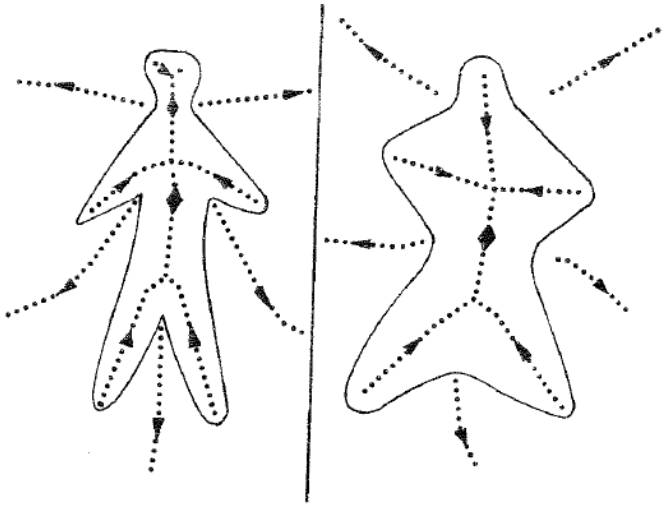

Figure 7.1: From Blum (1967). When calculated for humanoid shapes, the medial axis displays a perceptually appropriate skeletal structure, the branches of which correspond to the perceptual parts of the shape.

perturbations of the silhouette. Burdened by these limitations, the MAT never claimed its intended role as the foundation of human shape representation.

But despite considerable failings, the medial axis transform has not retreated to the back shelves of vision science. On the contrary, over the last four decades, the medial axis has become an essential component of the human vision, computer vision, graphics, and design researcher's toolbox. This is because, despite its imperfections, the MAT manages to capture a remarkably compelling distillation of the perceptually relevant structure of silhouettes with a surprisingly simple algorithm which can be run on a wide variety of inputs. This phenomenon - a tool which achieves limited but still impressive accuracy, and does so in a simple, intuitive, and versatile fashion - is not uncommon in the study of vision science, and may play a greater role in the human visual system than we realize. It is these strengths of the MAT, in addition to the MAT itself, that motivated our approach to the problem of silhouette analysis and representation.

Since the development of the MAT, the study of silhouettes has fractured into many different directions, both within human vision and into related fields such as computer vision and graphics. In the human vision community, the study of silhouettes has largely focused on how silhouettes may be used for object recognition; in particular, the question of segmenting silhouettes into perceptually relevant parts. This question is motivated by the hypothesis that silhouettes and real-world shapes cannot be represented and recalled in full formm; a more compact representation would encode objects and shapes as arrangements of easily describable geometrically simple parts.

Given a silhouette, it has been shown that humans do make part breaks quite consistently, even in silhouettes which are not identifiable as nameable objects (Siddiqi et al., 1996; De Winter and Wagemans, 2006). One class of approaches focuses on identifying the parts of a silhouette as belonging to an "alphabet" of primitives; Biederman's "geons" are perhaps the most influential example of this approach (Biederman, 1987). But geons do not seem to have a satisfying analogue for 2D shapes, and the identification of parallel contours which indicate their presence can be quite difficult. Pentland (1990) increased the flexibility and mathematical precision of the parts alphabet by proposing a technique that simultaneously breaks a shape into parts and reinflates the shape as a union of superquadric volumes, but his approach could still only be applied to shapes with relatively well-behaved, convex parts.

A second class of approaches instead focuses not on the parts themselves, but on the boundaries where the parts meet. The most influential of these approaches is likely the Minima Rule proposed by Hoffman and Richards (1984), which identified negative minima of principal curvature as the markers of part boundaries on three-dimensional surfaces; Hoffman and Richards theorized that the corresponding markers of part boundaries in two-dimensional silhouettes would be concave minima of contour curvature. Earlier work by Attneave supported the significance of curvature extrema (Attneave, 1954), and later psychophysical work seemed to confirm that humans place part-boundaries such that endpoints lie on or near minima of concave curvature (Braunstein et al., 1989; De Winter and Wagemans, 2006). However, the 2D Minima Rule proved difficult to develop into a working model; later work extended and elaborated on the approach, but numerous exceptions and seeming counterexamples to each system persist (Siddiqi and Kimia, 1995; Hoffman and Singh, 1997; Singh et al., 1999)). A different approach proposed by Mi and DeCarlo has shown promising results by moving away from the extrema of contour curvature, instead focusing axes of local

symmetry and scanning them for likely part transitions (Mi and DeCarlo, 2007; Mi et al., 2009).

Shape similarity is another application which has motivated a considerable amount of silhouette work, particularly in the computer vision community, driven by the demand for accurate content-based image retrieval systems. The variety of approaches to this problem is very wide, and cannot be completely covered here, but several classes of silhouette representations or models can be found in the shape similarity literature. Sharon and Mumford (2006) use a generative approach, modeling the difference between two shapes as the complexity or magnitude of a deformation required to map one shape onto the other. Others, noting the significance of parts to human vision, have proposed approaches that compare shapes by comparing non-metric part structure representations (Biederman, 1987); still others have blended these approaches into a hybrid analysis (Basri et al., 1998; Latecki and Lakämper, 2000). However, with effective methods for extracting parts from generic silhouettes lacking, these approaches have remained limited in applicability. Abbasi et al. (1999), perhaps inspired by the parts analysis work of Koenderink and van Doorn (1982), propose a method of silhouette analysis which locates curvature zero-crossings, or inflection points, in the shape's contour across a variety of scales.

Deformation of the shape contour is a popular approach, even in more general investigations of shape representation. The shocks approach evolves the boundary of the shape uniformly until singularities or "shocks" occur (e.g., a change in topology or the introduction of a cusp) (Kimia et al., 1995). Conformal mapping , on the other hand, represents a shape by analyzing the smooth deformation that takes that shape's contour to some canonical shape, such as the unit circle; and the similarity between two shapes can be viewed as a function of the deformation that maps one shape to the other (Sharon and Mumford, 2006).

Though the analysis of silhouettes and shape has received considerable attention, much of the work in this area has been restricted to the realm of computer science and applied computer vision. As a result, demand for data on human performance on tasks such as part segmentation and similarity judgments has been rather limited. Nevertheless, a number of valuable pieces of work have been done evaluating human silhouette judgments. These include investigations of human and mammalian similarity judgments (Scassellati et al., 1994; Op de Beeck et al., 2008) as well as human segmentations of silhouette parts (Siddiqi et al., 1996; De Winter and Wagemans, 2006). Any model of human silhouette analysis or of subproblems such as silhouette similarity and silhouette part structure must take these data into account. In addition, almost no work has been done relating human behavior to working models; for example, in De Winter and Wagemans (2006), the most extensive investigation of human part segmentation to date, the human part segmentation results are not compared with any working model of silhouette part segmentation.

Chapter 8

The New Idea: Puffball

8.1 Previous Work: Inflation

At the heart of the proposed approach to the representation and analysis of silhouettes is a simple tool for mapping two-dimensional silhouettes to three-dimensional shapes, a task known as inflation. As a formal problem, inflation is highly ill-posed: any given silhouette could have arisen from an infinite variety of possible three-dimensional shapes, and there is no clear ground truth against which to evaluate an inflation approach. Nevertheless, Tse (2002) has shown that certain silhouettes can elicit a clear perception of three-dimensionality, so it is not unreasonable to think that a process like inflation could occur in the visual system.

One early approach to inflation was developed by Terzopoulos et al. (1987); Terzopoulus and Witkin (1988). Their approach began with a user provided central axis of a given silhouette; the system then placed a deformable tube around the axis, and inflated or constricted the shape around the axis to match the silhouette. The results of the approach were intuitive and visually pleasing, but the physical surface model placed considerable limitations on the structure and topology of the surface; only simple shapes with a single major axis could be processed. Several more recent approaches also use the deformable surface model as their inflation mechanism (Pentland, 1990; Karpenko and Hughes, 2006) but the nature of the deformable surface approach requires careful selection of physical parameters (e.g. pressure, elasticity) and can severely limit the topological complexity of the inferred shape or shapes.

Much of the inflation work of the last 10 years has come out of the sketch interface community; because these algorithms utilize inflation as a design tool, rather than a shape inference technique, the ill-posed nature of the inflation problem is less of a concern. One of the best known techniques in this family is that of the Teddy system (Igarishi et al., 1999). Teddy, which inflates discrete polygons into polygonal meshes, builds on the axial strategy of Terzopoulos et al., but uses a more sophisticated axial structure which can be derived for an arbitrary shape. It then extends ribs from points along this axis to the edge of the shape, and places semicircular struts above and below these ribs. The inflated surface is then traced out by these semicircular struts. Teddy is a very powerful inflation tool, but is complex to implement; and, as pointed out in Alexe et al. (2004), piecing together the semicircular points can result in a bumpy surface which is less smooth than the output of some of its later competitors.

To solve this smoothness problem, several later systems have taken a more slightly sophisticated approach (Karpenko et al., 2002; Alexe et al., 2004; Tai et al., 2004). These systems calculate the same central axis structure as Teddy, but then use that structure to create an overall potential function on 3-D space, the level surface of which traces out the surface. This potential function is generally described by a finite set of parameters, which are carefully optimized to give the most consistent ant pleasing results. For example, in Alexe et al. (2004), spherical potential functions are placed along the central axis; to generate the final level surface, the distance parameters of each of these functions must be optimized to agree with the input silhouette.

8.2 Definition of Puffball Inflation

The grassfire height function, proposed by Blum (1967), can be thought of as a simple form of inflation. The silhouette is repeatedly eroded, resulting in a sequence of smaller and smaller silhouettes; these silhouettes can be summed over time to yield a height function on the interior points of the original silhouette, where the height at a given point is equal to the distance to the nearest edge (see Figure 8.1). Blum, of course, was not solving the problem of inflation, but rather calculating the medial axis transform, or MAT, in an effort to create a perceptually relevant skeletal shape descriptor.

The grassfire function forms the basis of many popular methods of creating beveled shapes in images, such as the Bevel and Emboss operation in Adobe Photoshop. A silhouette, such as the pair of Bs in Figure 8.2A, is passed through the grassfire function to give a beveled three-dimensional shape (Figure 8.2B). If a rounded silhouette inflation is desired, this height function can then be passed through a point-nonlinearity to give an appealing, rounded shape; but the result is scale-dependent, so the inflation of the smaller B is not simply a scaled-down version of the larger B (Figure 8.2C). The two Bs can be scaled properly if they are passed through different point non-linearities (Figure 8.2D), but in many situations a more desirable inflation approach is one which is inherently scale-invariant, and doesn't depend on post-hoc normalizations. We propose such an approach here, which we call Puffball inflation, or simply Puffball.

At the core of Puffball inflation is the principle: anywhere you can place a circle, place a

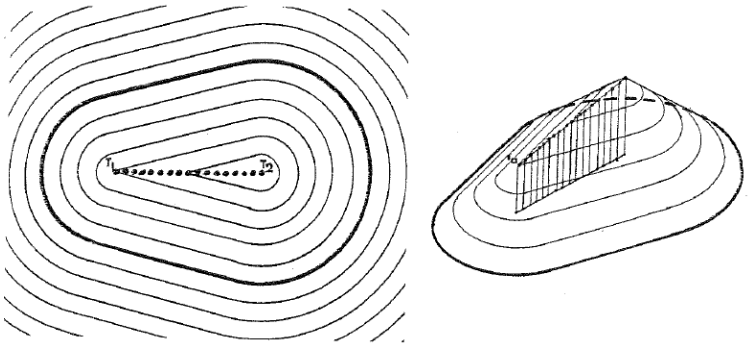

Figure 8.1: The grassfire height function. Note the ridge in the function lying above the medial axis of the shape.

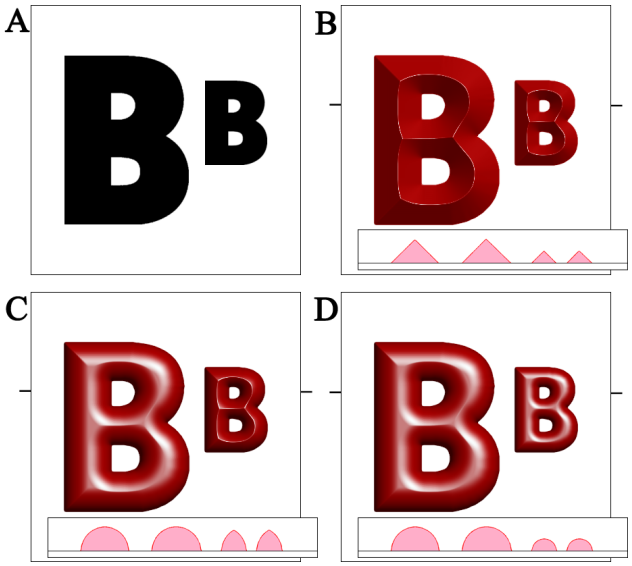

Figure 8.2: The bevel-nonlinearity inflation approach. (A) Two similar B silhouettes. (B) The grassfire height function of silhouette (A). A cross-section of the height function at the level marked by the hashes is shown in the inset. (C) Passing the grassfire height function through a non-linearity can yield a circular cross-section in one shape, but not both shapes simultaneously. (D) To get scale invariance, the two Bs must be passed through different point nonlinearities.

sphere. In fact, this principle can fully describe the output of Puffball inflation; in equation form, the Puffball inflation I of a silhouette S can be written:

$$I(S) = \bigcup\{B^3(p,r) \mid B^2(p,r) \subset S\} \tag{8.1}$$

where $B^3(p,r)$ is the spherical ball centered on point p with radius r, and $B^2(p,r)$ is the circular region centered on p with radius r contained in the plane of S. The set of such circles, however, is massive: at any interior point of S, infinitely many circles centered on that point lie entirely within S. So while Equation 8.1 is an elegant approach to silhouette inflation, it is deeply impractical in a computational setting. Fortunately, the process can be greatly accelerated by noting that

$$B^2(p_1,r_1) \subset B^2(p_2,r_2) \Rightarrow B^3(p_1,r_1) \subset B^3(p_2,r_2)$$

Thus, in calculating the Puffball volume, we need only consider those circles not contained in any larger circle which is also contained in S; that is, we need only consider the maximal circles of S. The centers and radii of the maximal circles of a silhouette S form the medial axis transform, or MAT, of the silhouette. As mentioned above, the MAT can be calculated by locating the ridges of the grassfire height function; this leads us to an alternative and much more practical definition of Puffball inflation:

$$I(S) = \bigcup\{B^3(p,r) \mid (p,r) \in \mathrm{MAT}(S)\} \tag{8.2}$$

This process is illustrated in Figure 8.3.

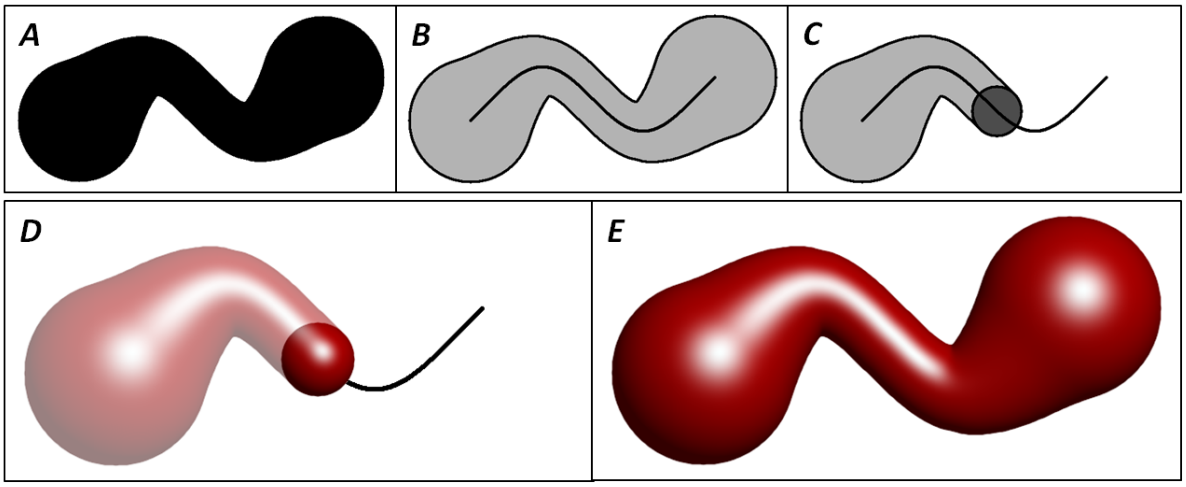

Figure 8.3: Puffball inflation. (A) Given a silhouette, (B) we can calculate the medial axis using morphological operations. (C) If we place circles of the appropriate radius along the medial axis, the resulting union completely reconstructs the original silhouette. (D) To calculate the Puffball inflation, we implement the same procedure, but we place spheres along the medial axis rather than circles. (E) The resulting inflated region.

MATLAB code implementing the algorithm can be found in Appendix B; the implementation takes a binary image as input and gives a height map image as output. Note that this does not calculate the union of spheres simply by taking the maximum; instead we use a soft maximum achieved by adding the exponential of each of the component spheres, and then taking a logarithm of the resulting sum. If a raw maximum is used, small numerical errors in the calculation of the grassfire height function (unavoidable in a discrete image) result in unsightly and perceptually inconsistent creases; the soft maximum eliminates these creases, while having a negligible effect on the overall shape of the output.

8.3 Strengths and Limitations of Puffball

The results of Puffball inflation on several simple silhouettes are shown in Figure 8.4. As the figure demonstrates, Puffball yields very intuitive results on all four shapes; a circle maps to a sphere, an ellipse maps to a prolate ellipsoid, etc. Puffball also achieves a high degree of scale-invariance, as shown in Figure 8.5. This scale invariance applies not only to similar silhouettes within an image but also to different parts of the same connected silhouette, and requires no normalization or post-hoc processing.

Figure 8.6 depicts a complex silhouette which was generated by thresholding random low-pass noise. The silhouette has multiple separate components, one of which is topologically very complex, and several of which extend beyond the boundary of the image. For inflation techniques which use physical models of the inflated surface, especially those dependent on deformable surfaces (Terzopoulos et al., 1987; Terzopoulus and Witkin, 1988; Pentland, 1990; Karpenko and Hughes, 2006), inflating a shape would require extracting the topology of the silhouette and constructing a surface with matching topology before any inflation

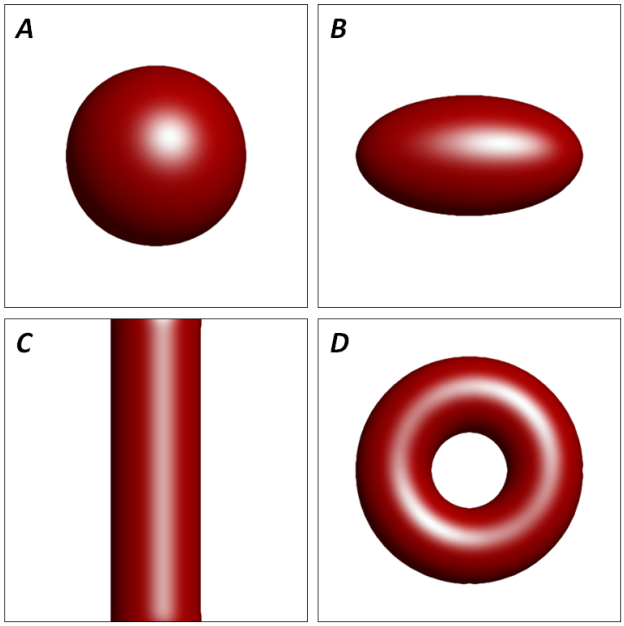

Figure 8.4: Results of Puffball inflation on several simple silhouettes.

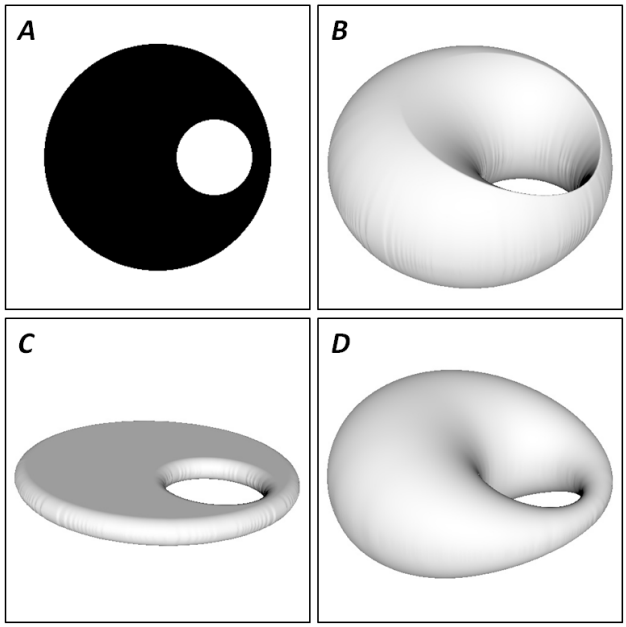

Figure 8.5: Scale invariance of Puffball inflation. (A) An offset annulus silhouette. The most intuitive inflation of silhouette would have a circular cross-section in both the large bend and in the small bend. (B,C) Passing the grassfire height function through a point non-linearity cannot achieve a circular cross-section in both bends simultaneously. Forcing a circular cross-section in the large bend introduces sharp ridges in the small bend; conversely, forcing a circular cross-section in the small bend forces the remainder of the inflation to be flat. (D) The Puffball inflation yields a circular cross-section in both bends of the resulting torus, and is a smooth, intuitive inflation of the original silhouette.

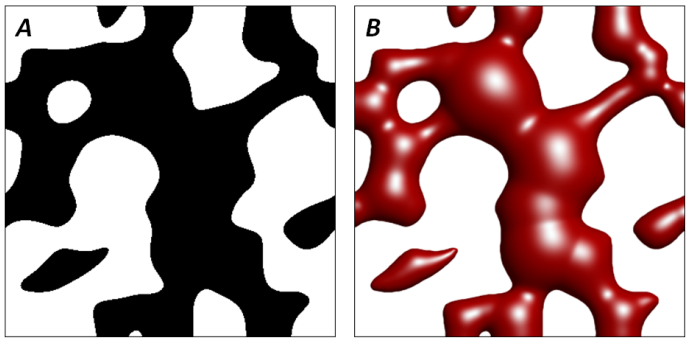

Figure 8.6: (A) A topologically complex silhouette generated by thresholding random low-pass noise. For most algorithms, a silhouette with this level of complexity would require considerable processing. (B) Puffball inflates the silhouette, with no additional machinery or extra processing.

could even begin. Even sketch-interface inflation approaches like Teddy (Igarishi et al., 1999; Karpenko et al., 2002; Alexe et al., 2004; Tai et al., 2004) which utilize a central axis structure still represent the resulting surface as a triangle mesh, in which the topology of the overall surface must be carefully monitored. But Puffball, because it operates entirely in the image domain, has no such limitations; the silhouette in Figure 8.6 presents no more difficulty to the algorithm than any other.

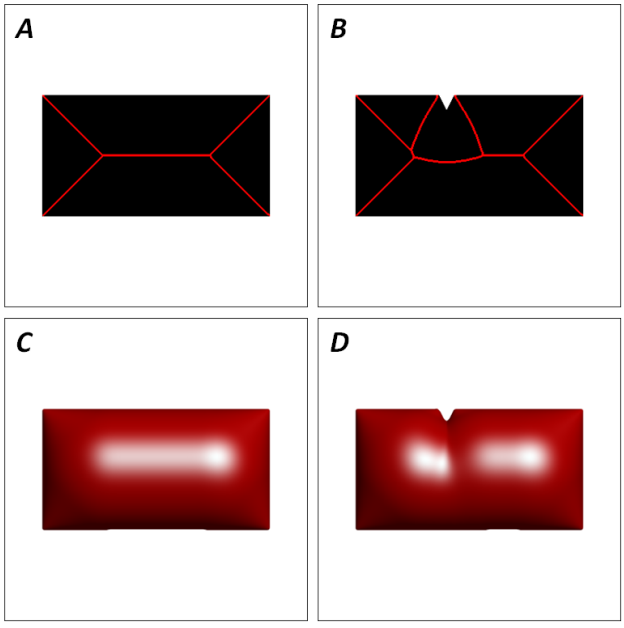

Figure 8.7: (A) The medial axis of a rectangle. (B) Perturbing the contour of the rectangle by removing a small piece from it causes significant, discontinuous changes to the structure of the medial axis. (C) The inflation of the rectangle. (D) Perturbing the contour of the rectangle does change the resulting inflation, but as the perturbation becomes smaller, so does the change in the inflated shape. So Puffball, unlike the MAT, is a continuous mapping.

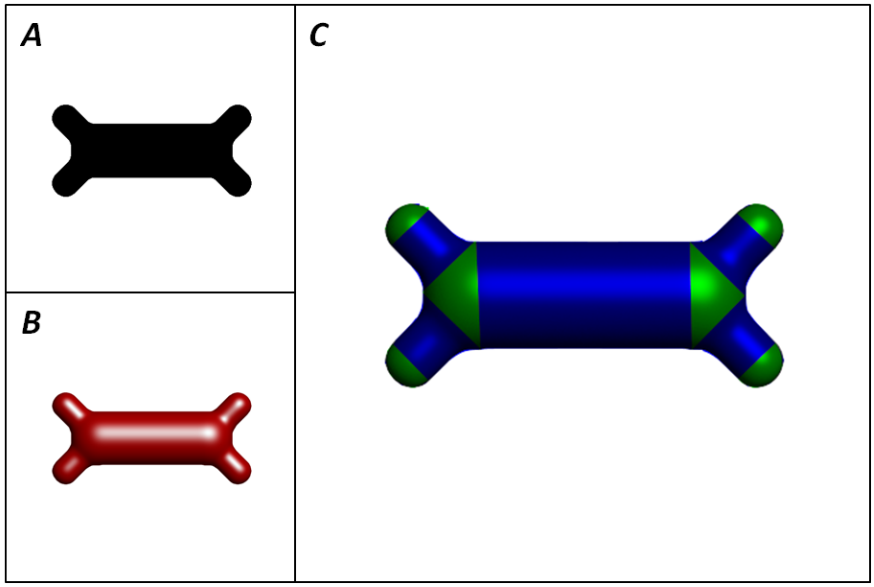

Figure 8.8: (A) A silhouette. (B) The Puffball inflation of silhouette (A). (C) The Puffball inflation can be geometrical broken into spherical regions (shown in green) and canal surface regions (shown in blue).

Because our implementation of Puffball is based around the medial axis transform, it is reasonable to ask if it suffers from similar limitations. In particular, is the resulting inflation robust to small perturbations of the input silhouette? Such perturbations can drastically alter the structure and appearance of the medial axis. However, the medial axis transform is not an essential component Puffball; Puffball may be defined with no mention of the MAT at all. As a result, small perturbations which disrupt the structure of the MAT have appropriately small effects on the resulting Puffball inflation (Figure 8.7). In short, Puffball is a robust, continuous mapping from two-dimensional silhouettes to three-dimensional regions.

Finally, due to its simplicity, Puffball can be analyzed more completely than many competing inflation approaches with existing mathematical knowledge. The surface that results from the Puffball inflation of a shape with a finite medial axis can be broken in two well-defined classes: spherical regions generated by a single branch-point, and intervening regions generated by the branches of the medial axis (Figure 8.8). These intervening regions fall into a class of surfaces known as canal surfaces, and they have been well studied in the mathematical literature (Garcia et al., 2006; Xu et al., 2006). This existing base of knowledge gives Puffball an additional advantage as a computational and modeling tool as its behavior can be analyzed mathematically as well as computationally.

Of course, Puffball is by no means the last word in inflation, nor is it a fully accurate model of human intuition about the relationship between 2D and 3D shape. For example, note how in Figure 8.6B, the shape contains noticeable bulges and creases. These are not the result of improper setting of the implicit soft-maximum parameter, but inherent properties of Puffball inflation. In addition, while most human observers interpret symmetric contours as bounding surfaces of revolution, Puffball make no such prediction. Finally, Puffball is constrained such that all silhouette contours are treated as extremal boundaries, where the

surface becomes tangent to the line of sight; but humans frequently interpret boundaries as hard corners or edges (Figure 8.9). But despite these limitations, Puffball's simplicity and intuitiveness - and ease of implementation - have yielded a number of illuminating applications and results.

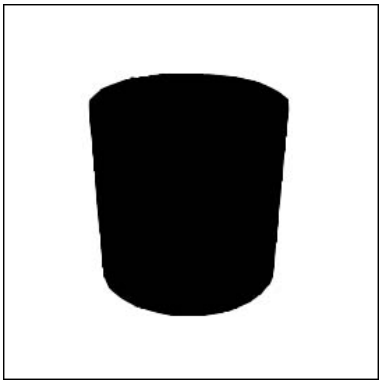

Figure 8.9: Humans can easily interpret this silhouette as a cylinder; in that interpretation, the top and bottom contours correspond to sharp edges of the shape, rather than extremal edges as would be predicted by Puffball inflation.

Chapter 9

Silhouette Part Segmentation

9.1 Puffball Part Segmentation

In their paper, "Parts of Recognition," Hoffman and Richards (1984) described two rules for locating parts in shapes. The second of these two rules, which is most often referred to as the "minima rule," and which I will refer to as the 2D Minima Rule, describes how part boundaries might be located in two-dimensional silhouettes using minima of contour curvature. This rule, as mentioned above, has led to considerable computational challenges. But the 2D Minima Rule was itself inspired by the first rule, also referred to as the "minima rule," which I will refer to as the 3D Minima Rule. The 3D Minima Rule is a much richer and more robust principle than its 2D cousin, derived from a generative principle of parts referred to as the principle of transversality. The 3D Minima Rule states that part boundaries on the surface of a 3D shape should be placed along loci of negative minima of principal curvature (Figure 9.1). Though far from a perfect description of real-world parts, the 3D minima rule is an elegant and intuitive principle which satisfies all three of the constraints that Hoffman and Richards placed on a valid approach to parts analysis: reliability, versatility,

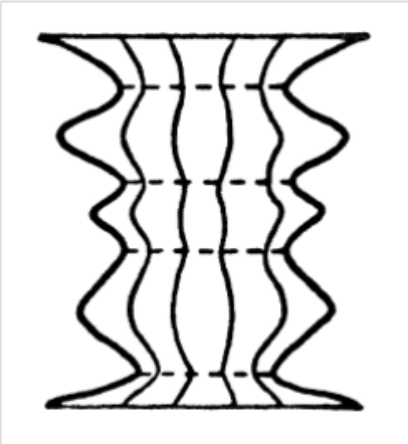

Figure 9.1: Visual part boundaries on a three-dimensional surface can be placed at the loci of minimal negative principal curvature.

and computability.

Of course, as Hoffman and Richards correctly pointed out, the strength of the 3D Minima Rule cannot be applied to silhouettes, as reconstructing a three-dimensional shape from a two-dimensional silhouette is a mathematical impossibility; thus Hoffman and Richards developed the 2D Minima Rule as an two-dimensional analogue. But the availability of Puffball suggests an alternate approach: perhaps one does not need to correctly reconstruct the three-dimensional surface to perform an effective parts analysis. It may be sufficient to infer a sufficiently intuitive three-dimensional pseudoshape, which can then be analyzed using the 3D Minima Rule. Indeed, I will later argue that using a regularly derived pseudoshape can actually yield better parts analysis than knowledge of the real three-dimensional surface.

Thus, our new approach to part-segmentation proceeds as follows: given a silhouette, such as one in Figure 9.2A, calculate the Puffball inflation (Figure 9.2B). On this surface, calculate the principal curvatures at each point, and locate bands of minimal negative principal curvature (Figure 9.2C). We are helped here by our understanding of the geometry of Puffball surfaces: of the two classes of regions that appear in Puffball surfaces, we will only find bands of negative curvature in the canal surface regions (spherical surfaces have constant positive principal curvature). And on canal surfaces, one principal curvature is always positive; thus we need only analyze the lesser principal curvature at every point.

According to the 3D Minima Rule, part boundaries are marked by loci of minimal negative principal curvature. In practice, however, locating these full loci can be challenging. First, not all real-world part boundaries are have negative principal curvature along their full length (consider the point where your arm joins your torso). In addition, the output

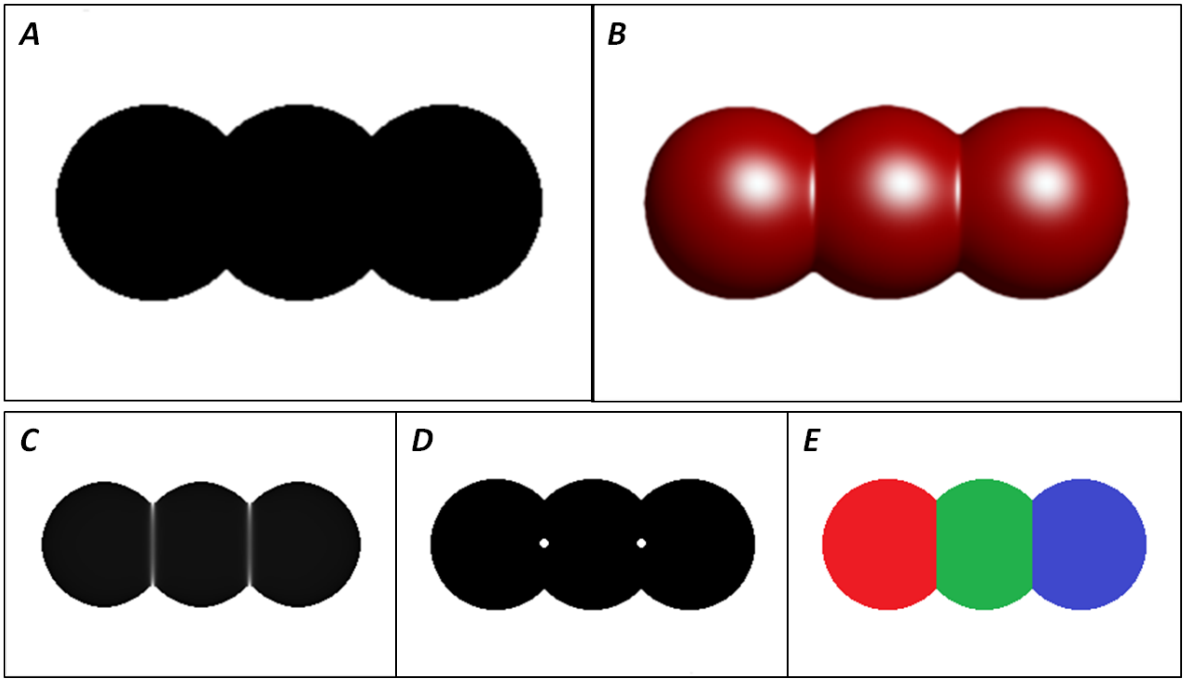

Figure 9.2: Puffball part segmentation. (A) An initial silhouette. (B) The Puffball inflation. (C) Bands of principal curvature on the Puffball inflation mark part boundaries. (D) Points of maximal principal curvature. (E) The resulting segmentation.

of Puffball is a height surface, which makes measurement of principal curvature near the extremal boundaries very unstable. Fortunately, a more robust, more reliable approach requires only locating points of minimal principal curvature "along the top" of the Puffball inflation (Figure 9.2D). Here, "along the top" means points where the derivative in the direction of maximum principal curvature is 0; if a canal surface section of a Puffball inflation is viewed as a series of circular ribs around a central spine, the points at the top of each rib lie along the top of the Puffball inflation. Once the top-most points of minimal negative principal curvature have been located, part-lines can be placed through them across the shape; there are several ways to do this, but the simplest is to locate the shortest line across the shape passing through the identified point. These lines segment the shape, yielding the final part-segmentation (Figure 9.2E).

It is important to note that small numerical errors in the calculated inflation will result in very shallow or transient principal curvature minima; it may also be the case that we wish to evaluate and compare the relative strength of part boundaries. The simplest choice is to threshold and rank part boundaries based on the magnitude of the principal curvature at the point the generated them; but this introduces a scale invariance, as a surface that is scaled up by a common factor will have all its principal curvatures scaled down by the same factor. However, if we simply normalize the principal curvature by the height of the Puffball height map at that point, the resulting normalized curvature is scale invariant. Using this value allows one to ignore noisy minima and rank the resulting part boundaries in a simple scale-invariant way.

9.2 Strengths of Puffball Part Segmentation

Consider the hand silhouette in Figure 9.3. This silhouette illustrates two of the issues any algorithm based on the 2D Minima Rule must resolve. First while the boundary between the palm of the hand and the index finger terminates at a curvature minimum (indicated by a red dot in Figure 9.3B) on one end, the other end does not lie at or near a significant curvature minimum. Though the majority of part boundaries terminate at or near two

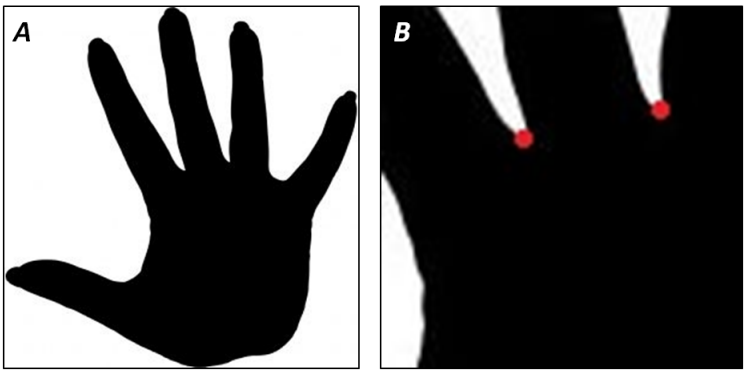

Figure 9.3: Though the silhouette has very clearly discernible part structure, the 2D Minima Rule leaves many questions unanswered. Though the curvature minima can be easily located, it is difficult to tell what to do next.

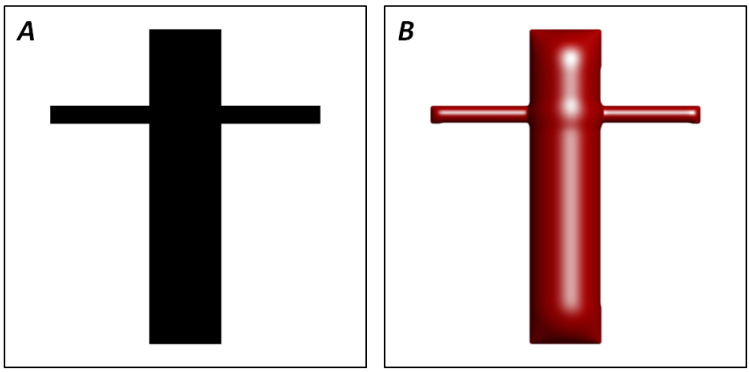

Figure 9.4: The Short Cut Rule. (A) According to the Short Cut Rule, given a silhouette with two possible part interpretations, a human chooses the part cuts that are shorter. (B) The Puffball inflation of the silhouette yields a higher principal curvature above the short part-cut than above the long part-cut. Thus Puffball part-segmentation implicitly obeys the Short Cut Rule.

part boundaries, it is still very common to have a part boundary with only one minimal endpoint. On the other hand, both the boundary of the index finger and the boundary of the middle finger terminate at the endpoint between the two fingers. If one has only the locations of the curvature minima, there is no way to know whether two part-lines meeting at the same endpoint are mutually inconsistent, or, as is the case here, are entirely consistent. Thus, while curvature minima do generally lie at or near part boundaries, and vice versa, curvature minima alone do not specify a part segmentation; any computable, reliable segmentation requires substantial additional computational machinery.

One such piece of machinery, proposed by Singh et al. (1999) is the Short Cut Rule. Though Singh et al. proposed several constraints on the selection of part-lines based on curvature minima, the most important was the principle that given two inconsistent but otherwise equally valid part-lines, humans will choose the part-line which is shorter. Thus, if one had to choose between drawing vertical cuts or horizontal cuts in the segmentation of the silhouette in Figure 9.4A, one would likely choose the shorter vertical cuts.

The Short Cut Rule does resolve some, if not all, of the ambiguities present in the 2D Minima Rule; but is this additional constraint necessary? Figure 9.4B depicts the Puffball inflation of the silhouette in Figure 9.4A; it is clear that the principal curvature above the shorter part-lines is considerably larger in magnitude than the principal curvature above the longer part-lines. Thus Puffball part segmentation will rate the shorter part-lines as better part boundary candidates than the longer part-lines. Thus Puffball part segmentation, unlike the 2D Minima Rule, implicitly includes the Short Cut Rule, making the extra computational machinery of the Short Cut Rule unnecessary.

Another addition to the 2D Minima Rule part boundary approach is the Necks and Limbs algorithm, proposed by Siddiqi et al. (1996). According to the Necks and Limbs theory, humans identify two classes of part boundary: necks which form the bridges or connections between two nearby components; and limbs, which separate larger components from adjoining smaller parts. Both kinds of part boundaries are illustrated in Figure 9.5. In the Necks and Limbs segmentation algorithm, necks and limbs are calculated through two

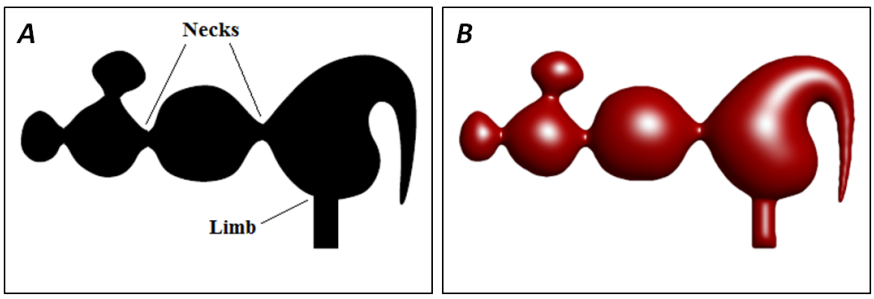

Figure 9.5: Necks and Limbs. (A) The algorithm proposed by Siddiqi et al. (1996) suggests different computational procedures to identify necks and limbs. (B) Bands of negative principal curvature on the Puffball inflation which mark part boundaries appear over both necks and limbs, making the division computationally unnecessary.

different computational methods.

If we inspect the Puffball inflation of a shape containing both necks and limbs, however, we see that both part boundary types exhibit the bands of negative principal curvature that Puffball part segmentation seeks (Figure 9.5B). Thus Puffball part segmentation identifies both limb-type and neck-type part boundaries, with no additional or separate machinery needed for either. While it is certainly possible that two separate methods are used by the visual system for the two boundary types, there is no strong psychophysical evidence for this, and Puffball part segmentation effectively proves that separate computational approaches are not necessary.

Given that the motivation of Puffball part segmentations was to generate a candidate 3D shape to which we can apply the 3D Minima Rule, one may wonder whether the ill-posed nature of the inflation task limits Puffball part segmentation's performance. That is, can Puffball hope to give a reasonable part segmentation if it cannot hope to consistently and correctly reconstruct a three-dimensional shape? If we return to the hand silhouette from Figure 9.3, however, we find that, counter-intuitively, Puffball's performance can be superior to that of a method which correctly inferred the existing three-dimensional shape; for though the 3D Minima Rule is a well-motivated and elegant approach to three-dimensional part segmentation, it does not always apply. If one inspects a human hand, one will find that the front of the hand does have creases or bands of negative principal curvature (Figure 9.6A), but that the back of the hand contains no such part markers (Figure 9.6B). Nevertheless, it is clear to us that a hand - and a hand silhouette - should be broken into parts, with part boundaries at or near the base of each finger. If, on the other hand, we inspect the Puffball inflation of a hand silhouette - which is clearly an incorrectly inferred three-dimensional shape - we see that the telltale bands of negative principal curvature are present all the way around the bases of the fingers (Figure 9.6C). Thus, searching for part boundaries on the Puffball inflated shape yields clearer, more robust part boundaries than inspecting the actual real world shape (Figure 9.6D).

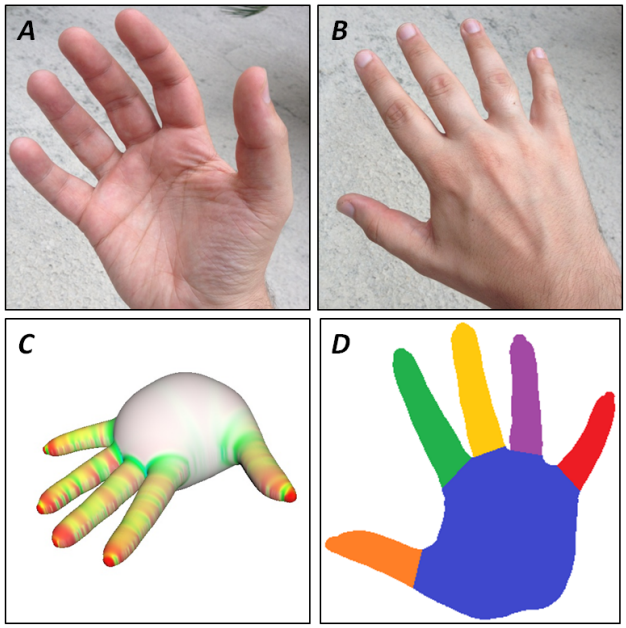

Figure 9.6: (A,B) On a real hand, bands of negative principal curvature can be seen at the base of the fingers on the palm side, but not on the back side. (C) The Puffball inflation displays much clearer and more complete bands of negative principal curvature (indicated by saturated green). (D) This allows the Puffball inflation to perform a more intuitive part segmentation than would result if one applied the 3D Minima Rule to an actual hand.

Chapter 10

Evaluation

10.1 Preliminary Results

As an initial evaluation of Puffball's part segmentation performance, a small study was run on Amazon Mechanical Turk. The experiment compared part segmentations generated by Puffball part segmentation with my best-effort implementation of the Necks and Limbs algorithm (Siddiqi and Kimia, 1995) and a part segmentation algorithm imposing the constraints of the Short Cut Rule (Singh et al., 1999). Necks and Limbs was implemented as described in Siddiqi and Kimia (1995); a description of my implementations of the Short Cut Rule can be found in the Appendix C.

The experiment was run using 24 silhouettes, 8 each in three classic part segmentation categories: animals, hand tools, and human figures. Each silhouette was run through all three part segmentation algorithms, and converted to a segmentation image in which different segments were indicated by blocks of different colors (Figure 10.1).

In each trial of the experiment, a subject was presented with one of the silhouettes and two segmentations of that silhouette, one of which was always generated by Puffball part

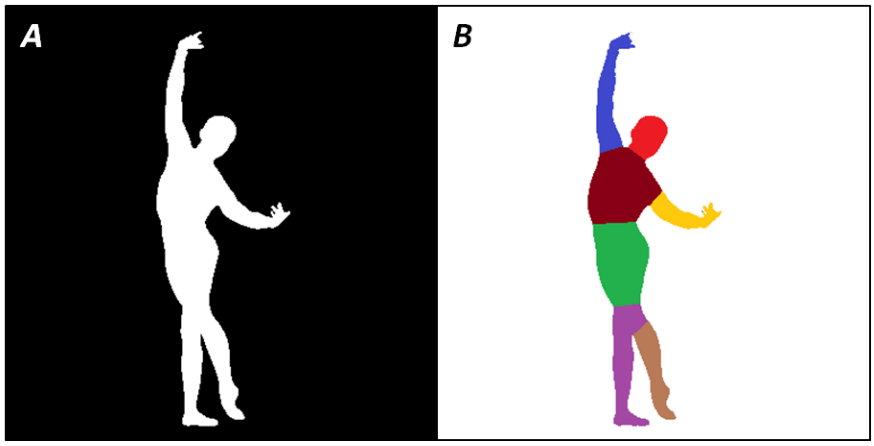

Figure 10.1: (A) A sample silhouette from the pilot experiment, from the category of human figures. (B) A depiction of a segmentation of the silhouette in (A), generated by Puffball part segmentation.

segmentation. They were told that the silhouette had been broken down into parts by two algorithms, and were asked to choose which of the two segmentations "looked more correct." The experiment was run by 40 total subjects; with 24 silhouettes and two comparison algorithms, each subject ran a total of 48 trials (every subject saw each silhouette twice). Hence, there were a total of 1920 trials, 960 each for each comparison algorithm and 320 each for each comparison algorithm and category combination. Results of the experiment are shown in Figures 10.2 and 10.3.

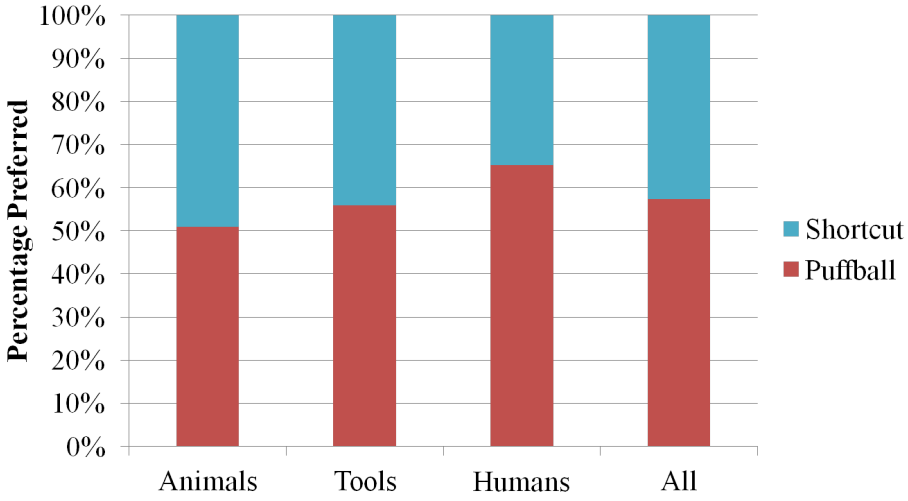

Figure 10.2: Results of pilot study comparing Puffball part segmentation with Short Cut Rule-based part segmentation. The preference for Puffball in the hand tools category was statistically significant with $p < 0.025$. The preference for Puffball in the human figure category and for all shape together was significant with $p < 0.001$.

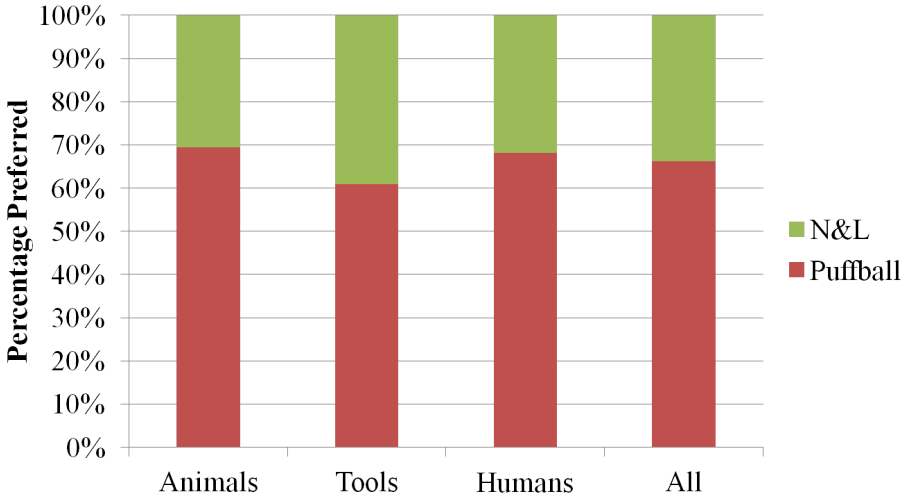

Figure 10.3: Results of pilot study comparing Puffball part segmentation with Necks and Limbs part segmentation. The preference for Puffball in all categories and overall was statistically significant with $p < 0.001$.

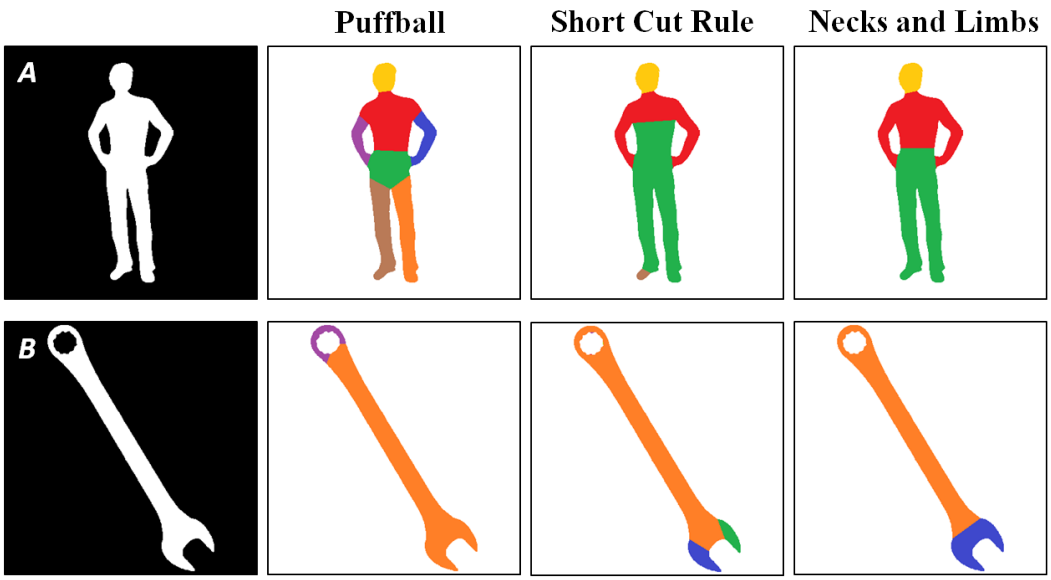

Figure 10.4: Comparison of behavior of the three segmentation algorithms on two silhouettes.

In every category, Puffball was chosen more often than the competing algorithm, with the preference being statistically significant in all but one category. Though the pilot study was quite small, the results do suggest that an inflation based part segmentation approach can perform on par with more traditional, contour-curvature-based approaches that utilize the 2D Minima Rule.

Figure 10.4 shows some sample segmentations, illustrating what mistakes Puffball segmentation was able to avoid, and a silhouette on which Puffball did not perform as well. In the human silhouette, the boundary between the torso and each arm terminates at only one curvature minima; thus the 2D Minima Rule-based algorithms are unable to segment the arms from the torso. Similarly, the boundaries of the legs have only one minimal endpoint; to make matters worse, both leg boundaries terminate at the same curvature minimum; so neither 2D Minima Rule based segmentation identifies the legs as separate segments either. Puffball, however, is able to segment the arms and the legs, as it does not seek pairs of contour minima. So, while the Puffball segmentation is not necessarily perfect (segmentation of the waist seems largely unnecssary) it is far more intutive than the segmentations produced by the 2D Minima Rule-based algorithsm. The wrench silhouette, however, presented some difficulty for Puffball. In this case, flaring of the wrench near the bottom has little effect on the inflation, because the negative space in the opening of the wrench prevents the inflated shape from becoming too large. So very little curvature is present on the top of the inflated shape, and the boundary between the shaft of the wrench and the "head" of the wrench is not identified, as it is in the Necks and Limbs segmentation. Again, none of the segmentations is perfect - a proper segmentation would identify both the "head" of the wrench and the loop at the other end as distinct functional parts - the Necks and Limbs algorithm is the most intuitive of the three.

10.2 Further Experimental Evaluation

Further attempts to evaluate the segmentations in an experimental setting yielded frustratingly inconclusive results. With a relatively small number of subjects, statistical significance was nearly impossible to achieve, and subject segmentation preferences were dominated by one particular confounding factor: segment number. Far more often than not, subjects selected the segmentation with fewer segments, independent of almost all other considerations. In the absence of any clear right answer, it was clear that the subjects were often settling on the simplest possible dimension along which to discriminate the segmentations they were presented with.

To counteract this confounding factor, segmentations generated for the next experiment used an adaptive threshold. Each algorithm includes a threshold which controls the number of parts it predicts. For Puffball part segmentation this threshold is the minimum scaled principal curvature along the top of the inflated shape; for Necks and Limbs and the Short Cut Rule, it is the minimum contour curvature magnitude that constitutes a relevant curvature minimum. For each silhouette and each algorithm, the threshold was chosen so that the number of parts predicted was equal to the rounded mean number of parts segmented by human subjects. In this way, the influence of segment number was eliminated from the preferences in the experiment.

Once again the experiment was run on Amazon Mechanical Turk; because each pair of segmentations needed to be compared multiple times, and it is less informative to show the same pair of segmentations to the same subject more than once, a large number of subjects were needed. Also, because the depiction of part segments is quite straightforward and largely display independent, the varying conditions of Mechanical Turk workers have little influence on the result.

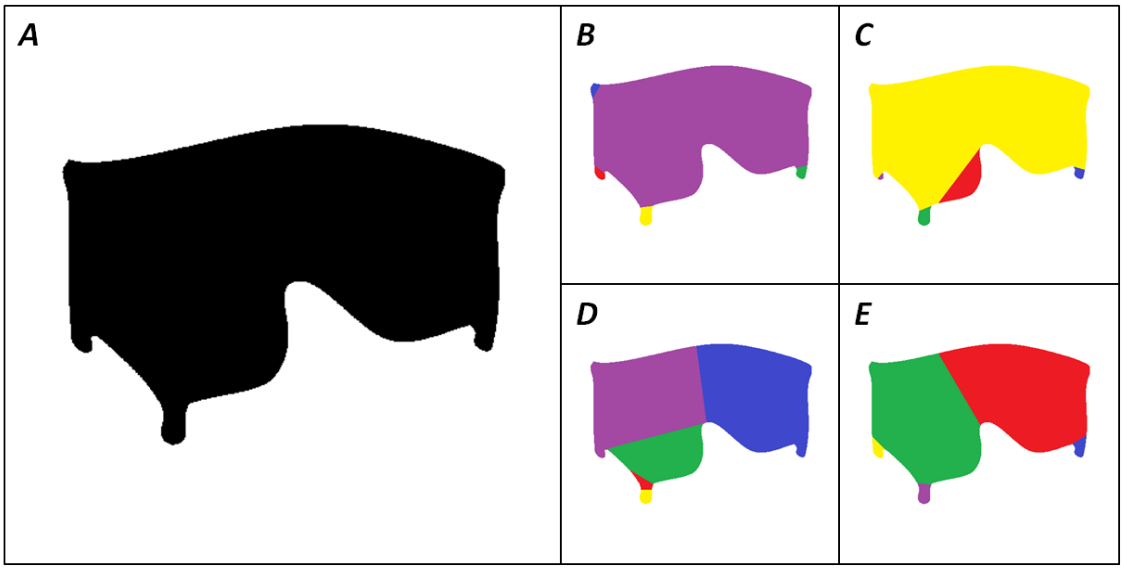

Figure 10.5: (A) A silhouette used in the Mechanical Turk experiment. (B) A sample human segmentation of silhouette (A). (C) The Necks and Limbs segmentation. (d) The Short Cut Rule segmentation. (E) The Puffball part segmentation.

The experiment was run using 44 silhouettes from the De Winter and Wagemans part segmentation dataset. Of the 88 silhouettes used in that study, 44 were deemed easy to identify, and 44 were deemed difficult to identify; our Mechanical Turk experiment focused on the difficult to identify silhouettes, to minimize the effect of top-down knowledge and object recognition. In every trial of the experiment, the subject was shown the silhouette, and then shown two segmentations of that silhouette. Each of these two segmentations was drawn from one of four possible sources: the three segmentation algorithms (Puffball part segmentation, Necks and Limbs, or the Short Cut Rule), or a human generated segmentation from the De Winter and Wagemans data set. Every subject saw all 44 silhouettes, seeing each silhouette only once. When presented with the two segmentations, they were prompted to select which of the two segmentations looked more correct. An example silhouette and segmentations are shown in Figure 10.5.

Also included in the experiment were 6 simple silhouettes which were paired with one intuitive segmentation and one highly counterintuitive segmentation. These silhouettes were included to ensure that subjects were attending to the task. Of the 96 subjects who completed the experiment, 5 subjects who chose the counterintuitive segmentation on at least 4 of these 6 test silhouettes were excluded. An example test silhouette and associated segmentations are shown in Figure 10.6.

The results of this experiment are shown in Table 10.1. On the whole, the results are encouraging; as with the pilot study, Puffball was preferred over the two competing algorithms, though in this case only the preference over Necks and Limbs was statistically significant. What is most troubling is that by a highly statistically significant margin, all three segmentation algorithms were preferred over the human segmentations. This does not mean that the human segmentations in the De Winter and Wagemans study were wrong; nor does it mean that the subjects in our Mechanical Turk experiment were incapable of evaluating the segmentations. It does, however, reveal a very deep inconsistency between the two sets of subjects understanding of what constitutes an appropriate part segmentation. Future work

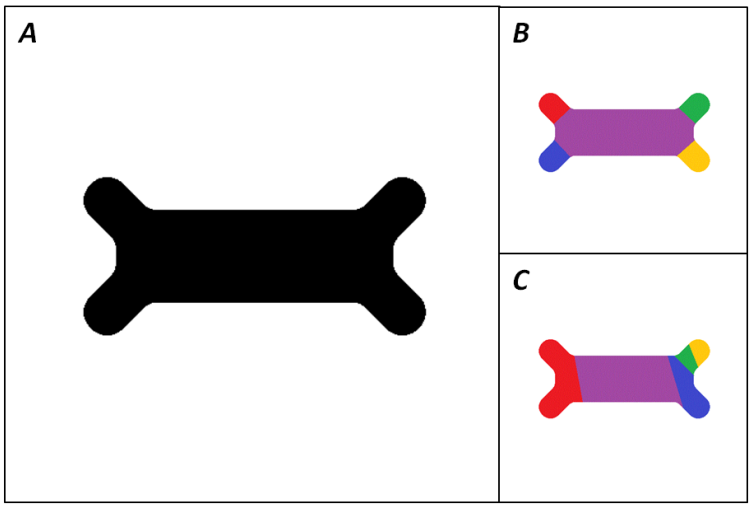

Figure 10.6: (A) A sample test silhouette. (B) An intuitive part segmentation of that silhouette. (C) A counter-intuitive segmentation of that silhouette.

	Puffball	Short Cut Rule	Necks and Limbs	Human Subjects
Puffball		0.5074	0.5446 (*)	0.6239 (**)
Short Cut Rule	0.4926		0.5474 (*)	0.6146 (**)
Necks and Limbs	0.4554 (*)	0.4526 (*)		0.5985 (**)
Human Subjects	0.3761 (**)	0.3854 (**)	0.4015 (**)	

Table 10.1: Results of the full Amazon Mechanical Turk experiment. Numbers indicate the proportion of trials in which the segmentation source labeling the row was preferred over the source labeling the column. Thus Puffball part segmentation was preferred over the Short Cut Rule in 50.74% of trials. Cell shading has been added for clarity, with red indicating positive preference, blue indicating negative preference, and saturation indicating degree of preference. (*) indicates statistical significance with $p < 0.025$; (**) indicates statistical significance with $p < 0.001$.

asking human subjects to generated part segmentations should take this fact into account, by carefully selecting how they instruct their subjects to generate their segmentations, and checking the reliability of those segmentations by having them evaluated by a different pool of subjects.

10.3 Numerical Evaluation

While the results of the experimental evaluations were encouraging, I also performed extensive mathematical and numerical evaluation of the competing part segmentations algorithms using the human-generated part segmentations in the De Winter and Wagemans dataset. For these analyses, all 88 silhouettes were run through the three part segmentation algorithms; for these comparisons, the relevant thresholds in the three algorithms were again selected for each silhouette such that the number of parts output by each algorithm was equal to the rounded mean number of parts in the human segmentations of the same silhouette. The numerical evaluations fell into three categories: comparison of full segmentations, comparison of individual part-lines, and comparison of part-line endpoints.

10.3.1 Comparison of Full Segmentations

The simplest and most direct approach to evaluating a part segmentation algorithm would appear to be to compare a full segmentation output by the algorithm with a number of segmentations generated by human subjects. Unfortunately, as was mentioned in the chapter on perceptual grouping, there is no accepted metric for comparing segmentations or partitions. Therefore, to evaluate the performance of the three algorithms implemented for this research, I tried several different metrics:

- One metric, which I derived, I will refer to as weighted incoherence:

$$WI(S_1, S_2) = \frac{2 \log N(S_1 \cap S_2)}{\log N(S_1) + \log N(S_2)} \tag{10.1}$$

where $N(S)$ is the numerosity of the segmentation S, defined as:

$$N(S) = \frac{(\sum_i |s_i|)^2}{\sum_i |s_i|^2}$$

where $|s_i|$ is the area of an individual segment of S. If all segments are of equal area, than the numerosity of a segmentation is the number of segments; otherwise, the numerosity is strictly less than the number of segments. Weighted incoherence is equal to 1 for identical segmentations, and larger for more dissimilar ones.

- Another metric, inspired by Martin et al. (2001), will be referred to as mean local consistency error:

$$mLCE(S_1, S_2) = \frac{1}{2A} \sum_p E(S_1, S_2, p) + E(S_2, S_1, p) \tag{10.2}$$

where A is the area of the silhouette and the function E is defined as:

$$E(S_1, S_2, p) = \frac{|R(S_1, p) \setminus R(S_2, p)|}{|R(S_1, p)|}$$

where $R(S, p)$ is the individual segment in the segmentation S containing the pixel p and $\setminus$ indicates set difference. When two segmentations are identical, the mean local consistency error is 0; more dissimilar segmentations have a higher $mLCE$. This measure is different from the local consistency error described by Martin et al. in that the mean of the two values of E is taken for each pixel, rather than the minimum. As a result of this change, a segmentation S_1 which is a refinement of another segmentation S_2 will not have an error of 0 when compared to that segmentation.

- A third metric, described in the perceptual grouping chapter, is Meilă's variation of information (Meilă, 2007):

$$VI(S_1, S_2) = H(S_1) + H(S_2) - 2I(S_1, S_2) \tag{10.3}$$

where H is entropy, and I is mutual information. The variation of information is equal to 0 for identical segmentations, and larger for more dissimilar ones.

For each silhouette and each segmentation metric, each pair of human segmentations was compared; all these measures were averaged to give a silhouette a mean inter-subject disagreement. Then for each silhouette, each part segmentation algorithm, and each segmentation metric, the resulting segmentation of that silhouette was compared with all the human segmentation, to give a silhouette and algorithm a mean algorithm-subject disagreement. The averages of these values for all three metrics are shown in Table 10.2.

Unfortunately the results of this analysis are by no means clear. While weighted incoherence indicated Puffball as having the best agreement with human subjects, mean local consistency error and variation of information measured a higher agreement with the Necks

	WI	mLCE	VI
Inter-Subject	1.4148	0.1745	0.6143
Puffball	1.4901	0.1953	0.6690
Short Cut Rule	1.5040	0.2185	0.7542
Necks and Limbs	1.5624	0.1749	0.6223

Table 10.2: Comparison of full segmentations. Table shows comparison values under three metrics: weighted incoherence (WI), mean local consistency error (mLCE), and variation of information (VI). The inter subject row reports the average value across silhouettes of the mean inter-subject disagreement. The remaining three rows show, for the three competing algorithms, the average value across all silhouettes of the mean algorithm-subject disagreement. Lower values indicate better agreement.

and Limbs theory. This may be in part due to the nature of the measures; variation of information and mean local consistency error tend to have higher values for silhouettes with on average more segments; thus the overall values will be skewed towards performance on more complex silhouettes, where Puffball has a tendency to oversegment. However, no metric declared Puffball part segmentation the worst of the three competing algorithms, lending further weight to the idea that Puffball achieves competitive results with a much simpler and more intuitive algorithmic approach.

10.3.2 Comparison of Part-Lines

Perhaps the clearest result of the previous analysis is that analyzing similarity of segmentations is as much art as it is science. To avoid some of this difficulty and ambiguity, we can analyze a simpler data structure: the placement of individual part-lines. Of course, analyzing individual part-lines does not capture the full complexity of a segmentation; after all,

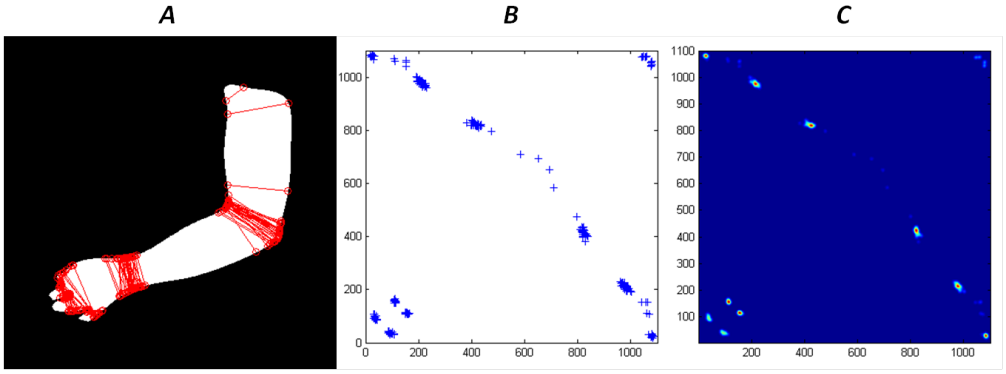

Figure 10.7: (A) The set of all human generated part-lines on a silhouette of a human arm. There are clearly several zones in which the part-lines cluster. (B) Plot of the same part-lines in a two-dimensional circular part-line space. The plot is symmetric because part-lines are equally valid in either direction. (C) The modeled distribution of part-lines for this silhouette. Several modes can be picked out, corresponding to the elbow, wrist, and fingers.

Mean Part-Line Likelihood ($\times 10^{-5}$)			
	σ = 2.5 pixels	σ = 5 pixels	σ = 10 pixels
Inter-subject	7.5541	3.3637	1.4411
Puffball	24.144	14.221	6.5024
Short Cut Rule	12.484	8.8323	4.6982
Necks and Limbs	16.352	10.788	5.3582

Table 10.3: Mean part-line likelihood for human subjects, Puffball part segmentation, the Short Cut Rule, and the Necks and Limbs algorithm, averaged over all 88 silhouettes.

a segmentation does not merely consist of a fixed number of independent samples from an underlying distribution of part-lines. But valuable information may still be gleaned from the distribution of part-lines for a given silhouette, and it is a much easier structure to analyze statistically.

Each part-line drawn in a silhouette can be considered an unordered pair of points along the contour of the shape. If we assume that the silhouette is a simple connected region, the contour of the silhouette can be viewed as a circular one-dimensional space, and the space of possible part-lines can be viewed as a circular two-dimensional space (Figure 10.7A and Figure 10.7B).

On the assumption that human subjects place part-lines with a certain amount of noise at locations around the silhouettes contour, we can model the distribution of part-lines by plotting all human subject-generated part-lines in this circular two-dimensional space, and filtering the space with a Gaussian window (Figure 10.7C). A new part-lines agreement with the existing set of part-lines can then be calculated as its likelihood in this modeled distribution.

For each silhouette, and each subject A, the distribution was modeled using all remaining subjects B and the likelihood of all subject A's part-lines was calculated; this way the

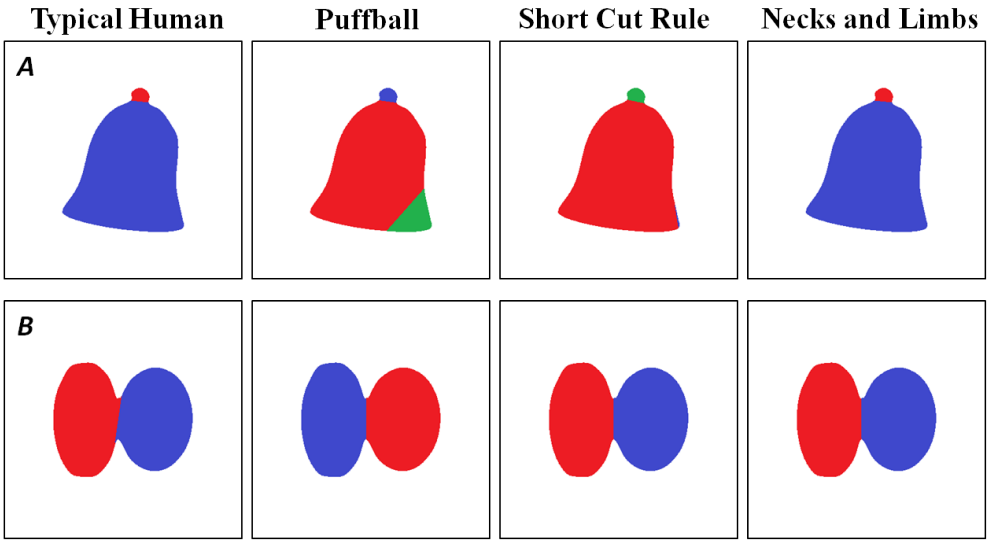

Figure 10.8: Example silhouettes on which all three algorithms performed well.

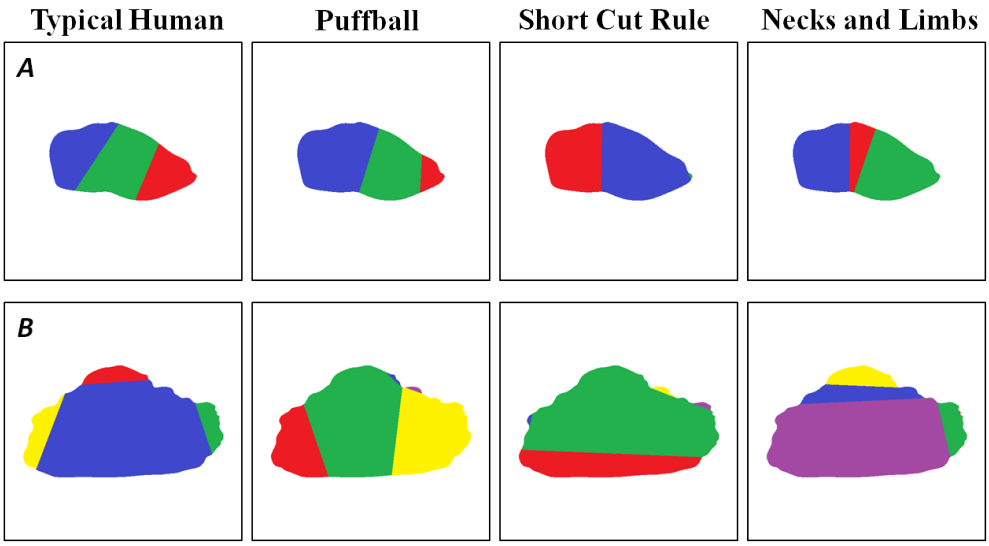

Figure 10.9: Example silhouettes on which all three algorithms performed poorly.

inter-subject likelihood of every subject's part-lines was calculated, to give a mean inter-subject part-line likelihood for each silhouette. In addition, the likelihood of all part-lines generated by each of the three competing part-segmentation algorithms was calculated using the distribution of all human part-lines for a given silhouette, giving each algorithm and silhouette a mean algorithm part-line likelihood. As the appropriate window width was not known, several values were tried, but all three values yielded the same relative ordering of the three algorithms. The averages part-line likelihood values across all silhouettes are shown in Table 10.3.

To get an idea of what these evaluations mean, let's consider a few examples. Figure 10.8 shows two examples where all three algorithms did rather well relative to mean human performance. In the case of the bell silhouette, all three algorithms easily locate the boundary between the knob at the top of the bell and the main body of the bell. The three algorithms place their second part boundary somewhat haphazardly, but many human subjects did the same; so on the whole, the performance of the algorithms is judged rather highly here. It is easy to see why performance was judged so highly on the other bi-lobed shape; the boundary between the two parts of the shape is perfectly clear, and all three approaches locate it easily. Figure 10.9, on the other hand, shows a couple silhouettes where all three algorithms were judged to perform poorly. Both silhouettes are largely convex, leading to no real global part structure; if any parts are present in these shapes, they are very subtle, very peripheral, or both. The human results were very inconsistent on these two silhouettes, and the likelihoods of the algorithm generated part-lines for all three algorithms were similarly very low.

Figure 10.10 shows a silhouette on which the likelihood of part-lines generated by the Necks and Limbs algorithm was much higher than those generated by the two competing algorithms. Nearly all human subjects gave a segmentation similar to the one shown, in which the handle of the teacup is partitioned from the main body of the teacup; the Necks and Limbs algorithm correctly identifies this part as a "limb". Puffball and the Short Cut Rule, however, do not locate this part boundary. The Short Cut Rule is drawn to a much

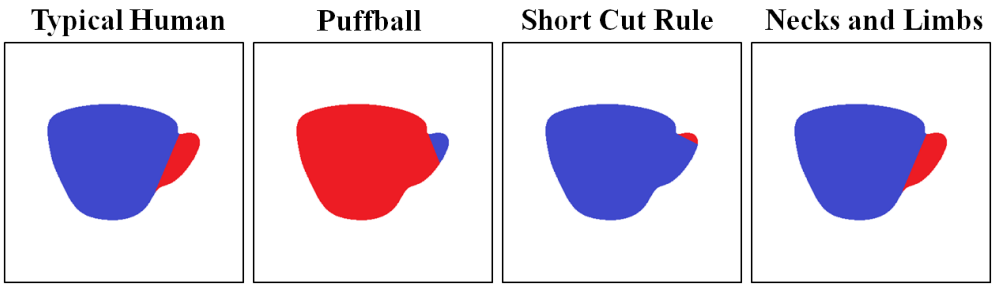

Figure 10.10: Example silhouette on which Necks and Limbs outpeformed its competitors.

shorter, but much less perceptually salient boundary, in large part because the negative curvature at the lower end of the teacup handle is so low magnitude. Puffball, on the other hand, is unable to inflate the teacup handle fully, because it does not extend out far enough; Puffball often has difficulty identifying "limb"-like parts which do not extend out very far.

Figure 10.11 shows a silhouette on which the Short Cut Rule similarly outperformed its competitors. In this case, the Short Cut Rules attraction to part-lines terminating on two curvature minima leads it to choose the vertical part-line which was also selected by the majority of human subjects. (The Short Cut Rule also chooses very non-intuitive part-line to the upper right, but this does not bring down its average likelihood below either of its competitors.) The Necks and Limbs algorithm, unable to classify the vertical part-line as either a "neck" or "limb", instead chooses two very shallow part-lines on the left pant-leg. Puffball chooses two part boundaries which would actually be quite intuitive if it were segmenting the lower half of a human form rather than simply a pair of pants. So, while it is less consistent with human subjects here, it is difficult to call the Puffball part segmentation incorrect.

Finally, Figure 10.12 shows two silhouettes in which Puffball part-lines were judged more consistent with human subjects than the two competing algorithms. In the first, Puffball gives a very intuitive segmentation, partitioning off the five arms of the starfish, as was done in the majority of human segmentations (one of which is shown here). But this silhouette presents problems for the 2D Minima Rule-based algorithms, because each inner corner of the starfish lies at the end of two part boundaries; this is why the Necks and Limbs algorithm fails to segment one of the arms, the Short Cut Rule places some boundaries very far from

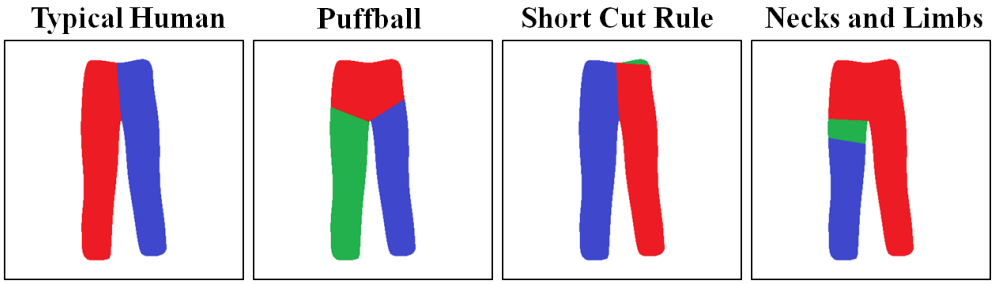

Figure 10.11: Example silhouette on which the Short Cut Rule outpeformed its competitors.

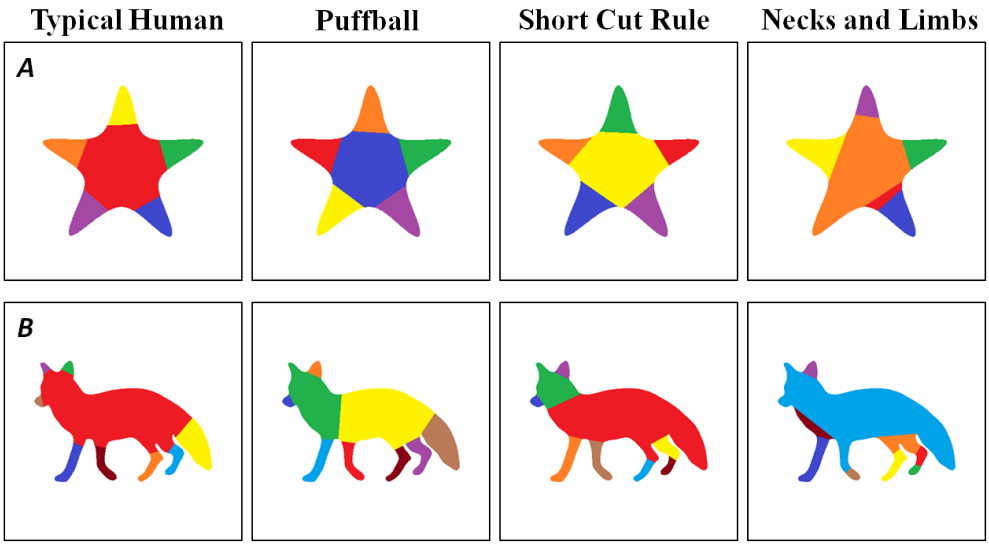

Figure 10.12: Example silhouettes on which Puffball outpeformed its competitors.

the center of the starfish. Another, more complex example of Puffball's success is shown in Figure 10.12B. In this silhouette, all three algorithms (as well as most human subjects) successfully segment the two front legs, and the larger of the two ears. But the two 2D Minima Rule-based algorithms have a great deal of trouble with the back limbs and the tail. The tail is bounded by a part boundary with only one curvature minimum, and the curvature minimum between the two legs is the endpoint of two common part boundaries in human subjects. Thus both algorithms fail to segment the tail, and either group the back legs together or only segment one of them.

Silhouette-dependent relative likelihoods were also calculated; after all, a part-line in a particular silhouette with low likelihood is less unimpressive if the average human part-line likelihood for that silhouette is also quite low, and much more unimpressive if the average human part-line likelihood for that silhouette is relatively high. So each part-lines likelihood was divided by the average likelihood of a human part-line for that silhouette; this relative likelihood could be averaged across a silhouette or all silhouettes. The overall average values are shown in Table 10.4; they produce the same ordering as the raw likelihoods.

This analysis is by no means a complete demonstration of success; the distribution of individual part lines does not tell us everything about human segmentations, and any dependence between part lines is lost in this representation. Nevertheless, Puffball does consistently

Mean Part-Line Relative Likelihood			
	$\sigma = 2.5$ pixels	$\sigma = 5$ pixels	$\sigma = 10$ pixels
Puffball	3.4798	4.8635	5.3822
Short Cut Rule	1.4027	2.5210	3.6252
Necks and Limbs	2.0461	3.1690	3.9688

Table 10.4: Average part-line relative likelihood of all three algorithms across all 88 silhouettes.

outperform the two competing algorithms; this result is robust to the smoothing parameter used to model the distribution of human part lines. So this analysis effectively demonstrates that part lines output by Puffball are numerically more consistent with human predictions than the two Minima Rule-based algorithms.

10.3.3 Comparison of Part-Line Endpoints

Given that the distribution of individual part-line endpoints is easier to model and analyze than complete segmentations, a logical next question is the behavior of individual part-line endpoints. Though this seem somewhat far removed from the behavior of overall segmentations, it is not an unreasonable question to consider, as the core principle of the 2D Minima Rule is that contour curvature minima are the foundation on which any part segmentation is built. In their analysis of human subject data, De Winter and Wagemans did perform several numerical analyses of the location of part-line endpoints, but only to test their consistency with general principles such as the 2D Minima Rule; no evaluation of complete segmentation algorithms was reported.

The analysis of part-line endpoints was performed much as the analysis of part-lines was in section 10.3.2. A distribution of end-points was modeled by placing the endpoints of all part-lines for a single silhouette from some number of subjects in a circular one-dimensional space parameterized by arc-length around the silhouette contour; this space was then filtered with a Gaussian window to estimate the distribution from which the end-points were drawn. The likelihood of an individual endpoint could then be calculated using this distribution.

Once again, each subject's part-line endpoints were evaluated with a distribution derived from all the remaining subject's part-lines for the same silhouette. Each algorithm's part-line endpoints were compared with the distribution derived from all human segmentations of a silhouette. Average likelihoods for human subjects and the three algorithms are shown in Table 10.5; average relative likelihoods, calculated as in section 10.3.2, are shown in Table 10.6.

Once again, Puffball clearly outperforms the two competing algorithms, independent of smoothing window size. The difference is not as pronounced as that of the likelihood of part-lines, but is nonetheless much more noteworthy. After all, both the Necks and Limbs algorithm and the Short Cut Rule are build on the basic assumption that however part-lines are formed and ambiguities are resolved, the endpoints of those part-lines, with some computable exceptions, can be found at minima of negative curvature. Thus, if these Minima

Mean Endpoint Likelihood ($\times 10^{-3}$)			
	$\sigma = 2.5$ pixels	$\sigma = 5$ pixels	$\sigma = 10$ pixels
Inter-Subject	2.2400	1.5381	1.0476
Puffball	4.3011	3.7656	2.9519
Short Cut Rule	3.9104	3.4367	2.7175
Necks and Limbs	3.8239	3.3727	2.6389

Table 10.5: Mean endpoint likelihood for human subjects, Puffball part segmentation, the Short Cut Rule, and the Necks and Limbs algorithm, averaged over all 88 silhouettes.

Mean Endpoint Relative Likelihood			
	$\sigma = 2.5$ pixels	$\sigma = 5$ pixels	$\sigma = 10$ pixels
Puffball	1.9202	2.4926	2.9777
Short Cut Rule	1.7798	2.3113	2.7719
Necks and Limbs	1.6696	2.1794	2.5941

Table 10.6: Average endpoint relative likelihood of all three algorithms across all 88 silhouettes.

Rule-based algorithms should outperform Puffball at any task, it would be the selection of part-line endpoints. But as we see in Tables 10.5 and 10.6, that is not the case. These results suggest, more strongly than any results thus far, that though contour minima as part-line endpoints are psychophysically consistent with human part segmentation behavior, they need not, and likely do not, play an explicit role in the visual systems computation of those part boundaries.

Chapter 11

Future Work

11.1 Improving the Evaluation Dataset

As the resuilts of our experimental evaluation show, there are considerable issues with the existing DeWinter and Wagemans dataset. First, while the silhouettes cover a wide variety of object categories and exhibit a broad range of identifiabilities, many of the silhouettes are very poor part-segmentation stimuli. These silhouettes are largely convex, and seem to exhibit little or no part structure; Figure 10.9 shows two examples. In addition, only segmentations in which at least one part-line was drawn were included; this does not allow for the possibility that the best segmentation of a silhouette is no segmentation at all. Finally, no effort has been made to evaluate the relative quality of the segmentations. Thus when algorithms are compared with the human dataset, they are penalized just as much for disagreeing with highly counter-intuitive segmentations as for disagreeing with highly intuitive ones; and many of the segmentations included in the dataset are very counter-intuitive. All of these factors make the evaluation of a part-segmentation algorithm or algorithms considerably more difficult.

For future research into the nature of part segmentation, a more robust, more controlled dataset is required. The existing dataset could be improved by a simple experiment designed to rate or evaluate the relative intuitiveness of the existing segmentations. Segmentations judged to be poor or counter-intuitive could be excluded from evaluation or simply given lower weight in the final calculation. In addition, silhouettes that yield only counter-intuitive segmentations could also be excluded as uninformative. However, it would also be very valuable to extend or replace the existing dataset; in particularly, effort should be made to generate a dataset which does not only include real-world or recognizable objects. Indeed, Siddiqi et al. (1996) demonstrated that human part segmentation is *more* consistent on unrecognizable nonsense silhouettes than on recognizable shapes; thus unrecognizable silhouettes are likely a better probe of the feed-forward mechanisms of part-segmentation.

11.2 Symmetry and Parts Analysis

The above results suggest that Puffball part segmentation is a powerful and effective model of human part segmentation; the next obvious question is why? We have seen that Puff-

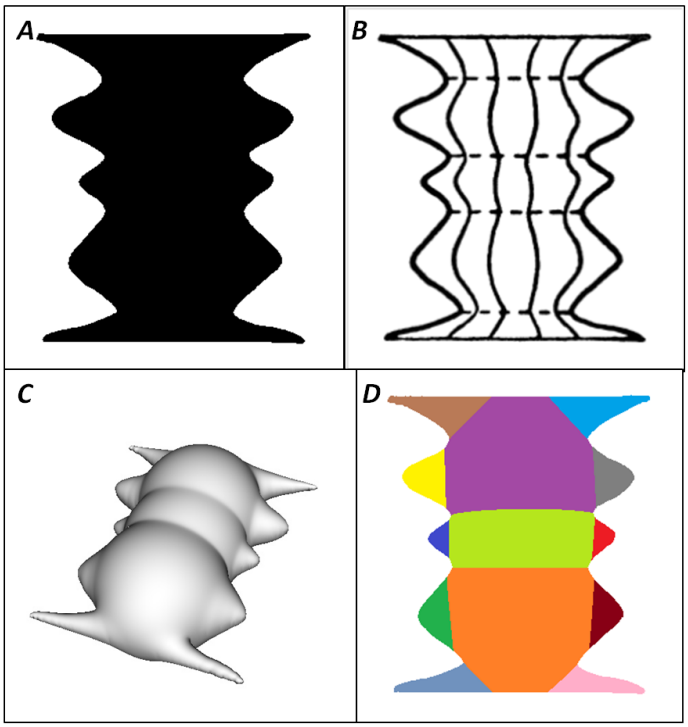

Figure 11.1: (A) A symmetric silhouette (Hoffman and Richards, 1984). (B) A three-dimensional interpretation of this shape, as a surface of revolution. (C) Puffball's inflation of the shape is unaffected by the symmetric contours, and does not yield a surface of revolution. (D) Puffballs part segmentation.

ball inflation seems to align closely with human intuition for the relationship between two-dimensional and three-dimensional shape, but this may be unrelated to its part segmentation success. Does the human visual system also use a 2D-to-3D mapping in its part analysis, or is Puffball merely a convenient mathematical shortcut to what is really a two-dimensional calculation? Put another way, does 3D shape play a role in human parts analysis of 2D shape?

To begin to answer this question, we must concentrate on those circumstances in which the agreement between Puffball inflation and our intuition about the relationship between two-dimensional shape and three-dimensional shape begins to break down. If Puffball is merely a proxy for a two-dimensional entity, then its performance in these cases should still be quite strong; if, however, Puffballs success derives from a similar approach in the visual system which uses the its own model of mapping 2D shapes to 3D regions, then performance in these cases will suffer. One class of silhouettes where this may be possible is silhouettes with a strong degree of symmetry. As mentioned in section 9.2, human subjects often interpret silhouettes with highly symmetric contours as surfaces of revolution (Figures 11.1A and 11.1B); Puffball, however, makes no such inference, instead yielding a counter-intuitive inflation with bulges and tabs (Figure 11.1C).

It is important to keep in mind that this is not strictly object recognition: the viewer has likely never seen this particular surface before, and similar interpretations can be achieved for shapes that are not as easily classified as an object. This effect appears to have more to

do with our intuitive understanding of the relationship between general shapes in the world and their contours than our experience with specific objects. It is best thought of as the result of an implicit application of the Helmholtz likelihood principle: symmetric contours in a silhouette would be highly coincidental if they were not generated by the same process, and the most likely process which generates both contours is a surface of revolution. The more complex the symmetric contour, the greater the coincidence, and the stronger the impression of a surface of revolution. But such an explanation is only meaningful if the silhouette is intuitively understood as the cross-section or projection of a three-dimensional shape. It is also worth noting that the Puffball inflation in this case is not, in any real sense, wrong. A three-dimensional shape such as the one given by Puffball is entirely consistent with the two-dimensional silhouette shown; it simply seems like a less intuitive explanation of the silhouette.

What happens to part segmentation in these circumstances? I hypothesize that when the presence of symmetric contours pushes the intuitive 3D shape away from the Puffball inflation, Puffball part segmentation will experience a similar drop in performance. To test this hypothesis, I propose an experiment to evaluate the part segmentation behavior of human subjects in the presence, or absence, of symmetry. Take, as an example, Figure 11.2A. If one assumes, as we did with Figure 9.4 that one can only draw vertical part-lines or horizontal part-lines, most existing part-segmentation models, including Puffball, would suggest that we draw vertical part-lines, separating the two side sections from the central section. This interpretation is consistent with a three-dimensional shape consisting of a thick central core with two smaller conical points extending in either direction. But suppose we instead attend

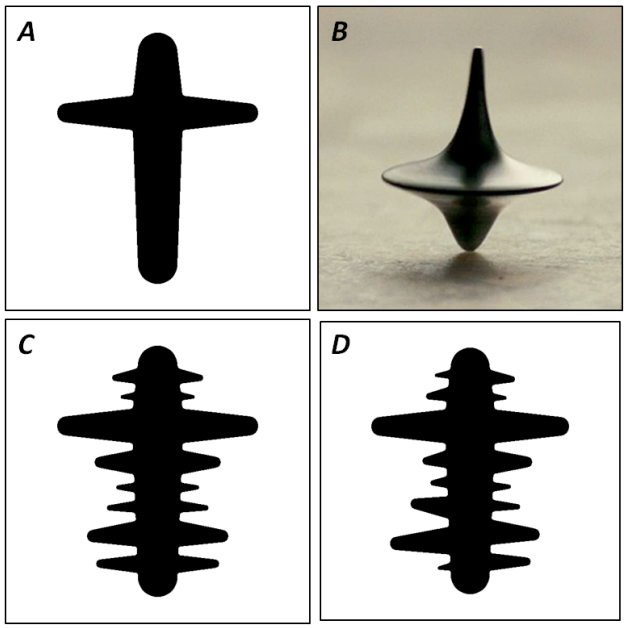

Figure 11.2: (A) A silhouette with ambiguous part structure, and ambiguous three-dimensional interpretation. (B) An illustration of how silhouette (A) could be interpreted as a surface of revolution. (C) A silhouette with more pronounced symmetry but identical local geometry. (D) A silhouette in which bilateral symmetry has been disrupted.

to the symmetry of the shape about the central vertical axis; with this in mind, an alternate three-dimensional explanation of this silhouette is a central core with a wide, tapering disk running all the way around it, like a top (Figure 11.2B). With this three-dimensional interpretation, the correct part segmentation, according to the 3D Minima Rule, is to place horizontal part boundaries, connecting the two side sections as a single part. Of course, neither three-dimensional interpretation is correct or incorrect; and neither part segmentation is correct or incorrect.

So how do we test the role that 3D interpretation might play? By influencing the effect of symmetry. Figure 11.2C shows a silhouette with identical local geometry, but much more pronounced overall symmetry; conversely, Figure 11.2D shows a silhouette in which the symmetry has been disrupted. If the above hypothesis is correct, then subjects should significantly more likely to infer horizontal part-lines in Figure 11.2C, and significantly less likely to infer horizontal part-lines in Figure 11.2D. If this result is confirmed, it will be a very strong affirmation of the hypothesis that intuitive three-dimensional interpretation plays a role in two-dimensional shape analysis and representation.

Of course, the question of how to probe human subjects' part segmentation is by no means an easy one. As the experimental results of section 10.3.1 demonstrate, human intuition about the parts of a silhouette is far from reliable. Indeed, one can see in the results of the De Winter and Wagemans study that humans can have very different conceptions of how to partition into important parts, as subjects were instructed. And this should come as no surprise. If one were asked to break an apple into parts (Figure 11.3A), one's understanding of the purpose of that segmentation would significantly affect one's choices. If one interpreted the task as identifying geometric parts of the apple shape, then the appropriate segmentation is likely to segment the stem and leaf (Figure 11.3B). But if one imagines breaking the apple into parts in everyday life, it very well might make more sense to simply remove the leaf and stem and split the apple in half, to make it easier to eat or to share with a friend (Figure 11.3C). So simply asking subjects to "break a shape into parts" is not guaranteed to yield consistent or satisfying results.

One possible solution is to give a richer context for what kind of parts we are looking for. For example, consider the following instruction: Break the shape into two pieces, so that each piece is simpler than the original shape. We have asked the subject to perform a segmentation, but we have encouraged the subject to focus not on the purpose or function of the shape, but only on its simplicity. Simplicity, though far from a well-defined concept, is

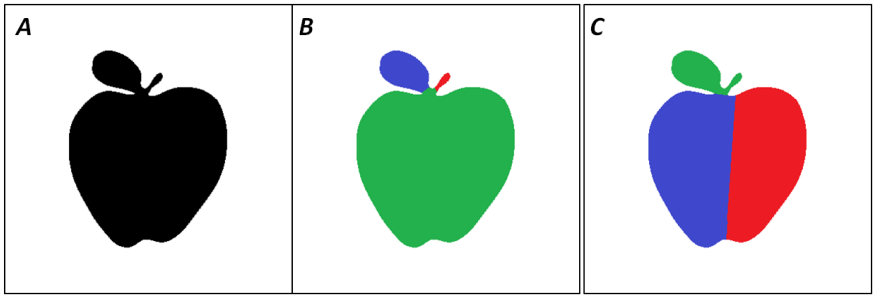

Figure 11.3: (A) An apple silhouette. (B,C) Human generated part segmentations of silhouette (A).

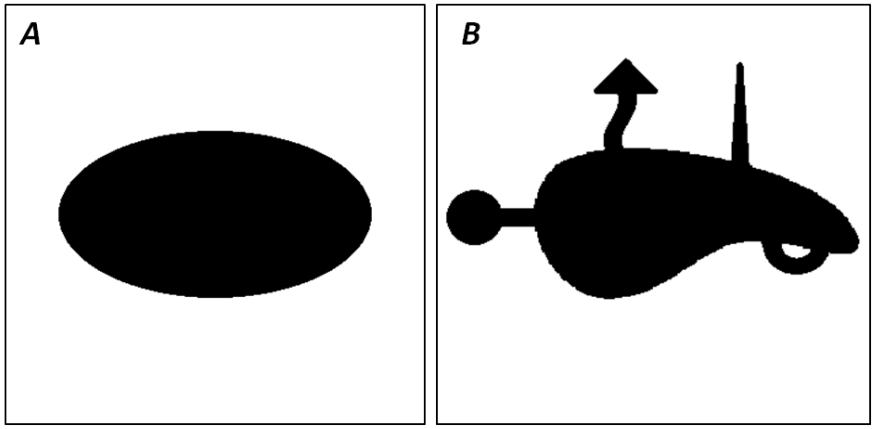

Figure 11.4: A demonstration that while we may not be able to define simplicity, we know it when we see it. (a) A simple shape. (b) A more complex shape.

certainly something we have some intuitive understanding of. Figure 11.4A is clearly simpler than Figure 11.4B, and that is likely because Figure 11.4B depicts a shape with many parts. A subject in an experiment might be asked to continue to break the pieces into smaller pieces until each piece was as simple as it could be. This would allow one to measure not only where people place their part boundaries, but in what order they select them.

11.3 Puffball and Silhouette Similarity

In the perception of silhouettes, humans are capable of perceiving differing degrees of similarity, even between abstract shapes, and such similarities can be both directly probed or tested implicitly through reaction times in shape recognition tasks. Indeed, such similarities play a very important role in the way humans relate different images as having similar content, making shape similarity a key component of the problem of content-based image retrieval, or CBIR. CBIR describes the challenge of searching for an appropriate image or images based on their inherent content; most existing image search algorithms still depend on tags, keywords or other text associated with an image file. Given the practical importance of the problem, it is no surprise that a considerable amount of work has been done on the problem of shape and silhouette similarity; proposed approaches have yielded varied but largely underwhelming success.

One problem is that almost all approaches have dealt with silhouettes as two dimensional forms, utilizing either global shape moments, or manipulations and statistics of the silhouette contour. Shape moments, though robust to many distortions and perturbations that only slightly affect the perception of a shape, are in fact too robust, filtering out almost all salient information about a shape. Measurements of the contour, though they capture almost all information present in a shape, seem poorly related to the human perceptual representation of shape. For example, if one represents a silhouette as a parametric function with curvature as a function of arc-length - a very common approach to silhouette analysis - a small perturbation of the curvature, even in a small region of the contour, can have drastic effects on the perceived shape. It was in part this gap between global moments, which are

in a sense too global, and local contour measurements, which are similarly too local, that Harry Blum was attempting to address with the MAT. The MAT, however, is too unstable to serve as a robust and reliable tool for the calculation of similarity in general.

Like the MAT, Puffball serves as a representation of the shape that lies between the purely local nature of measurements like contour curvature and the purely global measurements like shape moments; unlike the MAT, Puffball is a highly stable, continuous function of the input silhouette; small perturbations of the input silhouette yield similarly small perturbations of the inflated shape. Thus Puffball is an excellent candidate for representation of silhouettes and the measurement of their similarity. One possible form of such a representation is a map of Puffball surface normals. For example, if we inflate the shape shown in Figure 11.5A, and calculate the surface normal of the resulting inflated shape, we get a map in which every point is associated with a set of values related to where that point lies in the shape (Figure 11.5B). Points near the center of the shape have a near vertical surface normal, while points closer to the edge have surface normals closer to the have surface normals further from the line of sight, pointing in the direction those points lie away from the center.

The advantage of this map is that it assigns to every interior pixel of the silhouette a value which carries perceptually relevant information about the shape. So, when calculating similarity, instead of analyzing and comparing two very sparse binary signals, we are comparing two rich and highly informative images; rather than looking for deformations of contours leading from one silhouette to another, we can look for deformations of normal maps, which will constrain the intervening shapes to themselves be perceptually similar.

To investigate this approach to shape similarity, it would be best to begin with simple convex silhouettes. Though working only with convex silhouettes might seem too constrained, it will allow us to eliminate the influence of part structure. Part structure undoubtedly plays a key role in the representation, recognition, and similarity of shapes, but will add considerable complexity and difficulty to the problem of modeling shape similarity. It is thus best to begin with single part shapes; the best way to ensure that shapes have only one part is to ensure that all contour curvature is non-negative; hence, only convex silhouettes. However, even in this constrained silhouette domain, I believe that a great deal of insight

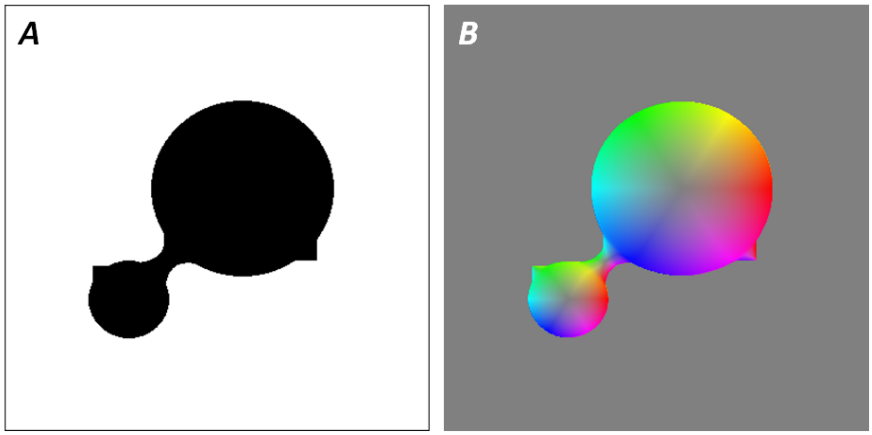

Figure 11.5: (A) An arbitrary silhouette. (B) The Puffball surface normal map of silhouette (A). Saturation represents the slant of the surface normal, and hue represents the tilt.

about shape similarity can be obtained; and I believe that the Puffball surface normal map may be an excellent candidate representation for unpacking the structures that the human visual system uses to perform these tasks.

Conclusion

Throughout this thesis, we have seen the critical differences between two senses of the word "model": the classical scientific sense and the more modern computational sense. The classical scientific model can be found throughout the fields of physics and chemistry: Newton's laws of motion and gravitation, Maxwell's equations of electromagnetism, Mendeleev's periodic table. Each of these scientists saw a pattern in nature, and posited an equation or structure which seemed to capture that pattern; these models each carry remarkable predictive power, and have been pivotal to the development and progress of their fields.

It was the power of these models that drove the earliest perceptual scientists to try and transplant their success into the study of vision and the study of the mind as a whole. Weber's law, Fechner's law, the Gestalt laws, and the creation of psychophysics (a revealing name choice if ever there was one) were all built around the hopes of unlocking the fundamental regularities of the mind; even modern researchers continue to describe constraints on Gestalt grouping as "forces" of "attraction" (Kubovy and Wagemans, 1995). But the study of grouping is not like the study of physics, and models of Gestalt grouping cannot work in the same way. The visual system is not a pocket universe with its own peculiar physics; the mind is a haphazard, complex, messily assembled network of computational tricks and inferential shortcuts. Over the millions of years in which it has developed, some computational structures that worked have been preserved, while those that haven't (and some that have) have fallen by the wayside. It is not enough to understand the physics of vision; we must also understand the engineering of the mind.

The study of silhouette parts shows another example of this disconnect. At its heart, the 2D Minima Rule proposed by Hoffman and Richards arose because the authors observed an apparent regularity in the visual world: part boundaries on three-dimensional shapes usually lie in areas of negative principal curvature; in the two dimensional projections of three-dimensional shapes, these areas of negative principal curvature appear as stretches of negative contour curvature; hence, part boundaries in two-dimensional shapes will tend to terminate at points of high negative contour curvature. This hypothesis was proposed, and psychophysical data confirmed that humans indeed placed part boundaries such that they very often terminated at or near minima of negative curvature. So why then has the 2D Minima Rule not succeeded?

The answer is because it is a scientific model, and not a computational one. Hoffman and Richards stipulate that a good model must not only be reliable and versatile, but computable; but while the contour minima themselves are easily found, the computability of the 2D Minima Rule stops there. Though the pattern described by the Minima Rule appears to be an accurate one, its computational power is very limited. The 2D Minima Rule is an

accurate description of what our visual system does; but it tells very little about what our visual system *will* do.

Puffball part segmentation, on the other hand, was not derived from first principles or inferred from an ideal observer; on the contrary, we came across the Puffball approach to part segmentation largely by accident. There was no reason a priori to believe that it would work well, but we tried it because as a tool Puffball was powerful enough to capture something important about shape, and computationally simple enough to be easy to experiment with. This process drew as much from engineering as it did from science, and more insights undoubtedly await the development of similarly unexpected innovations. Effort should be expended exploring unusual ideas, trying out computational tricks and playing with hacks. From such explorations, a wide array of valuable and powerful computational ideas have been - and will continue to be - discovered; but until recently, these explorations have remained largely the domain of engineers and computer scientists, and the results have often been ignored or put aside by those more invested in the scientific underpinnings of the human mind.

This is not, however, merely a matter of short-sightedness; for just as scientific thought without engineering cannot give us insight in to the computational structure of the mind, goal-driven engineering without consideration of the scientific knowledge and investigation of the visual system will leave us equally lost. Computational models which are too complex, too opaque, or too rigid will offer little scientific insight, even if their performance is unmatched. Development of the Gestalt grouping model framework was no small challenge, but an equally important and equally challenging step in my research was the development of a way to compare the results of that model with measurable human behavior. Puffball is a powerful and versatile inflation tool, but it is not the only, or even the most effective or efficient inflation tool available; its true value is derived from its conceptual simplicity and well-defined mathematical underpinning. This marriage of the two approaches, human vision and computer vision, is a relatively new approach in the world of vision research; but it is only through such efforts, combining creative computational ideas with scientifically and psychophysically grounded evaluation, that we can hope to truly understand the deeper aspects of the human visual system, and the human mind.

Bibliography

Abbasi, S., Mokhtarian, F., and Kittler, J. (1999). Curvature scale space image in shape similarity retrieval. *Multimedia Systems*, 7:467–476.

Alexe, A., Gaildrat, V., and Barthe, L. (2004). Interactive modelling from sketches using spherical implicit functions. In *International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61:183–193.

Babaud, J., Witkin, A. P., Baudin, M., and Duda, R. O. (1986). Uniqueness of the gaussian kernel for scale-space filtering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(1):26 –33.

Basri, R., Costa, L., Geiger, D., and Jacobs, D. (1998). Determining the similarity of deformable shapes. *Vision Research*, 38:2365–2385.

Beck, J., Rosenfeld, A., and Ivry, R. (1989). Line segregation. *Spatial Vision*, 4(2-3):75–101.

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

Blum, H. (1967). A transformation for extracting descriptors of shape. In *Models for the Perception of Speech and Visual Forms*, pages 362–380. MIT Press.

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

Braunstein, M. L., Hoffman, D. D., and Saidpour, A. (1989). Parts of visual objects: An experimental test of the minima rule. *Perception*, 18:817–826.

Callaghan, T. (1989). Interference and dominance in texture segregation: Hue, geometric form, and line orientation. *Attention, Perception, & Psychophysics*, 46:299–311.

Claessens, P. and Wagemans, J. (2005). Perceptual grouping in gabor lattices: Proximity and alignment. *Attention, Perception, & Psychophysics*, 67:1446–1459.

Comanciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence*, 24:603–619.

De Winter, J. and Wagemans, J. (2006). Segmentation of object outlines into parts: A large-scale integrative study. *Cognition*, 99(3):275–325.

Ehrenfels, C. (1937). On gestalt-qualities. *Psychological Review*, 44:521–524.

Elder, J. and Zucker, S. (1993). The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vision Research*, 33(7):981–991.

Elder, J. and Zucker, S. (1996). Computing contour closure. In Buxton, B. and Cipolla, R., editors, *Computer Vision ECCV '96*, volume 1064 of *Lecture Notes in Computer Science*, pages 399–412. Springer Berlin / Heidelberg.

Elder, J. H. and Goldberg, R. M. (2002). Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2(4):324–353.

Ernst, U. A., Mandon, S., SchinkelBielefeld, N., Neitzel, S. D., Kreiter, A. K., and Pawelzik, K. R. (2012). Optimality of human contour integration. *PLoS Comput Biol*, 8(5):e1002520.

Feldman, J. (1997). Curvilinearity, covariance, and regularity in perceptual groups. *Vision Research*, 37(20):2835–2848.

Field, D. J., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local "association field". *Vision Research*, 33(2):173–193.

Freeman, W. and Adelson, E. (1991). The design and use of steerable filters. *Pattern Analysis and Machine Intelligence*, 13:891–906.

Garcia, R., Llibre, J., and Sotomayor, J. (2006). Lines of principal curvature on canal surfaces in r. *Anais da Academia Brasileira de Ciências*, 78(3):405–415.

Geisler, W., Perry, J., Super, B., and Gallogly, D. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41(6):711–724.

Gigus, Z. and Malik, J. (1991). Detecting curvilinear structure in images. Technical report, Berkeley, CA, USA.

Grosof, D. H., Shapley, R. M., and Hawken, M. J. (1993). Macaque v1 neurons can signal 'illusory' contours. *Nature*, 365:550–552.

Grossberg, S. and Mingolla, E. (1985). Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psychological Review*, 92(2):173–211.

Hochberg, J. and Hardy, D. (1960). Brightness and proximity factors in grouping. *Perceptual and Motor Skills*, 10:22.

Hochberg, J. and Silverstein, A. (1956). A quantitative index of stimulus-similarity: Proximity vs. difference in brightness. *The American Journal of Psychology*, 69(3):456–458.

Hoffman, D. and Richards, W. (1984). Parts of recognition. *Cognition*, 18(1-3):65 – 96.

Hoffman, D. D. and Singh, M. (1997). Salience of visual parts. *Cognition*, 63(1):29 – 78.

Igarishi, T., Matsuoka, S., and Tanaka, H. (1999). Teddy: a sketching interface for 3d freeform design. In *International Conference on Computer Graphics and Interactive Techniques: ACM SIGGRAPH Courses*.

Jacobs, D. (1996). Robust and efficient detection of salient convex groups. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(1):23–37.

Karpenko, O., Hughes, J. F., and Raskar, R. (2002). Free-form sketching with variational implicit surfaces. *Computer Graphics Forum*, 21(3):585–594.

Karpenko, O. A. and Hughes, J. F. (2006). Smoothsketch: 3d free-form shapes from complex sketches. *ACM Transactions on Graphics*, 25(3):589–598.

Kellman, P. J. and Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cognitive Psychology*, 23(2):141–221.

Kimia, B. B., Tannenbaum, A. R., and Zucker, S. W. (1995). Shapes, shocks, and deformations i: The components of two-dimensional shape and the reaction-diffusion space. *International Journal of Computer Vision*, 15:189–224.

Koenderink, J. (1984). The structure of images. *Biological Cybernetics*, 50:363–370.

Koenderink, J. J. and van Doorn, A. J. (1982). The shape of smooth objects and the way contours end. *Perception*, 11:129–137.

Koffka, K. (1922). Perception: an introduction to the gestalt-theorie. *Psychological Bulletin*, 19:531–585.

Köhler, W. (1929). *Gestalt Psychology*. Oxford England: Liveright.

Kubovy, M., Holcombe, A. O., and Wagemans, J. (1998). On the lawfulness of grouping by proximity. *Cognitive Psychology*, 35(1):71–98.

Kubovy, M. and van den Berg, M. (2008). The whole is equal to the sum of its parts: A probabilistic model of grouping by proximity and similarity in regular patterns. *Psychological Review*, 115(1):131–154.

Kubovy, M. and Wagemans, J. (1995). Grouping by proximity and multistability in dot lattices: A quantitative gestalt theory. *Psychological Science*, 6(4):225–234.

Landy, M. S. and Bergen, J. R. (1991). Texture segregation and orientation gradient. *Vision Research*, 31(4):679–691.

Latecki, L. J. and Lakämper, R. (2000). Shape similarity measure based on correspondence of visual parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1185–1190.

Ledgeway, T., Hess, R. F., and Geisler, W. S. (2005). Grouping local orientation and direction signals to extract spatial contours: Empirical tests of "association field" models of contour integration. *Vision Research*, 45(19):2511–2522.

Lowe, D. G. (1985). *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Norwell, MA, USA.

Lowe, D. G. (1989). Organization of smooth image curves at multiple scales. *International Journal of Computer Vision*, 3:119–130.

Malik, J., Belongie, S., Leung, T., and Shi, J. (2001). Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43:7–27.

Marr, D. and Hildreth, E. (1980). Theory of edge detection. In *Proceedings of the Royal Society of London*, volume 207, pages 187–217.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 416–423.

Meilă, M. (2007). Comparing clusterings: an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Mi, X. and DeCarlo, D. (2007). Separating parts from 2d shapes using relatability. *Computer Vision, IEEE International Conference on*, 0:1–8.

Mi, X., DeCarlo, D., and Stone, M. (2009). Abstraction of 2d shapes in terms of parts. In *Proceedings of the 7th International Symposium on Non-Photorealistic Animation and Rendering*, NPAR '09, pages 15–24, New York, NY, USA. ACM.

Moulden, B. (1994). *Collator Units: Second-Stage Orientational Filters*, pages 170–192. John Wiley & Sons, Ltd.

Op de Beeck, H. P., Wagemans, J., and Vogels, R. (2008). The representation of perceived shape similarity and its role for category learning in monkeys: A modeling study. *Vision Research*, 48(4):598–610.

Oyama, T. (1961). Perceptual grouping as a function of proximity. *Perceptual and Motor Skills*, 13:305–306.

Oyama, T., Simizu, M., and Tozawa, J. (1999). Effects of similarity on apparent motion and perceptual grouping. *Perception*, 28(6):739–748.

Parent, P. and Zucker, S. (1989). Trace inference, curvature consistency, and curve detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(8):823–839.

Paris, S. and Durand, F. (2007). A topological approach to hierarchical segmentation using mean shift. *Computer Vision and Pattern Recognition*, 0:1–8.

Pentland, A. P. (1990). Automatic extraction of deformable part models. *International Journal of Computer Vision*, 4:107–126.

Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(7):629–639.

Quinlan, P. T. and Wilton, R. N. (1998). Grouping by proximity or similarity? competition between the gestalt principles in vision. *Perception*, 27(4):417–430.

Rosenholtz, R., Twarog, N. R., Schinkel-Bielefeld, N., and Wattenberg, M. (2009). An intuitive model of perceptual grouping for hci design.

Rush, G. (1937). Visual grouping in relation to age. *Archives of Psychology (Columbia University)*, 217:1–95.

Scassellati, B., Alexopoulos, S., and Flickner, M. (1994). Retrieving images by 2d shape: A comparison of computation methods with human perceptual judgments. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, pages 2–14.

Sharon, E. and Mumford, D. (2006). 2d-shape analysis using conformal mapping. *Int. J. Comput. Vision*, 70:55–75.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence*, 22:888–905.

Siddiqi, K. and Kimia, B. B. (1995). Parts of visual form: Computational aspects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:239–251.

Siddiqi, K., Tresness, K. J., and Kimia, B. B. (1996). Parts of visual form: pscyhophysical aspects. *Perception*, pages 399–424.

Sigman, M., Cecchi, G. A., Gilbert, C. D., and Magnasco, M. O. (2001). On a common circle: Natural scenes and gestalt rules. *Proceedings of the National Academy of Sciences*, 98(4):1935–1940.

Singh, M. and Fulvio, J. M. (2005). Visual extrapolation of contour geometry. *Proceedings of the National Academy of Sciences*, 102(3):989–944.

Singh, M., Seyranian, G., and Hoffman, D. (1999). Parsing silhouettes: The short-cut rule. *Attention, Perception, & Psychophysics*, 61:636–660.

Smits, J. T., Vos, P. G., and Van Oeffelen, M. P. (1985). The perception of a dotted line in noise: a model of good continuation and some experimental results. *Spatial Vision*, 1(2):163–177.

Tai, C.-L., Zhang, H., and Fong, C.-K. (2004). Prototype modeling from sketched silhouettes based on convolution surfaces. *Computer Graphics Forum*, 23(1):71–83.

Terzopoulos, D., Witkin, A., and Kass, M. (1987). Symmetry-seeking models and 3d object recognition. *International Journal of Computer Vision*, 1(3):211–221.

Terzopoulus, D. and Witkin, A. (1988). Physically based models with rigid and deformable components. *IEEE Computer Graphics and Applications*, 8(6):41–51.

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

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Graphics Pr.

Ullman, S. and Sha"ashua, A. (1988). Structural saliency: The detection of globally salient structures using a locally connected network. Technical report, Cambridge, MA, USA.

von der Heydt, R., Peterhans, E., and Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, 224(4654):1260–1262.

Watt, R., Ledgeway, T., and Dakin, S. C. (2008). Families of models for gabor paths demonstrate the importance of spatial adjacency. *Journal of Vision*, 8(7):1–19.

Wertheimer, M. (1923). Investigation of the principles of gestalt ii. *Psychological Research*, 4:301–350.

Witkin, A. P. (1983). Scale-space filtering. In *Proceedings of the Eighth international joint conference on Artificial intelligence - Volume 2*, IJCAI'83, pages 1019–1022, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Xu, Z., Feng, R., and guang Sun, J. (2006). Analytic and algebraic properties of canal surfaces. *Journal of Computational and Applied Mathematics*, 195(1-2):220–228.

Yen, S.-C. and Finkel, L. H. (1998). Extraction of perceptually salient contours by striate cortical networks. *Vision Research*, 38(5):719–741.

Appendix A

Dot Placement Algorithm

The algorithm began with two images, D and I. I was an binary indicator function which was updated throughout the process to denote which pixels could still be selected as dot centers; it began at value 1 for all pixels. D was a distance map which represented the *proximity parameter* at every pixel in the image domain: the minimum distance a dot placed at that point could be from any other dot. If the dots were placed uniformly, D would be the same everywhere.

In each cycle of the dot placement process, one of the remaining pixels, q, where the value of I was 1 was selected at random to be the center of a dot. This pixel q was added to a list of dot centers. Then every pixel p such that $d(p, q) < \max(D(p), D(q))$ had the value of $I(p)$ set to zero. This process repeated until I was zero everywhere with in a central region of the image. The result was a placement of dots such that every pixel p was less than $D(p)$ pixels away from at least one dot, and no dot placed at a pixel q was less than $D(q)$ pixels away from any other dot.

An illustration of such an image is shown in Figure A.1; this image was generated with two different proximity parameters denoted by red and blue. Note that the circles fully cover the central area of the image, but no dot center lies within another dot's circle.

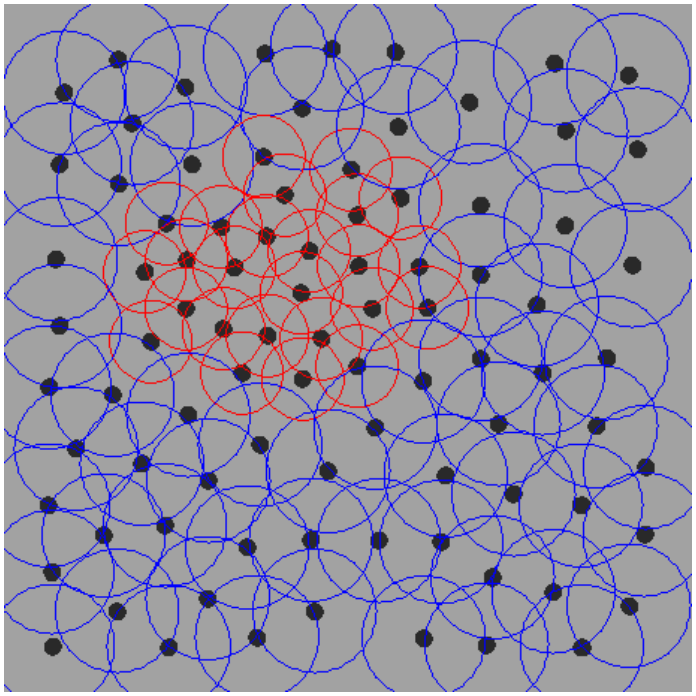

Figure A.1: An illustration of the dot placement process.

Appendix B

MATLAB Implementation of Puffball

```
function h = Puffball(mask)
    % CALCULATE GRASSFIRE HEIGHT FUNCTION %
    % A 3x3-tap filter to smoothly erode an anti-aliased edge
    fil = [0.1218 0.4123 0.1218; 0.4123 0.9750 0.4123; ...
        0.1218 0.4123 0.1218]/1.2404;
    nmask = double(mask);
    surf = zeros(size(mask));
    while ~isempty(find(nmask,1))
        surf = surf+nmask/1.67; % Each iteration erodes the edge .6 pixels
        nmaskpad = padarray(nmask,[1 1],'replicate');
        nmaskpad = conv2(nmaskpad,fil,'same')-1.4241;
        nmask = max(min(nmaskpad(2:end-1,2:end-1),1),0);
    end

    % LOCATE THE MEDIAL AXIS %
    [dx dy] = gradient(surf);
    dsurf = sqrt(dx.^2+dy.^2);
    % Medial axis points have a grassfire gradient measurably less than 1
    matr = bwmorph(dsurf<0.958&surf>2,'skel',Inf).*surf;

    % TAKE THE UNION (SOFT-MAX) OF MAXIMAL SPHERES %
    [X Y] = meshgrid(1:size(mask,2),1:size(mask,1));
    h = ones(size(mask));
    [y x] = find(matr);
    for i = 1:length(f)
        r = matr(y(i),x(i))^2 - (X-y(i)).^2 - (Y-x(i)).^2;
        h(r>0) = h(r>0)+exp(sqrt(r(r>0)));
    end
    h = log(h);
end
```

Box 1: MATLAB code implementing Puffball inflation.

Appendix C

Implementation of the Short Cut Rule

To implement the Short Cut Rule, I endeavored to follow three constraints on part-line choice described in Singh et al. (1999). First, that part-lines must cross an axis of local symmetry; second, that if two part-lines are in conflict, one should choose the one that is shorter; and three, that a sufficiently short part-line may be drawn even if it terminates at only one curvature minimum.

To implement these constraints, the first task was to locate the curvature minima, and hence to calculate the contour curvature. To do this, the edge pixels of a binay silhouette image were located, and arranged into a single, cyclical chain which proceeded counterclockwise around the shape. (We shall assume that the silhouette contour is simply connected; but the algorithm can be easily adjusted to handle shapes of more complex topology.) In this chain steps between fully adjacent pixels were given a value of 1, while steps between diagonally adjacent pixels were given a value of $\sqrt{2}$. These values were used to convert the edge to a parametric function of t, where each step corresponded to an increase in t of the appropriate size. This discrete parametric function was then fit at every point in a local Gaussian window ($\sigma = 5$ pixels), so that at every edge point the functions $x(t)$ and $y(t)$ were estimated as quadratic functions of t; these functions were then used to estimate the local orientation and curvature. Curvature minima were those edgepoint with curvature lower than the points around them and below some threshold curvature, which could be adjusted to change the coarseness or fineness of the distribution.

Next was the selection of the possible part-lines. Every pairing of a curvature minimum and any other edgepoint was considered a potential part-line, to allow for the possibility of short, one-minimum part-lines. Pairs of points such that the line passing between those points was not entirely within the shape were eliminated, as were pairs of points such that the line between them did not cross exactly one axis of local symmetry (represented by the medial axis, calculated as in Appendix B). All remaining possible part-line were ranked according to their adjusted length; lengths were adjusted according to the curvature at the two endpoints, such that a line between two highly concave endpoints would have its length cut by as much as a factor of two, while a line terminating at a point of highly convex curvature would have a very high adjusted length. This adjustment ensured that part-lines between two curvature minima would still be chosen more often than equally short part-lines with only one curvature minimum.

Once the part-lines had been collected and ranked, an iterative process began. In each iteration, the top-ranked remaining valid part-line was selected, and added to a list of part-line. Of the remaining part-lines, any lines that were in conflict with the recently selected part-line were removed. Two part-lines were considered in conflict if they intersected, or terminated at or near the same point; of course, we have seen that in reality part-lines can terminate at the same point, but without this constraint the Short Cut Rule produces an unreasonable number of part-lines and greatly oversegments. Once all conflicting part-lines had been removed from the potential part-line list, the process repeated. This continued until no valid part-lines remained.