SKELETAL SHAPE SIMILARITY AND SHAPE CLASSIFICATION

by

NATHAN DESTLER

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Psychology

Written under the direction of

Jacob Feldman

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2019

ABSTRACT OF THE DISSERTATION

Skeletal Shape Similarity and Shape Classification

By NATHAN DESTLER

Dissertation Director:

Jacob Feldman

What makes two shapes similar? Given two shapes, is there a mathematically principled way to predict human similarity judgments, and consequent shape classification judgments? To answer this question, an experimental framework was developed for rapidly collecting human judgments of whether shapes did or not not belong to novel shape categories. The subjects' judgments of category membership were compared against the predictions of several different shape similarity/classification models. Among these models, I propose a new *lattice similarity* model of shape similarity based on Bayesian shape skeletons. An earlier model of similarity based on the same Bayesian shape skeletons, the cross likelihood, has been shown to be an effective predictor of human shape discrimination, but this model applies only to shapes with similar part structures. The lattice similarity model is more principled and general, and is suitable for comparing arbitrary shape pairs in both 2D and 3D domains. This new model provides a better overall fit to human data than a number of competing models, including the cross likelihood model, an out-of-the-box convolutional neural network model, and a non-skeletal part-based similarity model proposed by Erdogan and Jacobs (2017). The lattice similarity model predicted human data more accurately than all other tested models in experiments that used 3D shapes, and most other models in experiments that used 2D shapes. Prototype-like and exemplar-like versions of the lattice similarity model were also compared using the same human data as above; the prototype-like version fits the experimental data better than the exemplar-like version.

Acknowledgements

To my advisers, Jacob Feldman and Manish Singh, for all the knowledge, wisdom, and support you have offered me over these years. I couldn't have asked for better guides on this journey.

To the other members of my committee, Melchi Michel and Ahmed Elgammal, for both your editorial feedback and for being instrumental in teaching me the computer vision and computational modelling skills I needed to get this far. When I struggled, each of you proved yourselves kind and attentive teachers.

To my labmates past and current, Vicky Froyen, Seha Kim, O. Dağlar Tanrikulu, Xiaoli "Lily" He, Serena De Stafani, and Ryne Choi, for walking beside me, and in some cases paving the way, as I walked this road.

To Anne Sokolowsky, for your kind support that was matched only by your invaluable expertise in helping me navigate university systems and requirements. I don't know how I would've made it here without you.

To my parents, Bill Destler and Rebecca Johnson, for being inspiring, interested, supportive, and always, always loving.

To my girlfriend, Marija Pushko, and my close friend, Cybil Artimovich, for keeping me grounded and motivated during the final stretch, and for listening to many a rant about code that wasn't working. You're both amazing, and I can only hope to one day repay the support you've given me.

And to the NIH and NSF, for your generosity in funding the early stages of this work.

From the bottom of my heart, thank you all.

Table of Contents

Chapter 1

Introduction

In order for the visual system to function, it must be able to identify shapes in the environment. A pie is round, or sometimes wedge-shaped; a book is rectangular; and a dog is a complex network of elongated parts. Although each of these object labels (pie, book, dog) contains far more semantic meaning than a simple shape, the shape-level identifications are a key step in the process that transforms patterns of light striking the retina into a complex visual scene containing complete perceptual objects such as pies and dogs (Kellman, Garrigan, & Erlikhman, 2015; Wagemans, 2015).

But what is a shape? How is it represented by the brain? And what makes two shapes similar to or different from one another? As yet, there are no concrete, universally-accepted answers. Although there are a number of human shape processing models in the literature, shape has proven difficult to concretely define.

Here, I will focus on shape similarity, and on how shape similarity is used to group shapes into categories. I will examine similarity and categorization outside the realm of known, nameable shapes, so as to allow for a focus on raw shape properties before semantics become involved in the categorization process.

1.1 What makes two shape similar?

Take a look at the two shapes shown in Fig. 1.1. Are these two shapes similar or different? If they're similar, how similar are they? If you are told one of them is a "blicket," would you think the other one was also a blicket? While humans can form these sorts of judgments, and some models can predict these judgments under specific circumstances (Briscoe, 2008), a full model of shape similarity and its components remains elusive.

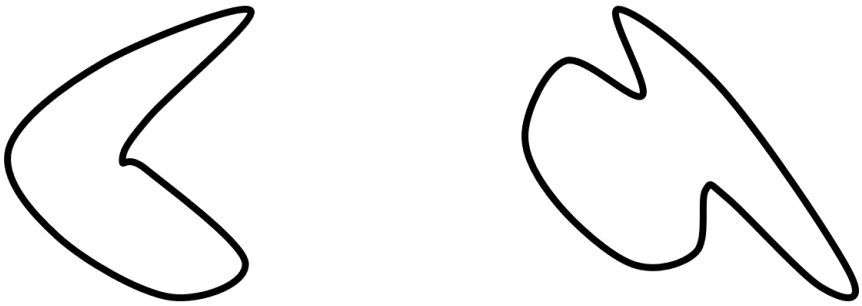

Figure 1.1: Two 2D shapes. How similar are they?

In both human and computer vision research, there are a number of measures used to compute shape similarity, including contour-based models, which compute similarity based on the geometry of the shape's bounding contour (Belongie, Malik, & Puzicha, 2002; Blake & Isard, 2012; Mustafa, Pugeault, & Krüger, 2013), skeleton-based models, which infer an internal part-based structure for each shape and compute similarity from there (Singh, Papanikolopoulos, & Cherkassky, 1998; Sebastian & Kimia, 2005; Feldman & Singh, 2006; Briscoe, 2008; Rezanejad & Siddiqi, 2015; Ayzenberg, Chen, Yousif, & Lourenco, 2019; Ayzenberg & Lourenco, 2019), and image-based models, which involve evaluating the raw properties of the pixel-value image and classifying based on those, often using state-of-the-art neural network algorithms (Farabet et al., 2010; Cireşan, Meier, Masci, & Schmidhuber, 2012; Simonyan & Zisserman, 2014; Radenovic, Tolias, & Chum, 2018). The next section will review each of these model classes in turn.

1.2 Types of shape models

1.2.1 Contour-based shape models

Perhaps the most intuitive class of shape models is contour-based models. These models represent the shape using properties of its bounding contour. Fig. 1.1 shows the bounding contours of two shapes and asks the reader to evaluate their similarity. One obvious way to compute this similarity is to compare the shape contours directly. In the simplest of these models, one might perform simple geometric operations, such as scaling the shapes to match in size and orientation and then evaluating the degree

to which their boundaries overlap.

Attneave (1954) famously observed that areas of high contour curvature contained large amounts of information about the shapes (a claim that has since been experimentally verified by Biederman, 1987 and Norman, Phillips, & Ross, 2001, and formalized by Feldman & Singh, 2005). These curvature extrema can be used to identify key shape locations like subjective part boundaries, which may play a key role in shape similarity judgments (Hoffman & Richards, 1984; Biederman, 1987; de Winter & Wagemans, 2006).

Yet other models compute similarity by the degree of deviation from one shape's contour to the other's. Blake and Isard (2012), Belongie et al. (2002), and Greene (2018) built and tested computer models that matched like points along the contours of each shape, and computed dissimilarity according to the error in the positions of matched points. For example, in the Blake and Isard active contour model, shapes are stored in memory as an ideal contour prototype plus a probabilistic error function. Classification, then, proceeds in a Bayesian fashion.

$$p(\text{IDEAL}|\text{OBSERVED}) \propto p(\text{IDEAL})p(\text{OBSERVED}|\text{IDEAL})$$

In this active contour approach, shape classification is presented as a curve-fitting problem, assessing the probability cost of fitting the observed shape contour to each candidate contour prototype. Note that this model, like all Bayesian models, requires the observer to have a prior estimate of the probability probability of seeing the shape in the environment. This is largely an advantage, as it allows observers to incorporate contextual knowledge when making shape classification judgments. After all, a cylindrical object tapered at the top is more likely to be a bottle when it's found in a liquor cabinet, and more likely to be a vase when it contains flowers.

However, these contour-based models are not without problems. In general, contour-based models account poorly for shapes that can and do articulate. Articulating shapes, such as the human arm which can bend at the elbow, are common in the real world. These articulations change the bounding contour of the shape dramatically, but humans appear to have little difficulty maintaining constant percepts in spite of articulations

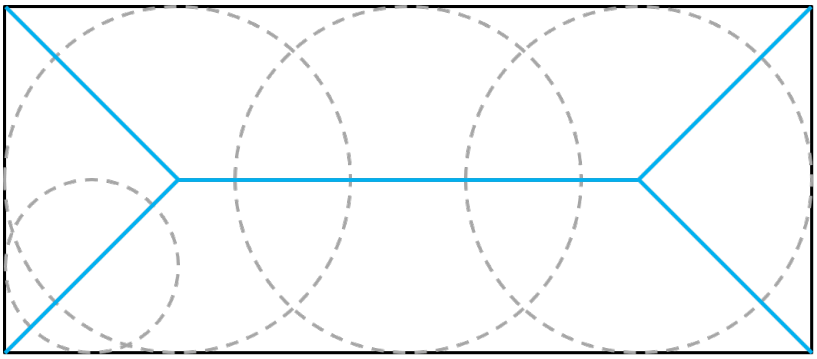

Figure 1.2: Blum's (1978) Medial Axis Transform of a rectangular shape. The shape contour is shown in black and the medial axes are shown in blue. Several of the shape's medial disks, whose centers define the medial axes, are shown in dashed gray. While only four are shown here for illustration purposes, there are an infinite number of these medial disks which are maximally-inscribed within the shape contour.

(Denisova, Feldman, Su, & Singh, 2016). While some contour-based models, particularly those based on Attneave's curvature extrema observation, can begin to account for these articulations (Barenholtz & Tarr, 2008), most contour models simply cannot maintain shape constancy across articulations.

1.2.2 Skeleton-based shape models

Another class of shape models, motivated in part by some of the weaknesses of the simpler contour-based models, is skeletal models of shape. Shape skeleton approaches characterize shapes as a series of parts, each of which is represented by an "axis" of the skeleton. Because each axis corresponds to a part, these skeletal models tend to reflect subjective human part decompositions. Additionally, shape skeletons can articulate along axis/part connection points while preserving almost all of their properties.

Most shape skeleton models are based on the Medial Axis Transform, or MAT (Blum & Nagel, 1978). In the MAT, the skeleton is computed by finding the centers of maximally-large disks inscribed within the shape's contours. The centers of all such maximal disks form the shape's medial axis (see Fig. 1.2). However, the MAT is also unstable to changes in the shape contour, producing substantially different skeletons as a result of only small contour perturbations (Blum & Nagel, 1978; August, Siddiqi, & Zucker, 1999; Katz & Pizer, 2003).

There are two main classes of solutions to this instability problem. The first class, put forward by August et al. (1999) and Kovács, Fehér, and Julesz (1998), is to characterize the medial points according to their relationship with the shape contour. This class of solutions systematically marks some axes of the shape skeleton as unimportant according to a contour-based criterion, and trims these points from their skeletons. These trimmed skeletons are much more stable than the MAT, and are suitable for computing similarity, usually by means of graph matching algorithms. Graph matching can be performed by a number of methods, including transforming one graph into the other by a series of operations with individual costs, or finding the largest possible subgraphs of both the candidate and comparison shapes that are isomorphic to one another (Siddiqi, Shokoufandeh, Dickinson, & Zucker, 1998; Torsello & Hancock, 2004; Demirci, Shokoufandeh, Keselman, Bretzner, & Dickinson, 2006; Bai & Latecki, 2008; Rezanejad & Siddiqi, 2015).

The other main class of solutions is to take the MAT as inspiration, but to step away from its geometric rigidity. Proposed by Feldman and Singh (2006), this approach instead uses the shape skeleton as a probabilistic, belief-based generator for the shape contour. In this belief-based approach, there is a distribution over skeletons that can generate a given shape, based on how well each skeleton explains the shape contour. In previous work, our lab has used models based on this approach to predict human shape similarity ratings (Briscoe, 2008) and discrimination ability (Destler, Singh, & Feldman, 2019).

1.2.3 Image-based shape models

While most models of shape are based on a defined shape region, which is presumed to have been separated from the rest of the image by some prior segmentation process, some modern classification approaches are based on whole unsegmented images. These approaches are usually focused on recognizing known objects in scenes, and so are somewhat outside the scope of shape as discussed here. However, under some circumstances, these approaches can classify objects into categories with such accuracy that they cannot reasonably be ignored.

The most common image-based approach to object recognition comes from the computer vision literature. These approaches apply deep learning algorithms to large datasets, training neural networks to correctly categorize images containing various objects. The most effective of these algorithms are called Convolutional Neural Nets, or CNNs.

CNNs are not shape classifiers per se, since they are applied to entire unsegmented images without regard to internal structure, but CNNs have been used to identify object shapes within images with considerable success. Heavily trained, extremely complex CNNs, such as those developed by Krizhevsky, Sutskever, and Hinton (2012), Simonyan and Zisserman (2014), and Farabet et al. (2010) have placed highly in image classification competitions using tens or hundreds of thousands of images (see Deng et al., 2009 for information on a common image set used in these competitions). More practically, these CNNs have also proven successful in automatically detecting warning signs in cancer screening images (Cireşan, Giusti, Gambardella, & Schmidhuber, 2013). These CNNs often contain hundreds of thousands of nodes, and millions of learned weight parameters, which allows them to take advantage of highly complex patterns in the relationship between image properties and shape identity. However, this complexity also leads to two major problems, one practical, and one theoretical.

The practical problem is that large neural networks require large sets of training data, which means they take a very long time to train. Although a number of methods for accelerating the training process have been proposed (Farabet et al., 2010; Cireşan et al., 2012; Dean et al., 2012; Krizhevsky et al., 2012), these methods generally still result in days of training even for powerful computers. Once trained, these networks can classify new images rapidly and effectively, but this long initial training time, and the volume of training images required, is incongruous with the rapid learning times observed in humans. Image sets such that those detailed by Deng et al. (2009) exist precisely because compiling training sets is extremely cumbersome. This presents a problem if we wish to use CNNs as a model for human shape classification. Humans can make categorical judgments based on a single instance, such as after seeing only one image (Macario, Shipley, & Billman, 1990; Greene, 2018). If CNNs require hundreds

of thousands of images to make similar judgments, they are not accurately modeling human vision.

The theoretical problem is that a model with millions of free parameters is inherently less compelling as an explanation for an existing phenomenon (e.g. human shape perception). Due to the bias-variance problem, we can expect a model with a large number of parameters to fit the data it is trained on, but such a model may be heavily variant, and may generalize poorly to new data (Geman, Bienenstock, & Doursat, 1992). Indeed, Nguyen, Yosinski, and Clune (2015) found that DNNs trained to recognize images can be easily fooled in ways no human can be. For example, an image that appears to be simple white noise to a human eye can be classified by multiple DNNs as a peacock with extremely high (>99%) confidence. It appears that the complexity of the neural net leads to serious errors when the model attempts to process data which differ from the training set, producing wildly erroneous results. CNNs are exceptionally complex models, and as such, their strong performances on standard image classification tasks cannot be taken as similarly strong evidence of substantial similarity to human visual processing.

1.3 Similarity and categorization in vision

So far I have described a number of approaches, many of which provide a general metric of the similarity between two shapes. Shepard (1957) proposed that similarity metrics, whatever they may be, represent the distance in some perceptual space of possible entities (in our case, shapes). Shapes that are very similar will be nearby in this perceptual space, while shapes that are very dissimilar will be far apart. Ashby and Perrin (1988) argued instead that similarity depends on the relative probabilities of two entities arising from a noisy error model. The correct similarity metric, then, relates geometric shape properties to probability distributions in a representational space within the brain.

This representation space isn't just a "soup" in which shapes float. Different regions are designated to different categories, and shapes can be classified according to where

they fall in relation to those categories (Ashby & Perrin, 1988). It is this very categorical structure that provides the probability distributions that Ashby and Perrin argue for. In order to use these probabilities to generate a similarity metric, we must first understand the forms these categorical generators take (Huttenlocher, Hedges, & Duncan, 1991).

1.3.1 Prototypes and exemplars

Modern research on human categorization has largely been driven by the contrast between two broad frameworks. In both frameworks, categories are inferred based on known members, entities that are understood to fall within the category. In the first framework, put forth by Posner and Keele (1968), Rosch (1973), Homa, Cross, Cornell, Goldman, and Shwartz (1973), and Minda and Smith (2001), information about the known members of the category is combined to form a single prototype member. This prototype becomes the archetypal member of that category, and all new candidates for category membership are compared to the prototype. The category, then, can be viewed as a distribution centered on the prototype, with some error inferred from the non-prototype members.

In the second framework, put forward by Medin and Schaffer (1978), Nosofsky (1986), and Hintzman (1986), every known member of the category becomes an exemplar for the category. Instead of comparing new candidates to a single prototype, new candidates are compared to *all* current category members. In exemplar models, the category takes the form of a mixture model, composed of error distributions around each exemplar of that category. The probability mass of this distribution is highest, and thus candidates are considered more "categorical," in areas where many exemplars are clustered tightly enough that their error distributions overlap substantially.

Here, I will explore how human observers infer shape categories, and which, if either, of these two approaches they use to combine information about multiple shapes.

1.4 My contribution

In what follows, I propose a new model of shape similarity. I begin by assuming that human similarity judgments and related task performance is not arbitrary. Instead, the system that performs these judgments is adapted to perform certain kinds of tasks. Many of these tasks, including shape classification and discrimination, can be represented as asking whether two shape images were generated by one noisy generative process or by two different such processes, in the style of noisy distributions put forth by Ashby and Perrin (1988). In classification tasks, observers are straightforwardly asking whether shapes "belong" to particular distributions, each of which have some uncertainly (not all dogs are identical, for example, but they tend to share certain traits). In discrimination tasks, observers are instead asking whether the retinal shape image was generated by a particular object, through a noisy process that includes factors from neuronal noise to object orientation. In both cases, we can see that the task being performed is an assessment of how likely a shape is to be a member of some probabilistically-defined class. If shape similarity is adapted to these tasks, then two shapes will be similar to the extent that they are believed to belong to the same distribution vs. different distributions.

The model of shape similarity proposed here is based on this observation. Using the belief-based shape skeletons put forth by Feldman and Singh (2006), and building on the models employed successfully by Briscoe (2008), Destler et al. (2019), and Ayzenberg and Lourenco (2019), a shape similarity metric has been constructed based on the probability that two shapes are generated by a common process, vs. by two different processes. The higher the probability that the two shapes are generated by the same process, the more the evidence of the shape contours favors a single-model explanation, and the more similar two shapes are.

This model is then used to predict human performance on a shape categorization task, using a new framework. This new framework allows for the collection of large amounts of data on novel category inference and membership in a relatively short amount of time, giving the model plenty of data to predict.

Finally, this new model's predictions are compared to those of other shape similarity models, including a purely-contour-based model, a non-probabilistic skeletal model, and an image-based deep learning model, so as to determine which similarity model has the greatest ability to predict human classifications of novel shapes into novel categories.

Chapter 2

Skeletal Models of Shape

In what follows, I will propose a similarity metric grounded in skeletal models of shape. In this chapter, I will explain where shape skeleton models came from, what variations currently exist, and how each variation represents shapes. Special care will be taken to focus on the belief-based models put forth by Feldman and Singh (2006), as these specifically form the basis of the similarity metric described here.

2.1 The Birth of Skeletal Models

The first skeletal model of shape was a developed by Blum in the 60s and 70s (Blum, 1967, 1973; Blum & Nagel, 1978). This model, which Blum called the Symmetrical Axis Transform, was the first of what are now called "medial axis" models of shape. In Blum's model, a shape's skeleton is defined as the set of points which are the centers of circles that are wholly within the shape's outer boundary, but which cannot be fully contained within any other circle that also lies wholly within the shape's outer boundary (Blum & Nagel, 1978). In this way, every point along an axis in Blum's model is the midpoint between two boundary points. Note, however, that the midpoint between two boundary points is not necessarily an axis point.

Blum's model is simple, and fully deterministic. For any given shape, it is mathematically simple to compute the shape skeleton (see Fig. 1.2). However, this advantage is also the main disadvantage of Blum's model, and indeed of most medial axis models. Due to the deterministic nature of the model, small perturbations in the shape contour can lead to large variations in the shape skeleton. In particular, when Blum's model is applied to shapes with even moderate contour noise, it often produces numerous small axes that do not correspond to structures a human would label as independent parts.

In this way, Blum's model fails to accurately predict human part judgments, and thus is an unlikely model of human shape perception.

However, while Blum's model is unlikely to accurately model human shape perception, it laid the groundwork for a general class of models which show substantial progress. Burbeck and Pizer (1995) proposed a model, based on Blum's medial axes, by which neurons in the visual system might identify shapes by their axis-like "cores". These cores were characterized by being equidistant from all of their nearest contour boundaries, as Blum's medial axes are, although the Burbeck and Pizer cores are not quite so deterministic as Blum's axes. More recently, in a study of the macaque inferotemporal cortex, Hung, Carlson, and Connor (2012) found evidence for the primate brain's sensitivity to differences in medial axis-like skeletal structure of 3-dimensional shapes, independently of (and in addition to) its sensitivity to the shapes' surface structure. This suggests that skeletal models are a plausible, if ultimately incomplete, model of primate shape perception. Here, it is reasonably assumed that this category would include human primates as well as macaques, as evidenced by a number of psychophysical studies described below.

2.1.1 Later Medial Axis Models

Many skeletal models of shape follow closely in Blum's footsteps, using contour geometry to build a more sophisticated form of Blum's Symmetrical Axis Transform. These approaches take a number of forms.

Kovács et al. (1998) proposed a less binary form of medial axes. Instead of each point within the shape contour being either a part of an axis or not a part of an axis, an information value is computed for each point within the shape. This value represents the degree to which that point is at the center of a local boundary segment around it. In this way, Kovács et al. can assign a measure of "axiality" to each point within the shape. While this model tends to assign high information values to regions that traditionally correspond to medial axis points, these information values are not uniform along the axes, and were found to correspond with points of high human sensitivity to shape changes. In other words, the Kovács et al. model predicts not only that the

medial axes themselves are informative, but which parts of the medial axes are most informative, both in theory and in empirical practice.

This practice of evaluating which parts of the medial axis transform are most informative is not unique to Kovács. Siddiqi and collaborators have proposed a *shock graph* version of the medial axis, in which each point on the axis is evaluated in terms of its relationship to the axis contour (Siddiqi & Kimia, 1996; Siddiqi et al., 1998; Siddiqi, Kimia, Tannenbaum, & Zucker, 1999). These *shocks* are classified into several types, which are then used to draw categorical distinctions between different areas of the shape. In further research, they also identified special axes called *ligatures* (August et al., 1999). When the same contour points correspond to many different axis points (as near some kinds of subjective part boundaries), these axis points are classified as a ligature. These ligatures can be thought of as connectors, attaching the main axis of one shape part to the main axis of another. Most of the instabilities inherent in the medial axis transform are concentrated in these ligatures, so if they are identified and discarded, the remaining skeleton is much more stable. The resulting ligature-less skeleton is called a "bone graph", which Macrini, Dickinson, Fleet, and Siddiqi (2011) use as the input to a graph-matching algorithm for categorizing objects, achieving generally better classification accuracy than shock graphs that include the ligatures.

A conceptually-similar method was suggested by Torsello and Hancock (2004). This model analyzes how quickly the maximal circles' intersection points along the contour change as their corresponding axis points move along the skeleton. If the axis points belong to a ligature, the rate of boundary change is zero. By characterizing the shape according to its skeleton's rate of boundary length change, Torsello and Hancock were able to both minimize the effects of small contour variations, as above, and render the shape representation invariant to simple deformations ("bending"). Although their algorithm is not explicitly compared to human performance, it does generally well at classifying shapes in a labelled database.

Tam and Heidrich (2003) used a similar approach as they attempted to generalize the medial axis approach to 3-dimensional shapes. Although this can theoretically be accomplished by simply using the centers of spheres to compute axis points, rather than

circles, in practice this approach is problematic. The 3-dimensional object structure often means that the resulting structure of medial axis points is extremely complex, since even a slight contour convexity can create two surfaces to bound such a sphere (Leymarie & Kimia, 2001). This means that the medial axis decomposition of a 3d shape is extremely complex, hardly a decomposition at all. Tam and Heidrich attempt to solve this problem by pruning less-informative axes. Their algorithm estimates the volume of the 3D shape represented by each axis, and removes axes that represent very small volumes. Although this approach does not suggest a computationally efficient way to generate the pruned skeleton without first generating the full skeleton, the resulting pruned shock graph skeleton is a much more robust to small contour perturbations than a simple medial axis skeleton. This technique is not perfect, however. Although shape details are preserved with remarkable detail, small separate parts, such as ears on a human head, are often removed by the pruning process.

Firestone and Scholl (2014) took a different approach, exploring whether Blum's original medial axis model can predict human behaviors when interacting with shapes. Firestone and Scholl simply observed human behavior when asked to touch a spot of their choosing on a given shape. Surprisingly, with no additional instruction, human subjects tended to touch the shape along its medial axes. Most surprisingly, when small perturbations where made in the contours of the shapes, thereby changing the medial axes with minimal change to the shape's contour, different subjects still touched along the new medial axes. When subjects were asked to guess where other subjects would tend to touch, however, they very rarely guessed a medial-axis-like pattern, instead favoring either a uniform distribution across the shape or a cluster at the shape's approximate center. These results suggest that medial axis models are not consciously available, but that they nonetheless play a role in how humans interact with shapes.

2.1.2 Medial axis models for shape similarity

Perhaps unsurprisingly considering their use as shape models, medial axes have also been used to model shape similarity. Most of these models have taken advantage of the tree-like properties of medial axes, computing similarity in terms of the mathematical

operations that these structures entail.

Some models of similarity, such as the one employed by Liu and Geiger (1999), observe that medial axes have a tree-like structure, with axes branches off from other (usually larger) axes. This allows Liu and Geiger to use a standard medial axis model (Liu, Geiger, & Kohn, 1998) and build on it, computing an approximate best-match between two medial axis trees using an $A*$ algorithm, which gives them a correspondence between the two sets of axes. Then, for each set of matched axes, they compute a similarity cost. If two axes are matched, the cost is based on the difference in curvatures between the two axes. Liu and Geiger further define *cut* and *merge* operations, which add additional costs when an axis in one skeleton is matched to two or more axes in a different skeleton. By summing the cost of these three operations for a given pair of skeletons, Liu and Geiger estimated the dissimilarity between the two shapes, and built a robust matching algorithm that can identify part relationships in animal- and plant-like objects that appear to match human intuitive judgments.

Other models have used similar tree matching techniques. Often described as the *edit distance*, these multi-operation dissimilarity metrics have become quite common in shock graph approaches to shape (Siddiqi, Tresness, & Kimia, 1996; Siddiqi et al., 1998; Sebastian, Klein, & Kimia, 2001). In each of these models, some set of operations is defined which can be said to "transform" one shape skeleton into another. Each of these operations has a cost, usually assigned ad-hoc (although costs tend to increase with the degree of a particular change, e.g. a large change in one angle has a higher cost than a small change in the same angle), as in the Liu and Geiger (1999) model described above, and the total cost of transforming one skeleton into the other is used to estimate the dissimilarity between the two shapes. In edit distance models, skeletons that differ greatly from each other will require many transformations to become identical, so they will be evaluated as highly dissimilar.

2.1.3 Belief-Based and Hierarchical Skeletal Models

Blum's medial axis and its direct descendants are geometrically deterministic. That is to say, for any given shape, there is precisely one medial axis skeleton (under a given

model) that corresponds to that shape. However, not all skeletal models are based on simple geometric relationships.

Feldman and Singh (2006) proposed instead that skeletons are not a simple geometric property of the shapes they represent, but rather that skeletons are a model of the observer's beliefs about the object. The skeleton is conceptualized in these models as the schematic by which the brain reconstructs the shape. This reconstruction can be said to be probabilistic in the Bayesian sense, in that there are many different shapes which the observer believes might be reconstructed by a given skeleton, with various degrees of reasonableness. This means that a given skeleton can reconstruct many different shapes, with some probability distribution over potential shapes resulting from the reconstruction process. A given shape, by the same token, can be constructed by different skeletons, although any given shape only has a single most-likely skeleton. This framework adds complexity to the concept of shape skeletons, as a given shape now has a potentially-infinite number of generative skeletons rather than a single skeleton. However, because the generative process is belief-based, rather than purely geometric, small perturbations in shape contour do not generally produce large deviations in skeletal structure (Feldman et al., 2013).

Feldman and Singh (2006) proposed a model as described above, which used Bayesian inference to compute the highest probability (strongest belief) skeletons as an inverse probability problem. That is to say, the degree of belief of a shape having a particular skeleton is related to the probability of a skeleton producing that particular shape according to Bayes' Rule, in the form

$$p(\textsc{skeleton}|\textsc{shape}) \propto p(\textsc{skeleton})p(\textsc{shape}|\textsc{skeleton}).$$

This approach allows for a hierarchical ranking of the best skeletons for a given shape, including a simplicity prior ($p(\textsc{skeleton})$), which helps to finesse the bias-variance problem. After its introduction, this model saw further use in shape similarity metrics. Under the Feldman and Singh model, shape similarity can be measured as the probability of one shape under another shape's skeleton (Briscoe, 2008; Feldman et al., 2013).

$$\textsc{similarity}(\textsc{shape}_1, \textsc{shape}_2) = p(\textsc{shape}_1|\textsc{skel}_2)$$

Wilder, Feldman, and Singh (2011) showed that a Bayesian classifier trained on the Feldman and Singh (2006) can, for certain naturalistic categories, provide a good approximation of human classification judgments. However, this research was limited to familiar objects from highly-skeletal categories, such as animal and plant shapes. This research alone is not enough to establish shape skeletons as an appropriate model for shape perception more broadly.

In the Feldman and Singh model, the belief-based nature of the shape skeletons is at the forefront. In other skeletal shape models, such as the one proposed by Telea, Sminchisescu, and Dickinson (2004), the hierarchy of possible representations takes the front. In the Telea et al. model, this is accomplished by evaluating the shape contour under different levels of shape simplification and skeletal abstraction, defined as minimum description lengths for defining the shape contour given the skeleton. This creates a hierarchy of skeletons, where skeletons at one end of the hierarchy provide extremely simple-but-inaccurate explanations of the shape, while skeletons at the other end of the hierarchy provide extremely complex-but-accurate explanations of the shape. Taken together, these different shape skeletons can be used to represent a shape at whatever level is required for the task being performed by the observer, with a corresponding computational cost.

Chapter 3

A Novel Bayesian Similarity Model

In this chapter, I propose a new approach to shape similarity based on the Bayes factor evidence in favor of a common generator for a given shape pair. In an influential paper, Edelman (1998) argued that similarity is the fundamental construct from which perceptual decision and model comparisons derive. The framework proposed here flips this argument on its head, arguing that, consistent with the principles of Bayesian inference, *model comparison* is fundamental, and human judgments simply reflect degrees of relative model fit.

This new approach takes similarity as a Bayes factor comparison between two alternative generative models that make use of the belief-based skeleton framework described above. In one model, the two shapes arise from a common generative process, while in the other model, each shape arises from its own generative process. Similarity is then taken as the degree to which the observed shape information supports the common-generator model versus the separate-generator model. This Bayes factor similarity approach is highly similar to the likelihood ratio approach used by Erdogan and Jacobs (2017), which in turn is partly based on the early Feldman and Singh (2006) skeleton model, though the new model that will be presented here differs from theirs substantially in terms of how the component probabilities are calculated.

The model described here is based primarily, but not entirely, on the internal skeleton-like structure of the shape. Internal skeletons play an important role in shape processing (Burbeck & Pizer, 1995; Briscoe, 2008; Firestone & Scholl, 2014; Destler et al., 2019; Ayzenberg & Lourenco, 2019), but they are not the only models that exist. As described in Ch. 2, models based on the outer contour of the shape (Belongie et

al., 2002; Blake & Isard, 2012; Greene, 2018), or on the pixel properties of a shape image (Farabet et al., 2010; Krizhevsky et al., 2012; Simonyan & Zisserman, 2014), have also seen success in similar domains. In what follows, I introduce a similarity model that incorporates both skeleton and contour components, and in subsequent chapters compare it to models based on skeleton, contour, and image properties.

3.1 Bayes factor evaluation

In order to determine the degree to which the observed shapes support each model, we must first establish the principles by which we will evaluate the evidence in favor of each model. We will use uppercase letters to represent generative models, and lowercase levels to represent individual shapes. Therefore, under a belief-based generative framework such as the skeletal frame described in Ch. 2, there is a probability $p(s|M)$ for a given shape model M to generate a particular shape s. Using Bayes' rule, we can then evaluate probability $p(M|s)$ that an observed shape s was generated by a candidate model M.

$$p(M|s) = \frac{p(s|M)p(M)}{p(s)} \tag{3.1}$$

This equation allows us to evaluate the extent to which a particular shape model is supported by the observed shape. However, the shape similarity approach proposed here requires that we evaluate the extent to which a model is supported by evidence from two observed shapes, a and b.

$$p(M|a,b) = \frac{p(a,b|M)p(M)}{p(a,b)} \tag{3.2}$$

Assuming the shapes are generated independently conditioned on the model, we can rewrite this likelihood as the product of two separate probabilities, $p(a|M)$ and $p(b|M)$.

$$p(M|a,b) = \frac{p(a,b|M)p(M)}{p(a,b)} = \frac{p(a|M)p(b|M)p(M)}{p(a,b)} \tag{3.3}$$

Now that we can evaluate the extent to which the evidence from a pair of observed shapes supports a given shape model, we can compare different shape models to one

another. For the proposed similarity measure, we must evaluate the probability of a pair of shapes being generated by a common generative model $\hat{C}$, or by two separate generative models $\hat{A}$ and $\hat{B}$. Here, we use uppercase letters with hats to indicate models with parameters fitted to particular shapes, e.g. $\hat{A}$ is fitted to shape a, $\hat{B}$ is fitted to shape b, and so on. $\hat{C}$ is specifically used to indicate a common model whose parameters are fitted to both shapes a and b. Note also that $\hat{A}$ and $\hat{B}$ combine to form a single model $M_{\hat{A}\hat{B}}$, and that $p(a|M_{\hat{A}\hat{B}})p(b|M_{\hat{A}\hat{B}}) = p(a|\hat{A})p(b|\hat{B})$, since $M_{\hat{A}\hat{B}}$ proposes that a is generated only by $\hat{A}$ with no influence from $\hat{B}$, and vice versa. Therefore, the Bayes factor in favor of the common model takes the form

$$BF = \frac{p(a|\hat{C})p(b|\hat{C})}{p(a|\hat{A})p(b|\hat{B})} \tag{3.4}$$

This resulting BF value can be taken as the degree of similarity between the two shapes a and b, because it quantifies the degree to which the shape data support the common model relative to the separate models.

3.2 The shape lattice

In order the evaluate this Bayes factor, we must first define the generative models $\hat{C}$, $\hat{A}$, and $\hat{B}$. To do this, we introduce the concept of skeleton topology and the shape lattice. A given shape skeleton can be viewed as a tree, with a root axis that serves as a parent to one or more child axes, which may themselves be parents to additional child axes (see Fig. 3.1).

While there are an infinite number of skeletons with a given number of axes, there is a finite set of possible parent-child relationships between those axes (again, for a given number of axes). By defining these parent-child relationships, along with the total number of axes, we can create a representation of the skeleton's course topology. "Topology" is not used here to refer to all of a skeleton's physical features, but instead to refer to the set containing all of the parent-child axis relationships within the skeleton. For example, the skeleton in Fig. 3.1 would have the topology set $\{(\text{BLACK}, \text{BLUE}_1), (\text{BLACK}, \text{BLUE}_2), (\text{BLUE}_1, \text{RED})\}$, where e.g. $(\text{BLACK}, \text{BLUE}_1)$ means

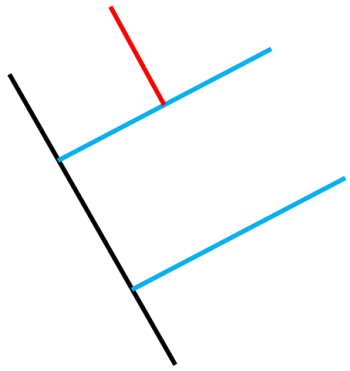

Figure 3.1: Parent-child relationships within a shape skeleton. The black axis is the root, and the blue axes are its children. One of the blue axes, in turn, is the parent to the red axis.

that the black axis is the parent of the first blue axis. As long as this set remains the same, two skeletons are considered topologically equivalent, even if the axes themselves are otherwise substantially different across the skeletons. From this definition, we can construct a theoretical lattice of skeleton topologies, with each node representing one possible topology. This shape skeleton lattice is itself defined as a set with a partial order on topology sets, such that $S \leq T$ if topology set S contains topology set T plus some additional elements. The lattice set is a lattice because, for any two topology sets S_1 and S_2, we can find both a meet (greatest common descendent) and a join (least common ancestor). The lattice meet between any two topology sets within the lattice is simply the union of the two topology sets (which is unique, and therefore the meet is unique), while the join is similarly the intersection of the two topology sets (which is also unique). See Fig. 3.2 for an illustration of the first four levels of the shape lattice.

But how does this shape lattice relate to the model probabilities for $\hat{C}$, $\hat{A}$, and $\hat{B}$? Each of these models represents a distribution over possible shape skeletons within a certain lattice node. If a shape's Maximum A Posteriori (MAP) skeleton (under the Feldman & Singh, 2006 model described in Ch. 2) falls within the same node as the model, determining the skeleton's probability under the model distribution is straightforward; it's simply defined by the distribution.

However, when a shape x's MAP skeleton S_x falls outside of the model's node, this simple definition is no longer sufficient. Each shape model $\hat{Y}$ is defined within a lattice

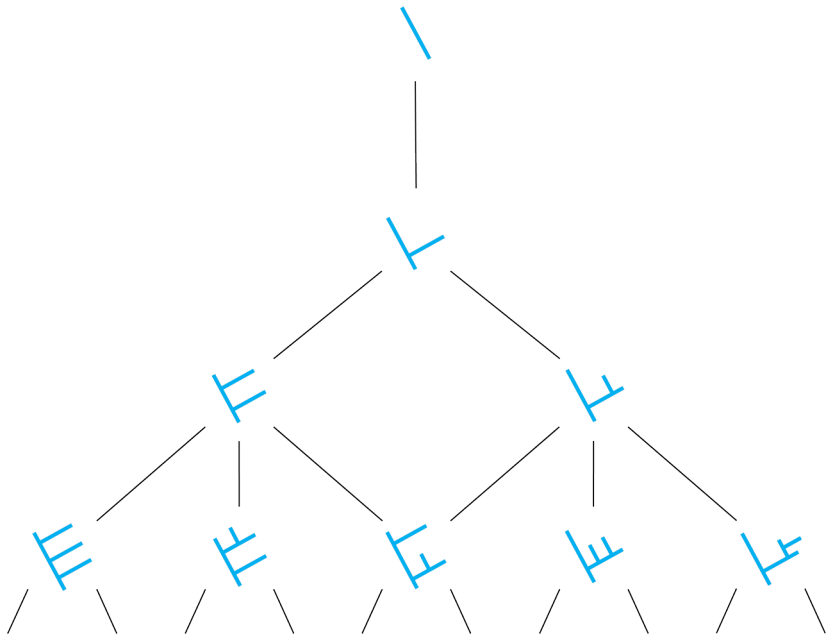

Figure 3.2: The first four levels of the shape lattice. Each node represents a class of parent-child skeleton topologies. At each level of the graph, every shape has the same number of axes. The shape lattice is not restricted to these four levels, and continues downward infinitely as more axes to added to the skeletons.

node Y, and describes a distribution over the parameters of each axis within that node. If there is a mismatch in the number or topological identity of the shape axes, the model distributions can no longer be straightforwardly applied. To solve this problem, we add to each model a distribution over the lattice position. The more topological differences (axis additions or removals) between the skeleton node X and the model node Y, the lower the probability of the skeleton's lattice position. Note that these topological difference probabilities are independent of the particular model parameterizations, so $p(\hat{X}|\hat{Y}) = p(X|Y)$. In other words, the probability of one topological structure given another remains the same, regardless of the parameter distributions used.

$$p(S_x|\hat{Y}) = p(S_x|\hat{X})p(\hat{X}|\hat{Y}) = p(S_x|\hat{X})p(X|Y) \tag{3.5}$$

From here, we can determine the final probability of any shape given any model within the lattice according to the following equation:

$$p(x|\hat{Y}) = p(x|S_x)p(S_x|\hat{X})p(X|Y) \tag{3.6}$$

Since the probability of a shape given a skeleton $p(x|S_x)$ is well-defined by the skeleton framework, and the other probability terms are well-defined as described above, we can now compute the full probability of any shape given any potential generative model within the lattice.

3.3 Model parameterization in practice

Now that we have described how to compute the probability of a shape under an arbitrary model within the lattice, our next task is to describe which non-arbitrary models to use for our evaluation. Recall that our goal is to solve for BF in the Bayes factor shown in Eq. 3.4. To accomplish this, we need to compute probabilities of shapes a and b under shape models $\hat{C}$, $\hat{A}$, and $\hat{B}$. However, while a and b are observed shapes, $\hat{C}$, $\hat{A}$, and $\hat{B}$ are latent models that must be estimated.

Our goal in defining $\hat{C}$, $\hat{A}$, and $\hat{B}$ is to represent the best possible one-generator and two-generator models for shapes a and b. In other words, we must create a set of models that maximize probabilities $p(a, b|\hat{C})$ and $p(a|\hat{A})p(b|\hat{B})$ (the numerator and denominator of the Bayes factor respectively). To do this, we will use the skeletons S_a and S_b to infer the axis parameters of $\hat{C}$, $\hat{A}$, and $\hat{B}$. For the common model $\hat{C}$, we will use both S_a and S_b to infer the parameters of the model distributions, such that $\hat{C}$ maximizes the total probability of both skeletons. For the purposes of generating $\hat{C}$ and calculating $p(a, b|\hat{C})$, S_a and S_b are normalized in terms of the scale and position of their root axes. In other words, their root axes are set to the same size and orientation, and any child axes undergo the same scale and orientation changes as the root axes. In this way, the model can be considered scale- and orientation-invariant to the base part of each shape, but not to the scale and orientation of additional parts. Some degree of scale and orientation invariance is commonly attributed to human shape perception both behaviorally (Loomis & Philbeck, 1999; Palmer, 1999) and neurologically (Riesenhuber & Poggio, 2002), though recent evidence has raised questions as to the degree of this

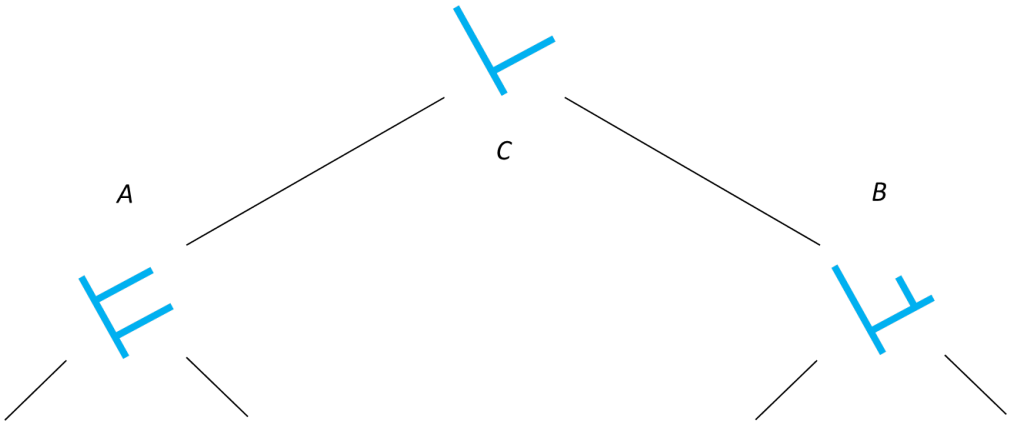

Figure 3.3: Three nodes on the lattice. Note that C is the join to both A and B.

invariance (Lehky & Tanaka, 2016). By enforcing these invariant properties for the base part, and only for the base part, we hope to reflect the totality of these findings while also adhering to the intuition if e.g. a leg is removed from a chair, the chair's remaining parts are not perceived as changing in size; an intuition supported by research on perceptual segmentation and the crucial role of the base part by Cohen, Singh, and Maloney (2008). The positions of the base parts are also normalized, to again reflect the intuition of the chair. If a leg is removed, the chair's position in space does not appear to shift, presumably because its base part remains stationary.

In addition, if S_a and S_b are located in nodes A and B on the shape lattice, respectively, $\hat{C}$ will exist within a node C which is the join of A and B. See Fig. 3.3 for a (simple) possible arrangement of these nodes.

For the separate models $\hat{A}$ and $\hat{B}$, we will use only S_a to parameterize $\hat{A}$ and only S_b to parameterize $\hat{B}$, because each generator only needs to maximize the probability of its respective skeleton. If we again take S_a and S_b to be located in nodes A and B respectively, then $\hat{A}$ and $\hat{B}$ will also be located in A and B respectively, because the "join" of A alone is simply A.

Because each model $\hat{C}$, $\hat{A}$, and $\hat{B}$ is parameterized on some set of a and b, we can more precisely notate them as $\hat{A}_a$, and $\hat{B}$ as $\hat{B}_b$, to illustrate where their parameters are inferred from. Carrying this further, we can notate new models that represent other combinations of parameterization and lattice node. Model $\hat{A}_{ab}$, for example, notates a

model within node A that is parameterized on both shapes a and b.

Now, at last, we can compute the Bayes factor BF for the skeletal model. The expansion of the skeletal Bayes factor as used here is

$$BF = \frac{p(a|\hat{C})p(b|\hat{C})}{p(a|\hat{A})p(b|\hat{B})} \tag{3.7}$$

$$BF(\text{SKEL}) = \frac{p(a|S_a)p(S_a|\hat{A}_{ab})p(A|C)p(b|S_b)p(S_b|\hat{B}_{ab})p(B|C)}{p(a|S_a)p(S_a|\hat{A}_a)p(b|S_b)p(S_b|\hat{B}_b)} \tag{3.8}$$

$$= \frac{p(S_a|\hat{A}_{ab})p(A|C)p(S_b|\hat{B}_{ab})p(B|C)}{p(S_a|\hat{A}_a)p(S_b|\hat{B}_b)} \tag{3.9}$$

3.4 Curvature component

While the main focus of the model presented here is the skeleton lattice described above, previous work suggests that both shape skeletons and shape contour properties play key roles in shape perception (Wilder, Feldman, & Singh, 2015a, 2015b). To that end, the probabilities used in the Bayes factor also include a component for the relative probabilities of different contours. Based on the measure of contour information proposed by (Feldman & Singh, 2005), we sample the contour curvature at a series of points along the contour of each shape, and use a distribution inferred from these curvatures to estimate a probability of one shape's contour curvature given another shape's contour curvature. For any shape a and b, we take K_a and K_b to be the sets of samples points along the contours of a and b respectively. The relative curvature of a to b is given by the equation

$$p(K_a|K_b) = \sum_i p(K_{ai}|K_{bi}) \tag{3.10}$$

where K_{ai} and K_{bi} represent matched points on the contours of a and b. For each pair of shapes, all possible order-preserving contour point correspondences are tested, and the best (maximum likelihood) is used in the similarity formulation.

As with the skeletal similarity components described above, we also infer a common distribution of contour curvatures from both shape contours, which we will refer to as

K_c. This allows us to compute $p(K_a|K_c)$ and $p(K_b|K_c)$, which we use to compute the curvature component of the Bayes factor.

$$BF(\text{CURVE}) = \frac{p(K_a|K_c)p(K_b|K_c)}{p(K_a|K_a)p(K_b|K_b)} \tag{3.11}$$

This curvature component is then combined with the skeletal component described above to form the full Bayes factor used by our lattice similarity model.

$$BF = \frac{p(a|\hat{C})p(b|\hat{C})}{p(a|\hat{A})p(b|\hat{B})} \tag{3.12}$$

$$= BF(\text{SKEL})BF(\text{CURVE}) \tag{3.13}$$

$$= \frac{p(S_a|\hat{A}_{ab})p(A|C)p(S_b|\hat{B}_{ab})p(B|C)p(K_a|K_c)p(K_b|K_c)}{p(S_a|\hat{A}_a)p(S_b|\hat{B}_b)p(K_a|K_a)p(K_b|K_b)} \tag{3.14}$$

This combined similarity measure expresses how similar the contours are as contours with how similar the enclosed regions (represented skeletally) are as regions.

3.5 Edit distance approximation

Thus far, the lattice similarity metric has been presented in terms of the Bayes factor between a common-generator model and a separate-generator model. However, the skeleton component of this metric can also be described as an approximation of the edit distance between two shapes. Recall that the edit distance is a measure of the number and complexity of transformation operations that must be performed to make two skeleton identical, and that edit distance has been used as a model of shape dissimilarity in the past (Siddiqi et al., 1996, 1998; Sebastian et al., 2001).

Here, we observe that an edit distance expressed in terms of generative probabilities is approximately equivalent to the numerator of the skeletal component of the Bayes factor, while the denominator serves as a normalization.

There are two components of the edit distance between two shapes a and b in the lattice similarity framework. The first component is the minimal path through the lattice from node A to node B (see Fig. 3.4). This path has a probability dependent

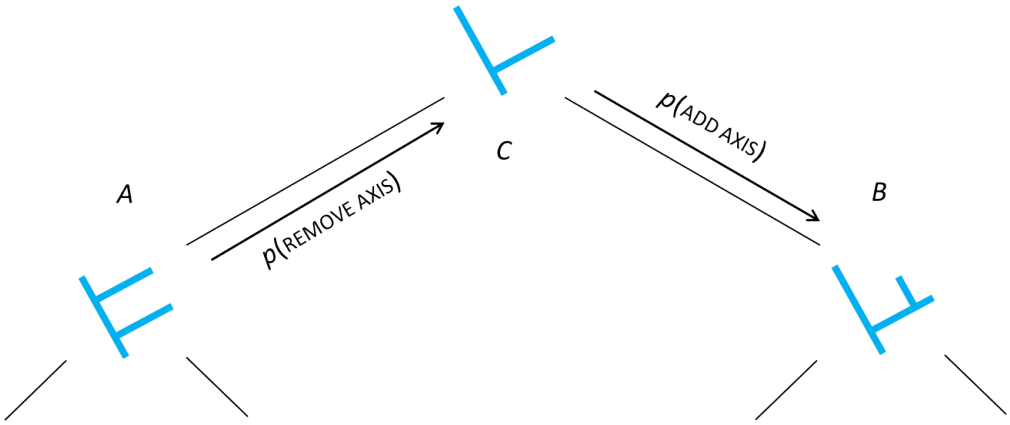

Figure 3.4: The edit path between two nodes A and B on the shape lattice. Each time an axis is added to or removed from a skeleton S_a in A, the skeleton "moves" to a new node, and the probability of the resulting transformed skeleton decreases. In this way, longer edit distances have lower probabilities associated with them.

on the number of skeleton transformations required, or in other words, the number of nodes traversed, with longer paths having lower probabilities. This is because each topological transformation has a certain probability, described in the form $p(X|Y)$. The probability of two such transformations is the joint probability of both transformations. For example, to traverse the lattice from X to Y and then from Y to Z, we compute probability $p(Y|X)p(Z|Y)$. Since both component probabilities are less than 1, the resulting joint probability is lower than either component probability. As the length of the path increases, the number of these component probabilities increases, and the joint probability of all these transformations decreases further as a result.

The shortest possible path between two shapes necessarily passes through their meet (greatest common descendent) or, more saliently for our purposes, their join (least common ancestor). In the case of A and B, their join is C, which means we can split $p(B|A)$ into the joint probability $p(C|A)p(B|C)$. If we assume axis additions and removals are equally probable, any path probability $p(X|Y)$ is equal to $p(Y|X)$. Therefore, $p(B|A) = p(A|C)p(B|C)$. This joint probability is a key component of the numerator of the Bayes factor described in Eq. 3.9.

Of course, these probabilities only describe the cost of moving from one node to another. However, shape skeletons within a single node can still be highly dissimilar, as illustrated in Fig. 3.5. Here, the edit distance diverges somewhat from the Bayes

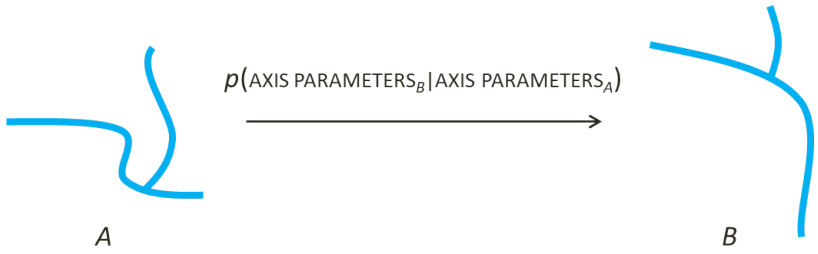

Figure 3.5: Even when two skeletons have the same topological structure, they can still differ significantly. To account for these, the lattice similarity model also incorporates additional dissimilarity costs for differences in axis parameters, such as curvature, length, and origin on the parent axis.

factor model. In an edit distance framework, we transform skeleton S_a into S_b directly after topological additions and subtractions. This means that to compute this final stage of the transformation, we have to compute $p(S_b|\hat{B}_a)$, the probability of S_b given a parameter model $\hat{B}_a$ that has S_b's topological structure, but S_a's axis parameter values. By contrast, the Bayes factor model asks us to compute the probability of both S_a and S_b under parameter models based on both of their parameter values, resulting in a probability $p(S_a|\hat{A}_{ab})p(S_b|\hat{B}_{ab})$.

While both these probability equations represent related transformations based on skeletons S_a and S_b, they are fundamentally different equations, and should not be expected to resolve to the same values. For this reason, the edit distance can be described as an approximation to the numerator of the partial Bayes factor over shape skeletons, but the two approaches are not equivalent in a strong sense.

Chapter 4

Experimental Framework

In order to test this lattice model of skeletal shape similarity, a shape classification task was used. As described above, similarity is conceived of here as a psychological system adapted to classifying mental objects. As such, a metric that models shape similarity ought to be able to predict human shape classification judgments with good accuracy.

In order to evaluate our model's predictions, a novel framework was developed that allows researchers to obtain a large number of human shape classification judgments in a relatively short amount of time, and to obtain these judgments in the context of different imposed structures on the shape categories.

4.1 General Method

Under this experimental framework, each subject was asked to classify shapes as belonging to, or not belonging to, a particular unfamiliar category based on a set of category example shapes. Subjects were shown one or more example shapes that were labelled as belonging or not belonging to a novel named category (e.g. "this is a blicket," or "this is not a blicket"), and were asked to choose additional shapes from a list of candidates that they judged as belonging to that same category (e.g. "which of these are blickets?"). The candidate shapes were displayed in white in a 6x6 grid at the center of a computer screen, while the positive examples were shown on the left in blue, and the negative examples were shown on the right in white, as seen in Fig. 4.1. Text instructions were displayed at the top and bottom of the screen.

Subjects selected shapes that they judged to be members of the novel category by clicking on them. When they clicked on a shape, its color changed from white to blue,

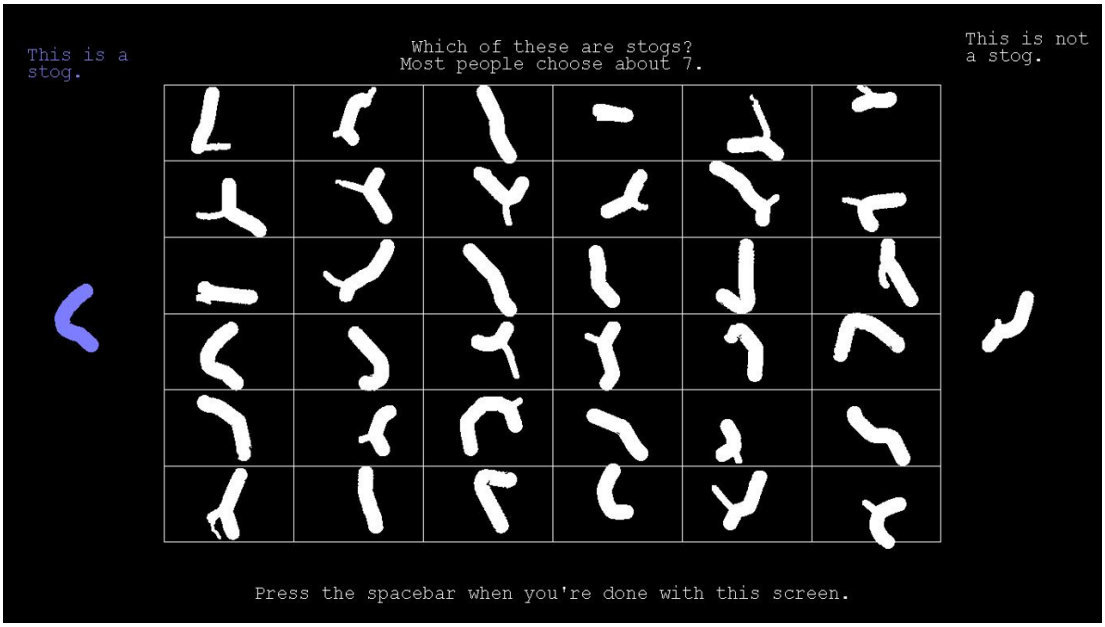

Figure 4.1: An example screen from one trial of our experiments. On this trial, one shape is shown from a novel category (left, blue), and that is not from a novel category (right, white). The 36 shapes in the center grid are the candidate shapes, which the subject clicks on to indicate each shape's judged membership or non-membership in the blue novel category.

the same color as the positive example, indicating its judged membership in the novel category. Additional clicks on the same shape would toggle the shape's membership status and color.

Subjects were instructed that most people chose approximately seven shapes on each trial to be members of the known category. This instruction was given in order to create a norm that would have the subjects pick an informative number of shapes, without strictly requiring them to pick a certain number, which might cause them to select shapes that they did not judge as very similar.

Once the subject had picked all the shapes they judged as members of the category, they would press the spacebar to lock-in their selections and move to the next trial. Each subsequent trial used a different set of randomized shapes, and a different novel category name. In order to minimize the chance that the category names would be confused, each category name began with a different letter of the alphabet. Each subject saw 26 trials, each with its own novel category, so no starting letters were repeated for any given subject.

Note that this experiment does not have ground truth correct answers. The shapes are not a-priori grouped into categories; only the subjects' judgements divide them into different classes. While there are structural differences between the shapes, these differences only affect category membership insofar as they influence the subjects to classify shapes differently. Because there is no ground truth, it is impossible to objectively evaluate a subject's performance. Instead, the models were evaluated in terms of how well they predicted the subjects' choices.

4.2 Shape generation process

4.2.1 2D shape generation

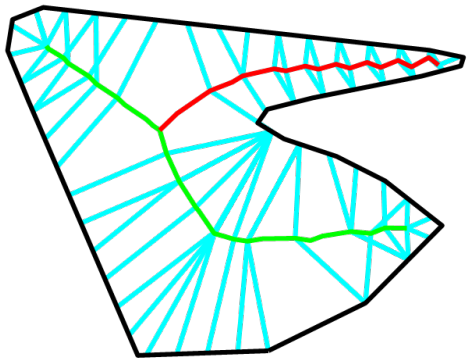

Figure 4.2: A sample shape and its MAP (highest posterior probability) skeleton. The shape contour is shown in black, the two axes of the MAP skeleton are shown in green and red respectively, and the ribs are shown in blue. Note how the ribs project outward from the axes to generate the shape outline.

Shapes for the 2D classification experiments were randomly generated using a three-step process. During the first step, for each shape, the generator randomly selected a series of coordinate points from a uniform X-axis distribution and a uniform Y-axis distribution. These randomly-selected points formed the first rough contour outline of the shape. This outline was defined further by computing the concave hull over these contour points.

Once the concave hull was computed, a shape skeleton was extracted from the

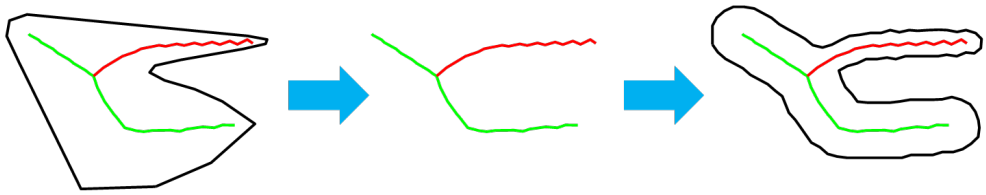

Figure 4.3: An illustration of the shape generation process. First, a concave hull (the shape contour on the left, in black) is generated from a series of random points, and a shape skeleton is computed based on that (the green and red axes on the left). That skeleton is then extracted from the concave hull (middle), and used to generate a new axial shape (the shape contour on the right, in black) using the methods described in Feldman and Singh (2006).

contour using the method described in Feldman and Singh (2006). As described above, this skeleton forms a belief-based explanation of the shape contour. Under the Feldman and Singh skeleton model, any shape can have a number of skeletons, each of which is believed to a different degree $p(\text{SKELETON}|\text{SHAPE})$. For the purposes of this experiment, only the MAP skeleton was used for each shape, which is the single most plausible (highest posterior probability) skeleton given the shape contour.

Finally, rib projection probabilities stored in the MAP skeleton were used to generate a final shape contour by projecting ribs out from the skeleton axes, effectively reversing the skeleton inference process to generate a new shape from the skeleton by sampling from the generative model. This is possible due to the fundamentally Bayesian nature of the Feldman and Singh skeletons. Just as there is a probability of a given shape being generated by a particular skeleton, so too is there a probability of a given skeleton generating a particular shape (see Fig. 4.2). Here again, only the most likely shape was computed for the purposes of these experiments, although a theoretically-infinite number of shapes can be generated by any skeleton. Once generated by this reverse skeletal process, the new shape contour was used directly for my 2D shape experiments. See Fig. 4.3 for a visual summary of the full shape generation process.

4.2.2 3D shape Generation

The shapes used in the 3D experiments were generated similarly, but with a twist. The skeleton-estimation process does not have a counterpart in 3D as yet, but the likelihood model does have a 3D version, the inflation model described in Feldman et al. (2013, and see also Twarog, Tappen, & Adelson, 2012). The inflation model is a 3D generalization of the likelihood model developed by Feldman and Singh (2006), which constructs the 2D shape contour by projecting ribs outward from the skeleton axes. In the 3D inflation model, the shape's surface is constructed in a similar fashion, by projecting ribs outward from the axes in three dimensions, in a manner reminiscent of the spokes of a wheel (see Fig. 4.4). The endpoints of these axes form the vertices of the 3D shape surface.

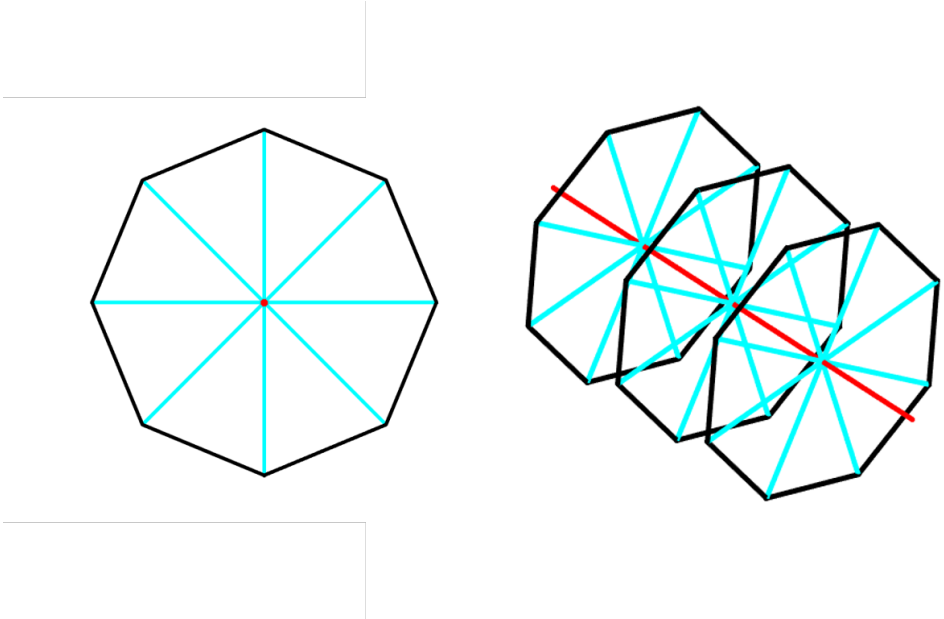

Figure 4.4: An example of the 3D inflation process. Left: a lengthwise view of the axis (red), which projects ribs outwards like spokes of a wheel (blue). The endpoints of these ribs form the vertices of what will become the 3D surface (black). Right: an axis (red) projecting ribs (blue) from three different points, creating three sets of vertices (black) which will define the 3D surface.

In the 3D experiments, the shape surfaces generated by this inflation process from known skeletons (inferred from the 2D concave hulls described above, so the skeletons are always planar) were used in place of the estimated skeletons.

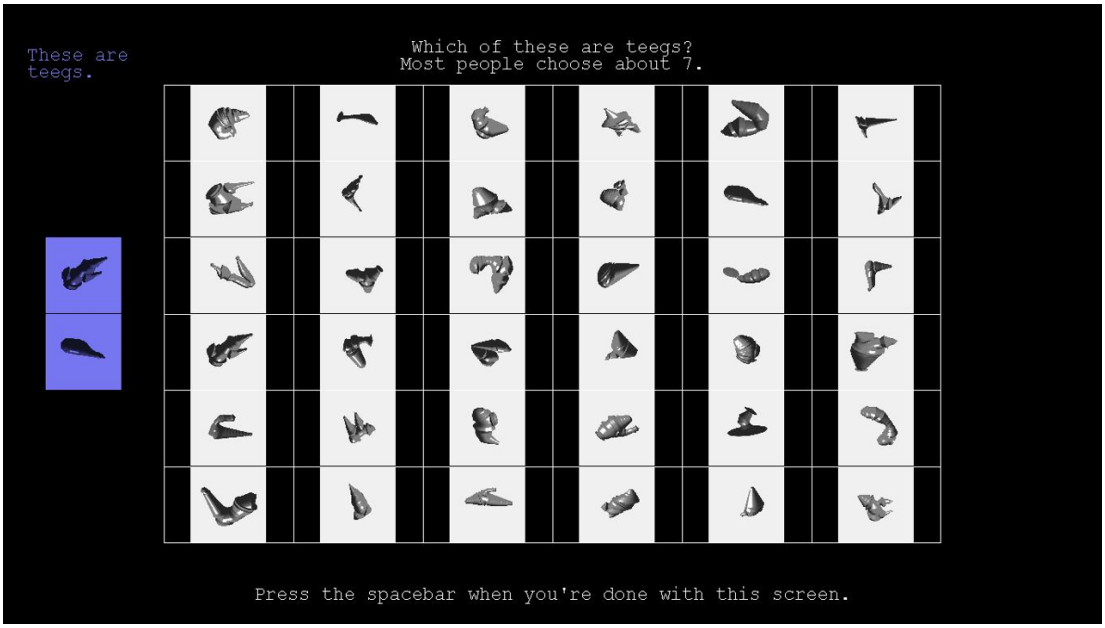

Figure 4.5: An example screen from one trial of the 3D shape experiments. On this trial, two shapes from a novel category were shown (left, blue). The 36 shapes in the center grid are the candidate shapes, which the subject clicks on to indicate each shape's membership or non-membership in the blue novel category.

In this way, new 3D shapes were created with corresponding skeletons that could be used in the lattice similarity model. The shapes were always rendered from the same fixed viewpoint, 15° off from the planar view, so that the overall 3D structure would be visible to each subject. See Fig. 4.5 for an example experiment screen showing these rendered 3D shapes.

When computing similarities of the 3D shapes, the planar skeleton that was used to generate a shape was substituted for that shape's MAP skeleton under the Feldman and Singh skeletal model. In this way, the generative skeleton serves as a reasonable approximation of the MAP skeleton for the 3D shape, which allows us to sidestep the practical problem of computing the 3D skeleton from a 3D shape, which is outside the scope of the current project.

When computing similarities of the 2D shapes, the MAP skeleton was computed using the ordinary Feldman and Singh method described for 2D shapes, above.

4.3 Data collection process

Amazon Mechanical Turk was used to collect subject responses. All participants had at least 95% positive ratings on the Mechanical Turk system, and were recruited from within the United States.

After seeing an instruction screen describing the experimental task, each subject saw one full round of 26 novel shape categories across 26 screens, each with its own example shape(s) and candidate shape grid. Subjects selected shapes as members of the category by clicking on them, as described above. Subjects usually responded to the full set of shapes in approximately 20-30 minutes, and were compensated $4 each for their participation.

4.4 Advantages of this experimental framework

The primary advantage of this experimental method is the speed with which large amounts of data can be collected. Subjects are able to judge shapes as belonging to the category or not belonging to the category very quickly when shapes are presented simultaneously. Even when presented with 36 shapes at once, subjects respond in systematic patterns (as will be shown in Ch. 6) in times averaging less than a minute per set of 36. This new framework takes advantage of this rapid processing and response time to collect data on many shapes at once, cutting down on the total time needed to run classification experiments.

Furthering the speed advantage offered by this framework, the instructions are simple, and the task can easily be performed by non-expert subjects on a computer. This allowed the data collection process to be automated using Amazon Mechanical Turk, making it even faster.

This high collection speed allowed the testing of a large number of shape categories in each experiment. Even more importantly, it enabled the running of ten different experiments (described in Ch. 5) using different patterns of category example shapes in both the 2D and 3D domains.

Chapter 5

Experiments

Using the experimental method described above, ten experiments were performed testing subjects' shape classification judgments based on a small number of positive and negative examples. Exps. 1-5 were performed with 2D shapes such as those seen in Fig. 4.1. Each of these five experiments used a different pattern of positive and negative example shapes, e.g. Exp. 1 had only a single positive example and no negative examples, while Exp. 5 had two positive examples and one negative example. These experiments are denoted x:y, where x is the number of positive examples and y is the number of negative examples. For example, the 2:0 experiment had two positive examples and no negative examples.

Different numbers of positive and negative examples were used for several reasons. First, to investigate the question of how we combine information from multiple positive examples to infer a category. What is the essential information that e.g. all chairs have in common? Humans can easily identify very different styles of chairs from very different viewpoints, and classify them as chairs, in spite of large differences in the physical structure and retinal projections of different chairs. What information do we use to accomplish this? The different numbers of positive examples were used in order to both explore that question, and to examine categorization in the absence of that additional level of complexity. Second, to explore the effects of a negative example on the inference of the positive category. It was expected that the negative example would be treated as an example from a second category, as in common two-alternative-category experiments (see e.g. Ashby & Gott, 1988, Ashby & Maddox, 1992, and Waldron & Ashby, 2001), and that shapes would be classified as belonging to the named category to the extent that they were more likely to be generated by the named category than

by this negative category. However, the negative example is not explicitly a member of a second category in these experiments, and instead simply represents a set of shapes disjoint from the named category. These negative examples were tested in order to explore how members of such a disjoint set would affect the inference of the named category.

In addition to these five experiments with 2D shapes, another five experiments, Exps. 6-10, were performed using 3D shapes. Each of the 3D shapes was rendered and shown from a single canonical view as described in Ch. 4 (see Fig. 4.5).

10 subjects participated in each experiment ($N = 10$), based on pilot studies suggesting this would be sufficient to reveal consistent trends in subjects' shape classification judgments. Across all ten experiments, 100 subjects participated in total. The subjects were recruited using Amazon Mechanical Turk, and each subject was compensated $4 for their time. All subjects were recruited within the United States. The experiment took 20-30 minutes for each subject.

Informed consent was obtained from all participants, and all research described here was conducted under the supervision of the Rutgers Institutional Review Board and in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

5.1 Exps. 1 and 6 (1:0 Case)

The simplest experiments were the 1:0 experiments, with a single positive example shape and no negative example shapes. These experiments are the purest measure taken here of how shape similarity relates to shape classification. Since the novel category is only informed by a single example shape, classifications should be based primarily on the similarity between the example shape and each candidate shape, with a minimum of influence from other category inference processes.

5.2 Exps. 2 and 7 (2:0 Case)

The 2:0 experiments had two positive examples and no negative examples. These are the tasks that most cleanly test how a category is inferred from multiple category examples. Under the Ashby and Perrin (1988) framework of similarity as a function of relative probabilities under some noisy model, similarity to the category should be based on the category distribution. Therefore, we can evaluate the shape of this distribution by testing the classification patterns under different positive example pairs, and investigating how the distance between the positive example shapes informs the distribution that subjects infer.

5.3 Exps. 3 and 8 (3:0 Case)

The 3:0 experiments had three positive examples and no negative examples. These experiments were largely designed to serve as a check for the 2:0 experiments, to test whether the same models that predicted two-example category inference also predicted category inference for higher numbers of examples, and to further challenge the combination rule.

5.4 Exps. 4 and 9 (1:1 Case)

The 1:1 experiments had a single positive example and a single negative example. These experiments were a partial return to the more "pure" similarity measured in the 1:0 experiment. This time, however, a single negative example was also included, in order to test the effects of the negative example on the similarity distribution around the positive example.

5.5 Exp. 5 and 10 (2:1 Case)

Lastly, the 2:1 experiments had two positive examples and one negative example. These experiments served as a combination of the 2:0 and the 1:1, testing how category inference from multiple positive examples was influenced by the presence of a negative

example.

5.6 2D and 3D

As described above, each of these five experiment types was performed twice. Once using only 2D shapes (Exps. 1-5), and once using only 3D shapes (Exps. 6-10) viewed from a fixed perspective, for ten experiments in total.

5.7 Data Collection

In each of these ten experiments, the example shapes shown to the subjects were varied on different trials, and which shapes each subject judged as a member of the positive category for that trial were recorded. These classification judgments were then used to test whether each of several different similarity models could predict the human classifications, as I will describe below.

Chapter 6

Results and Model Comparison

Once the shape classifications were collected, the extent to which each of several models predicted the human classifications was analyzed. In order to do this, each model was used to predict which choice(s) a subject would make on a given trial, modelling each choice made by each subject on each trial. Recall that there is no "right answer" to how similar one shape is to another or which shape belongs in which named category; the models can only be considered good or bad relative to human behavior, and to the predictions made by other models.

For each model, the prediction was computed as follows: For each choice the subject made on a given trial, a probability was assigned to each (not already chosen) shape in the grid. The probabilities assigned to each shape were proportional to the negative exponential of the similarity to the positive example(s) times a sensitivity factor ($e^{-c*\text{SIMILARITY}}$), in accordance with the Luce choice rule (Luce, 1959; Shepard, 1987). The probabilities were then normalized such that the probabilities summed to 1, since for each choice, the subject picked exactly one shape, by definition. The probability of the choice the subject actually made was recorded, and the analysis moved on to the next choice. Each model was penalized according to the log probability of these choices under that model, summed over trials under the assumption that the trials were conditionally-independent given the model and the previous choices of the individual subject. Data from all 10 subjects in each experiment were combined to compute an AIC value for that model, in that experiment.

The AIC values for each model are shown in Figs. 6.1 through 6.6, sorted according to experiment. Lower AIC values indicate higher model probabilities (Akaike, 1973, 1974, 1979), and all plots show AIC values relative to the lowest AIC value of any

model analyzed, as suggested by Wagenmakers and Farrell (2004). As such, the model that best predicts the data in a given experiment has a relative AIC of 0, and cannot be seen in the graphs. All visible models are worse predictors, and the higher their relative AIC values, the worse they are at predicting human classifications in this experiment.

In the 2:0, 2:1 and 3:0 experiment graphs, similarity was usually computed for each candidate shape based on its similarity to the most similar positive example shape. This approach is comparable to an exemplar model of similarity (Medin & Schaffer, 1978; Nosofsky, 1986; Hintzman, 1986), which does not involve the explicit construction of a category central tendency. The lattice similarity model, however, includes a built-in model of category inference, as described in Ch. 3. As such, for the lattice similarity model, a prototype-like similarity measure was tested and included which infers a common generative distribution for all positive examples and then computes the similarity of candidate shapes based on that distribution. This prototype model was used to produce the lattice similarity predictions used here.

In the 1:1 and 2:1 cases, the similarity to the negative example was also incorporated into the final classification probability for each model, though weighted by a fitted parameter. In these cases, a shape was assigned a predicted probability of being selected proportional to the log likelihood ratio between its similarity to the positive example (see above), and its similarity to the negative example times the weight parameter w. This ratio is expressed as subtraction because of the log scale used for these probabilities. For a given shape a, model similarity between two shapes sim, positive example e_p, and negative example e_n, the final classification probability is given by

$$p(\text{SELECTED}) \propto log(sim(a, e_p)) - w * log(sim(a, e_n)) \tag{6.1}$$

6.1 The models

6.1.1 Lattice similarity

The lattice similarity model (dark blue on the graphs) is the primary model proposed herein, and has been described in detail in Ch. 3. It is based on the contour and

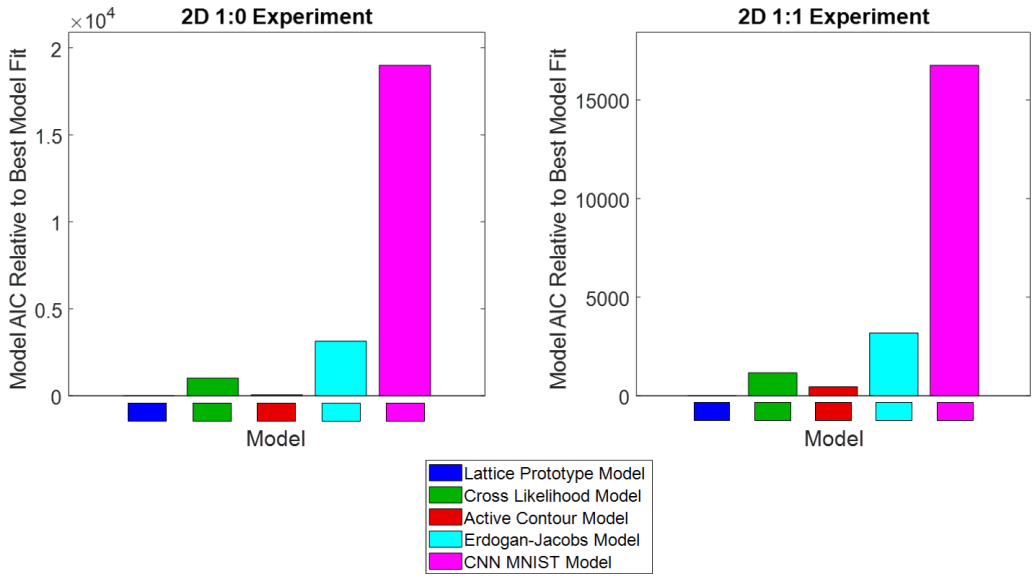

Figure 6.1: Model AIC values for the 2D 1:0 and 1:1 experiments. Each AIC is shown relative to the best predictor, which in turn is 0, and therefore not visible on the bar graph. Each visible bar represents a model which performs worse than the best predictor for that experiment, with higher bars representing worse performance.

skeleton models proposed by Feldman (1997) and Feldman and Singh (2006), and treats similarity as a Bayes factor comparison between a model of two shapes being generated by a common process, and two shapes being generated by different processes.

As presented here, this model includes 8-9 fitted parameters (depending on the experiment, see below) that are used primarily to weight the component probabilities used to evaluate the Bayes factor. Four of these parameters are scalar weight factors on the components of the skeletal similarity probability. These weight factors control the extent to which the model values differences in axis length, turning angle, initial branch angle from the parent axis, and position of the branch on the parent axis, when comparing a skeleton to a skeletal distribution (e.g. $p(S_a|\hat{A}_{ab})$). Another parameter weights differences in the contour curvature pattern between shape contours (e.g. $p(K_a|K_c)$), while another weights the probability of each shape under its own skeleton $p(a|S_a)$. Another fitted parameter serves as the probability cost of lattice additions and removals, which are assumed here to be equally costly. The next, and usually final, fitted parameter is the sensitivity c used to turn model similarity ratings into predictions of human performance ($p(\text{SELECTED}) \propto e^{-c*\text{SIMILARITY}}$, as described above). In the 1:1 and 2:1

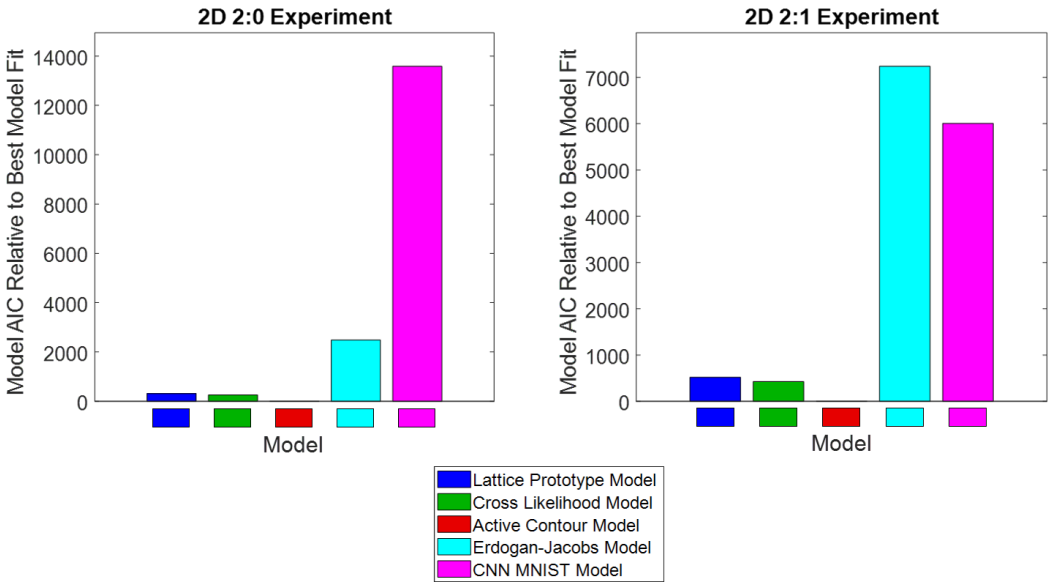

Figure 6.2: Model AIC values for the 2D 2:0 and 2:1 experiments. Each AIC is shown relative to the best predictor, which in turn is 0, and therefore not visible on the bar graph. Each visible bar represents a model which performs worse than the best predictor for that experiment, with higher bars representing worse performance.

experiments, a ninth and final parameter was added, which governed the weight the model put on the similarity to the negative example. Negative example similarity was otherwise computed using the same weights as positive example similarity.

These parameters were fitted based on data from the first half of the subjects in each condition, and tested on the full dataset for that condition.

6.1.2 Cross likelihood model

The cross likelihood model (green on the graphs) is a previous skeletal model of shape similarity introduced by Briscoe (2008). Although it is still based on the skeleton work by Feldman and Singh (2006), the cross likelihood model does not incorporate the shape lattice, the Bayes factor, or any contour component. Instead, it simply asks, for two shapes a and b, what is the probability of shape a being generated by shape b's skeleton, and vice versa.

$$CL = \frac{p(a|S_b) + p(b|S_a)}{2} \tag{6.2}$$

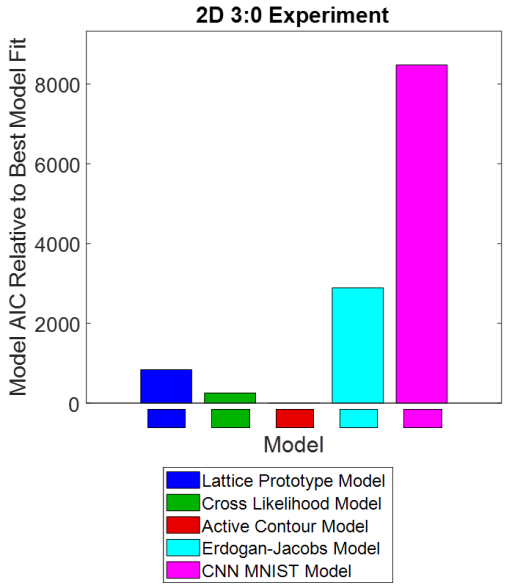

Figure 6.3: Model AIC values for the 2D 3:0 experiments. Each AIC is shown relative to the best predictor, which in turn is 0, and therefore not visible on the bar graph. Each visible bar represents a model which performs worse than the best predictor for that experiment, with higher bars representing worse performance.

This model has proven effective at predicting human ability to discriminate highly similar shapes (Destler et al., 2019), and has shown some promise for predicting similarity ratings (Briscoe, 2008), but its predictions began to deteriorate as the two shapes become less similar (as happens when two shapes have qualitatively different part structures) and the probability of one shape's skeleton producing the other shape approaches 0.

The cross likelihood model is not defined for 3D shapes, due to the limitations of the skeleton model described in Ch. 4. As such, the cross likelihood model is excluded from the analyses of the 3D experiments.

6.1.3 Erdogan and Jacobs model

The model proposed by Erdogan and Jacobs (2017, light blue on the graphs) is similar in some respects to the lattice similarity model and shares its generative formulation. As in the lattice similarity model, shapes are compared as a likelihood ratio between their probability of being generated by a single generator, and their probability of being generated by two separate generators. The Erdogan-Jacobs (EJ) model

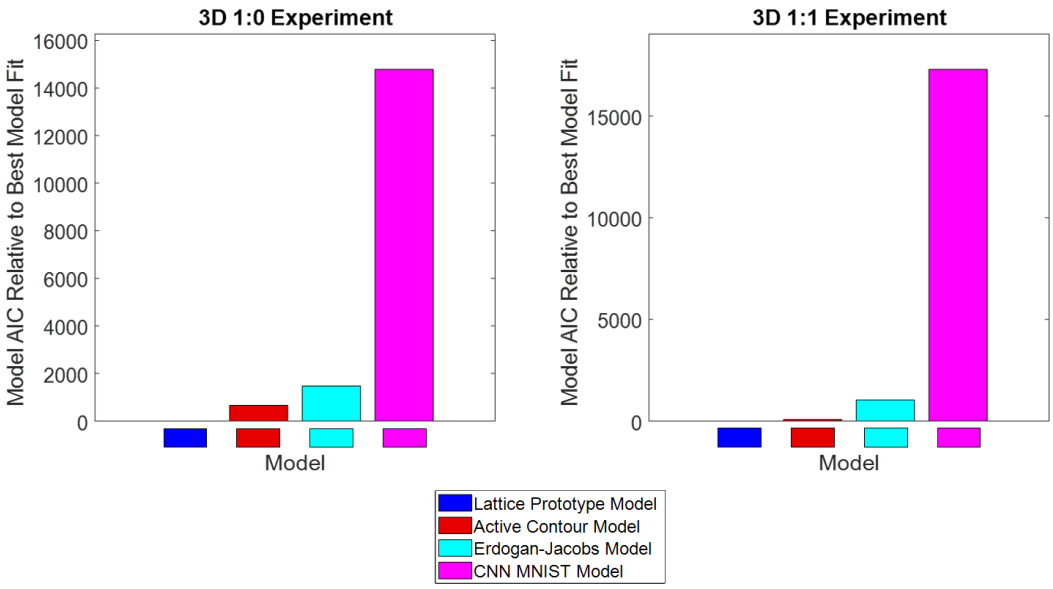

Figure 6.4: Model AIC values for the 3D 1:0 and 1:1 experiments. Each AIC is shown relative to the best predictor, which in turn is 0, and therefore not visible on the bar graph. Each visible bar represents a model which performs worse than the best predictor for that experiment, with higher bars representing worse performance.

begins by inferring the 3D structure of a shape from a single rendered image. Shapes are represented as sets of segments, somewhat analogous to shape skeleton axes, and segment endpoints. A bounded-uniform prior probability over number of segments and endpoint positions is combined with a likelihood for 3D shape inference that assumes the image reflects the 3D stimulus plus some Gaussian error and a given viewpoint of 2D projection.

Once this 3D shape has been inferred, the EJ model computes predicts shape classifications using the likelihood described above. The model compares shapes in terms of their rendered images, and computes a posterior probability that the second images was generated by the same 3D shape as the first, under the error and viewpoint distributions described in the likelihood function. The EJ model then computes the posterior probability that the second image was generated by any other 3D shape. These two posterior probabilities are then compared in a posterior ratio, which serves as the EJ measure of similarity.

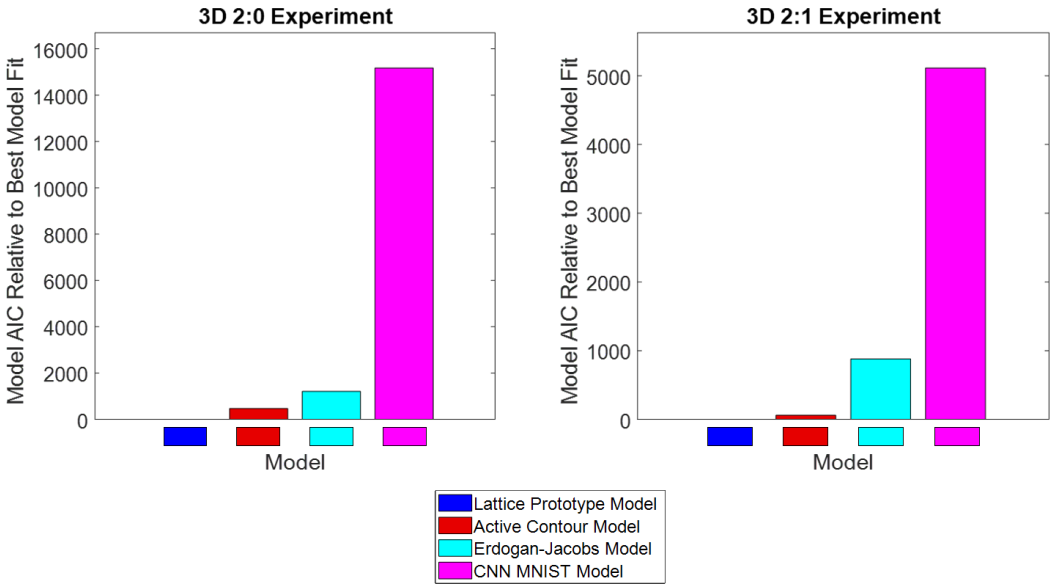

Figure 6.5: Model AIC values for the 3D 2:0 and 2:1 experiments. Each AIC is shown relative to the best predictor, which in turn is 0, and therefore not visible on the bar graph. Each visible bar represents a model which performs worse than the best predictor for that experiment, with higher bars representing worse performance.

6.1.4 CNN MNIST model

Given the recent success of convolutional neural net (CNN) models in classifying images in standard databases (Jacobs & Bates, 2019, see also Farabet et al., 2010, Krizhevsky et al., 2012, Simonyan & Zisserman, 2014), a generic CNN model was included as a comparison to the other, more typically psychophysical, models evaluated here. For this purpose, generic CNN model packaged with the MatConvNet MATLAB toolbox (available at http://www.vlfeat.org/matconvnet/) was selected. The model used here (pink on the graphs) was pre-trained on the MNIST handwritten digit database, which consists of 60,000 training images and 10,000 test images. Each image is of a single handwritten digit, 0 to 9. A handwritten digit database was chosen because it was judged that handwritten characters were similar in structure to the axial 2D shape stimuli used here. The handwritten character images are of single, isolated objects with some degree of variation, and the ten digit categories are defined largely by the arrangement of parts relative to one another.

The CNN MNIST model, like most CNN models, produces a series of confidence

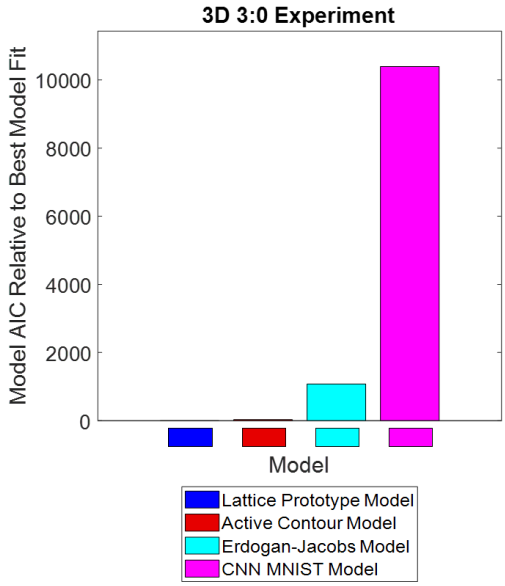

Figure 6.6: Model AIC values for the 3D 3:0 experiments. Each AIC is shown relative to the best predictor, which in turn is 0, and therefore not visible on the bar graph. Each visible bar represents a model which performs worse than the best predictor for that experiment, with higher bars representing worse performance.

ratings as its output. Each confidence rating corresponds to one of the model's learned categories, in this case, the digits 0 through 9. For the purposes of the AIC evaluation, these confidence ratings were treated as a probability distribution, and defined similarity between two shapes as the intersection of their respective distributions; that is to say, the model's confidence that both shapes would be classified as belonging to the same category. For shapes a and b and set of classes i, the similarity between a and b is described by

$$sim(\text{CNN}) \propto \sum_i p((a \in i) \cup (b \in i)) \tag{6.3}$$

6.1.5 Active contour model

The active contour model (yellow on the graphs) is a model for deforming contours step-by-step until they match other contours. In the context of shape similarity, that means deforming one shape's bounding contour to match another shape's bounding contour. This is done by finding a minimum-distance order-preserving correspondence between contour vertices in the two shapes, and then adjusting the positions of vertices

in one shape until they have the same positions. as the vertices in the other. Each of these adjustments has a certain probability, given as a Gaussian function of new vertex position, and the similarity between two shapes is proportional to the combined probability of these adjustments. This is similar in concept to the contour component of the lattice similarity model, but uses vertex positions instead of turning angle curvature.

6.2 Model comparisons

In the 2D 1:x experiments, the lattice similarity model has the lowest (best) AIC value of all evaluated models. In the 2D multiple-positive-example experiments (2:x and 3:0), the active contour model has the lowest AIC value. In all 2D cases, the lattice similarity, cross likelihood, and active contour models have substantially lower AIC values than the Erdogan-Jacobs model on this dataset. The Erdogan-Jacobs model, in turn, has a substantially lower AIC value than the CNN MNIST model in all experiments except the 2:1 experiment.

Note that these AIC values relative to one another can be converted into effective likelihood ratios, expressing the relative probability of the models given the data. Because the AIC values exist in log scale, absolute differences in AIC space correspond to exponential differences in likelihood ratio. Two AIC values that differ by even 100 points, far smaller than many of the differences seen here, correspond to a likelihood ratio of e^{100} (roughly 10^{21}) in favor of the lower-AIC model.

6.3 Exemplar vs. prototype models

When computing the model predictions for the 2:0, 2:1, and 3:0 experiments, the model had to combine information from multiple positive examples when computing similarity. This combination was implemented in two different ways, corresponding roughly to "prototype-like" and "exemplar-like" rules. In the prototype-like method, the positive examples are combined into a shape prototype and similarity is computed from the prototype to the target shape (Posner & Keele, 1968; Rosch, 1973; Homa et al., 1973; Minda & Smith, 2001). In the exemplar-like method, similarity is computed

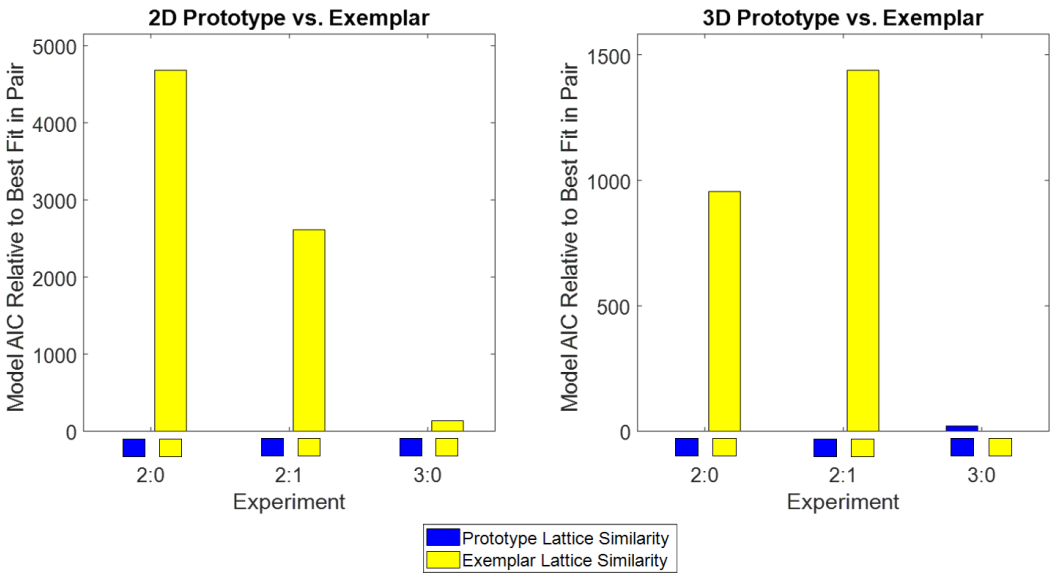

Figure 6.7: Model AIC values for the prototype vs. exemplar versions of the lattice similarity model in the experiments with multiple positive examples. The 2D experiments are shown on the left, while the 3D experiments are shown on the right. Each AIC is shown relative to the best predictor for that experiment, which in turn is 0, and therefore not visible on the bar graph. Each visible bar represents a version of the model which performs worse than the best predictor for that experiment.

from the target to each positive example, and the highest of those similarities is used for the metric (Medin & Schaffer, 1978; Nosofsky, 1986; Hintzman, 1986). Most similarity models here do not inherently include a method for inferring shape prototypes, so the exemplar version was used instead. The lattice similarity model, however, inherently proposes a formulation of the prototype, by inferring the parameters of the positive shape distribution from all positive examples. In the prototype version of the lattice similarity model, the parameters of the common distribution $\hat{C}$ are inferred using all the positive example shapes, not just one of them. Otherwise, this inference proceeds as described in Ch. 3, and the similarity is computed between this full common distribution and the target shape.

In the 2:0, 2:1, and 3:0 experiments, the performance of this prototype lattice similarity was analyzed versus the exemplar version of the lattice similarity model. Model predictions were computed for each version as with comparisons between models. The prototype version of lattice similarity out-performed the exemplar version in every experiment except for the 3D 3:0 experiment (see figure 6.7). The AIC values were

relatively close in both the 2D and 3D 3:0 experiments.

6.4 Role of the negative example

When computing the model predictions for the 1:1 and 2:1 experiments, the model needed to incorporate, or choose not to incorporate, the negative example shape. When it was incorporated, the negative example shape was included in a likelihood ratio, as described by Eq. 6.1. The negative example was also weighted by a scalar value, which was one of the fitted parameters in the model. The results of adding the negative example can be seen in Fig. 6.8. As can be seen from the figure, the incorporation of the negative example substantially improved the model fit in the 2D 2:1 experiment and the 3D 1:1 experiment, while adding little in the 2D 1:1 and 3D 2:1 experiment. Adding the negative example never negatively impacted the model's predictions, in part because the parameter fitting effectively discards variables without predictive power.

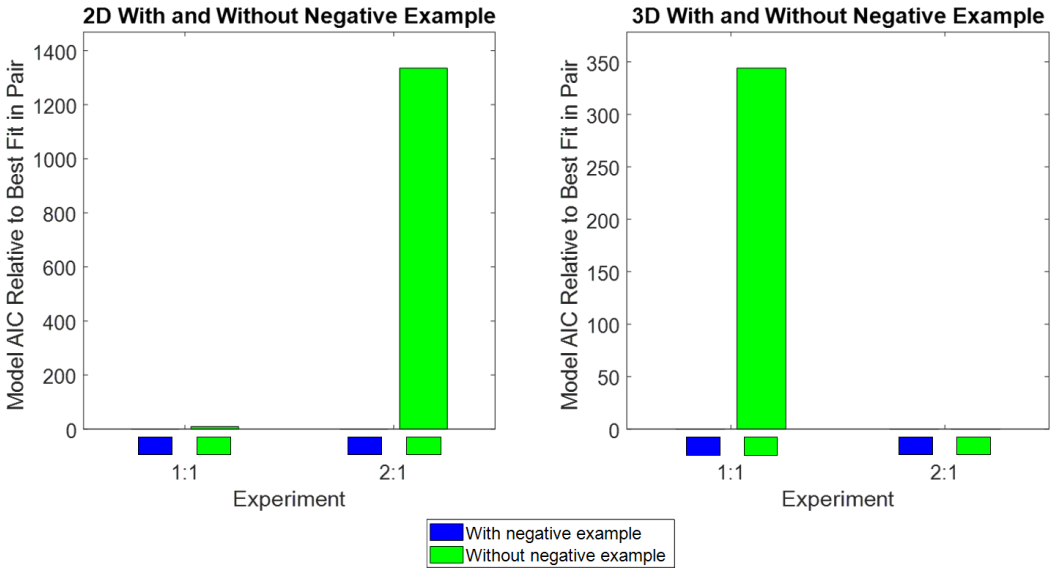

Figure 6.8: Model AIC values for versions of the lattice similarity model that included vs. did not incorporate information from the negative example shapes, in the four experiments with negative example. The 2D experiments are shown on the left, while the 3D experiments are shown on the right. Each AIC is shown relative to the best predictor for that experiment, which in turn is 0, and therefore not visible on the bar graph. Each visible bar represents a version of the model which performs worse than the best predictor for that experiment.

6.5 Overview of results

Here, lattice similarity model has predicted human classification data more accurately than other models overall, including in all of the 3D experiments and in the one-example 2D experiments. It substantially outperforms the Erdogan-Jacobs and MNIST CNN models in all experiments, but is sometimes outperformed by the Active Contour model and the Cross Likelihood model in 2D experiments.

In the experiments with multiple positive example shapes, the prototype version of the lattice similarity model strongly out-performed the exemplar version of the same model overall. However, in the 3:0 experiments specifically, the performance of the two versions was more evenly matched.

In the experiments with a negative example shape, the inclusion of the negative example in the similarity metric was sometimes, but not always, able to substantially improve the prediction of the lattice similarity model.

Chapter 7

Summary and Conclusion

In this research, I set out to create and test a new shape similarity measure, based on the simple principle that shape similarity is essentially a model comparison task. From this grounding idea, and based on skeletal- and contour-based models that have successfully predicted human shape similarity and classification judgments in past work (Briscoe, 2008; Wilder et al., 2015a, 2015b; Destler et al., 2019), I constructed a new shape similarity model that treats shape similarity as a Bayes factor model comparison, assessing similarity as the probability that two shapes were generated by a common model. In this lattice similarity model, shapes are considered similar to the extent that they are likely to be generated by this single common model of shape region and contour, vs. by two separate models. This similarity measure includes a component evaluating the change in skeletal structure, like previous graph-matching skeleton comparison techniques (Siddiqi et al., 1998; Torsello & Hancock, 2004; Demirci et al., 2006; Bai & Latecki, 2008; Rezanejad & Siddiqi, 2015), but embeds this in a full generative probabilistic model, allowing the more principled Bayes factor formulation developed above.

In order to test this new model, I created a new experimental framework for rapidly testing shape classification judgments. Subjects were shown example shapes that were either labelled as members of a novel shape category, or as non-members of a novel shape category. The subjects were then asked which target shapes from a grid of 36 were also members of this category. This technique allowed the rapid collection of data from subjects over the internet using Amazon Mechanical Turk, which in turn allowed the running of ten different experiments.

These ten experiments varied in both the dimensions of the stimuli (2D shapes vs.

3D shapes) and the number of positive and negative example shapes shown to the subjects. By varying the numbers of positive and negative shape examples, I hope to shed light on how category examples are used to infer mental shape categories. And by experimenting in both 2D and 3D, I aimed to demonstrate the extent to which common principles govern both 2D and 3D shape classification.

The 3D experiments also put into practice the skeleton inflation techniques proposed by Feldman et al. (2013) and Twarog et al. (2012). While further work is required to develop a full 3D version of the Bayesian skeletons used in the lattice similarity model, I have created a way to infer 3D shapes from known skeletons, and here it has been shown that these skeletons can then be used as reasonable approximations of the true skeletons in shape similarity calculations. In fact, even relying on this approximation, the lattice similarity model performed better relative to other models in the 3D experiments than in the 2D experiments.

7.1 Key findings

In these experiments, the lattice similarity model made the best predictions of all tested models in the most straightforward 2D classification experiments, 1:0 and the 1:1. This success is important, because it demonstrates that the model is highly successful at predicting how shape similarity relates to shape classification in the purest case. The lattice similarity model also made the best predictions in all of the 3D classification experiments, which lends strong support not only to the lattice similarity model, but to the practical applicability of the new new 3D inflation technique. Overall, the lattice similarity showed the best performance of all models in seven out of the ten experiments, though it was out-performed by the active contour and cross likelihood models in three of the 2D experiments.

While the lattice similarity model was out-performed by the active contour model on the 2D experiments with multiple positive examples, the lattice similarity model still substantially out-performed the Erdogan-Jacobs and CNN MNIST models in all conditions, 2D and 3D. The relative success of both this skeleton-contour hybrid model

and the purely contour-based active contour model suggests that further exploration may be needed into the role of contour geometry in the inference of shape categories from multiple examples. However, the consistent success of the lattice similarity model in the 3D domain suggests that shape skeletons play an important role in 3D shape perception.

The Erdogan-Jacobs model and especially the CNN model had difficulty predicting human shape classifications in these experiments. The relatively poor performance of the Erdogan-Jacobs model is especially surprising given that model's success on a previous classification task (Erdogan & Jacobs, 2017), and may result from differences in the shape structure of the stimuli. The more serious failure of the CNN model in these experiments is less surprising. CNNs are traditionally trained on upwards of tens of thousands of category examples (Farabet et al., 2010; Cireşan et al., 2012; Dean et al., 2012; Krizhevsky et al., 2012), while humans can infer categories from as few as one example (Macario et al., 1990; Greene, 2018), as subjects were asked to do in this experiment. The failure of the CNN model to approximate human performance can thus easily be attributed to it being given an inapt training set, but is still significant in light of how strongly CNN models have been advocated as models of human visual classifications (Yamins et al., 2014), and the lack of task-specific training data for the human subjects beyond the specific category examples.

More subtly, these experiments showed that the lattice similarity prototype model out-performed the lattice similarity exemplar model decisively in the 2:0 and 2:1 experiments. This suggests that subjects do, at least in part, infer shape categories by constructing a common generative model based on the example shapes, as the prototype version constructs an overt "common model" while the exemplar version does not. However, the 3:0 experiments were less conclusive, indicating that subjects may use a more exemplar-like technique to infer categories with larger numbers of positive examples.

Finally, performance was improved for the lattice similarity in some of the 1:1 and 2:1 experiments by the incorporation of negative example information. Specifically, incorporating the negative example similarity as the denominator of a likelihood ratio

led to improved model prediction in two of the four experiments, and effectively-equal model prediction in the other two. This suggests that humans may incorporate negative example information according to the rules of rational model comparison, but if so, there are additional factors that impact when and how the negative shape example is used. The exact nature of these factors remains an important subject for future research.

7.2 Future directions

This work has highlighted the need for, and benefits of, a similarity model grounded in the mathematical principles that underlie model comparison and belief updating. I have shown here that the math of Bayesian model comparison is an effective framework from which to understand shape similarity, and opened the door for mathematically-motivated models of other key aspects of shape and object perception.

The details of the lattice similarity model have also lent support to previous findings that both shape skeletons (Briscoe, 2008; Destler et al., 2019) and contour curvature (Wilder et al., 2015a, 2015b) play an important role in shape similarity and classification. While skeletons and contours are often represented as different models, a full model of human shape perception can and must account for both. Indeed, in the lattice similarity framework, skeleton and contour models are treated in an exactly parallel fashion, with an analogous degree of fit comparison (the Bayes factor), but with different generative models.

Another attractive feature of the lattice similarity framework is that a shape "prototype," the common model, can be calculated for any group of shapes, using only the same operations that the model applies for "ordinary" shape similarity calculations. The experimental findings reported here have supported the psychological reality of this prototype model, in that it predicts human performance more accurately than the exemplar version of the lattice similarity model. However, as the number of category example shapes increases, the advantage of this prototype process appears to decrease. This pattern may suggest that exemplar models are favored as the amount of available category information increases (see McKinley & Nosofsky, 1995, Minda & Smith,

2001, and Briscoe & Feldman, 2011) but further testing on larger numbers of category examples is necessary to draw any definitive conclusions.

While these results have illuminated much about how positive category examples are used, the role of negative category examples remains largely unclear. I have demonstrated that information from negative examples, incorporated in the form of a likelihood ratio, is often useful in predicting human classifications, but this utility is inconsistent. Further investigation will be required to fully illuminate when, and how, humans make use of negative category examples.

Finally, I developed a new experimental framework for rapidly testing human classifications of large numbers of shapes. This framework can be used either in-person or, as it was used here, over an online data collection service such as Amazon Mechanical Turk. I hope that in the future, this framework can be used to build on what I have demonstrated here, and bring new insight into the domain of shape and object perception.

Bibliography

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265.

Akaike, H. (1974). A new look at the statistical model identification. In *Selected papers of hirotugu akaike* (pp. 215–222). Springer.

Akaike, H. (1979). A bayesian extension of the minimum aic procedure of autoregressive model fitting. *Biometrika*, *66*(2), 237–242.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 50.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological review*, *95*(1), 124.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, *61*(3), 183.

August, J., Siddiqi, K., & Zucker, S. W. (1999). Ligature instabilities in the perceptual organization of shape. In *Computer vision and pattern recognition, 1999. ieee computer society conference on.* (Vol. 2, pp. 42–48).

Ayzenberg, V., Chen, Y., Yousif, S. R., & Lourenco, S. F. (2019). Skeletal representations of shape in human vision: Evidence for a pruned medial axis model. *Journal of vision*, *19*(6), 6–6.

Ayzenberg, V., & Lourenco, S. F. (2019). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific Reports*, *9*(1), 9359.

Bai, X., & Latecki, L. J. (2008). Path similarity skeleton graph matching. *IEEE*

transactions on pattern analysis and machine intelligence, *30*(7), 1282–1292.

Barenholtz, E., & Tarr, M. J. (2008). Visual judgment of similarity across shape transformations: Evidence for a compositional model of articulated objects. *Acta psychologica*, *128*(2), 331–338.

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, *24*(4), 509–522.

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

Blake, A., & Isard, M. (2012). *Active contours: the application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion*. Springer Science & Business Media.

Blum, H. (1967). A transformation for extracting new descriptors of shape. *Models for the perception of speech and visual form*, *19*(5), 362–380.

Blum, H. (1973). Biological shape and visual science (part i). *Journal of Theoretical Biology*, *38*, 205–287.

Blum, H., & Nagel, R. N. (1978). Shape description using weighted symmetric axis features. *Pattern recognition*, *10*(3), 167–180.

Briscoe, E. (2008). *Shape skeletons and shape similarity* (Unpublished doctoral dissertation). Rutgers University.

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, *118*, 2–16.

Burbeck, C. A., & Pizer, S. M. (1995). Object representation by cores: Identifying and representing primitive spatial regions. *Vision research*, *35*(13), 1917–1930.

Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention* (pp. 411–418).

Cireşan, D. C., Meier, U., Masci, J., & Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. *Neural Networks*, *32*, 333–338.

Cohen, E. H., Singh, M., & Maloney, L. T. (2008). Perceptual segmentation and the perceived orientation of dot clusters: The role of robust statistics. *Journal of vision*, *8*(7), 6–6.

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

Dean, J., Corrado, G. S., Monga, R., Chen, K., Devin, M. D., Le, Q. V., ... Ng, A. Y. (2012). Large scale distributed deep networks. In *Advances in neural information processing systems* (pp. 1223–1231).

Demirci, M. F., Shokoufandeh, A., Keselman, Y., Bretzner, L., & Dickinson, S. (2006). Object recognition as many-to-many feature matching. *International Journal of Computer Vision*, *69*(2), 203–222.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 248–255).

Denisova, K., Feldman, J., Su, X., & Singh, M. (2016). Investigating shape representation using sensitivity to part- and axis-based transformations. *Vision Research*, *126*, 347–361.

Destler, N., Singh, M., & Feldman, J. (2019). Shape discrimination along morph-spaces. *Vision research*, *158*, 189–199.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Science*, *21*(4), 449–467.

Erdogan, G., & Jacobs, R. A. (2017). Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psychological Review*, *124*(6), 740–761.

Farabet, C., Martini, B., Akselrod, P., Talay, S., LeCun, Y., & Culurciello, E. (2010). Hardware accelerated convolutional neural networks for synthetic vision systems. In *Proceedings of 2010 ieee international symposium on circuits and systems* (pp. 257–260).

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

Feldman, J., & Singh, M. (2005). Information along contours and object boundaries. *Psychological review*, *112*(1), 243.

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences*, *103*(47), 18014–18019.

Feldman, J., Singh, M., Briscoe, E., Froyen, V., Kim, S., & Wilder, J. (2013). An integrated bayesian approach to shape representation and perceptual organization. In *Shape perception in human and computer vision* (pp. 55–70). Springer.

Firestone, C., & Scholl, B. J. (2014). please tap the shape, anywhere you like shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological science*, 0956797613507584.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, *4*(1), 1–58.

Greene, E. (2018). New encoding concepts for shape recognition are needed. *AIMS NEUROSCIENCE*, *5*(3), 162–178.

Hintzman, D. L. (1986). " schema abstraction" in a multiple-trace memory model. *Psychological review*, *93*(4), 411.

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

Homa, D., Cross, J., Cornell, D., Goldman, D., & Shwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology*, *101*(1), 116.

Hung, C.-C., Carlson, E. T., & Connor, C. E. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, *74*(6), 1099–1113.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological review*, *98*(3), 352.

Jacobs, R. A., & Bates, C. J. (2019). Comparing the visual representations and performance of humans and deep neural networks. *Current Directions in Psychological Science*, *28*(1), 34–39.

Katz, R. A., & Pizer, S. M. (2003). Untangling the blum medial axis transform. *International Journal of Computer Vision*, *55*(2-3), 139–153.

Kellman, P. J., Garrigan, K., & Erlikhman, G. (2015). Challenges in understanding visual shape perception and representation: Bridging subsymbolic and symbolic coding. In S. Dickinson & Z. Pizlo (Eds.), *Shape perception in human and computer vision* (p. 249-274). Springer.

Kovács, I., Fehér, Á., & Julesz, B. (1998). Medial-point description of shape: a representation for action coding and its psychophysical correlates. *Vision research*, *38*(15), 2323–2333.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lehky, S. R., & Tanaka, K. (2016). Neural representation for object recognition in inferotemporal cortex. *Current Opinion in Neurobiology*, *37*, 23–35.

Leymarie, F. F., & Kimia, B. B. (2001). The shock scaffold for representing 3d shape. In *International workshop on visual form* (pp. 216–227).

Liu, T.-L., & Geiger, D. (1999). Approximate tree matching and shape similarity. In *Computer vision, 1999. the proceedings of the seventh ieee international conference on* (Vol. 1, pp. 456–462).

Liu, T.-L., Geiger, D., & Kohn, R. (1998). Representation and self-similarity of shapes. In *Computer vision, 1998. sixth international conference on* (pp. 1129–1135).

Loomis, J. M., & Philbeck, J. W. (1999). Is the anisotropy of perceived 3-d shape invariant across scale? *Perception & Psychophysics*, *61*(3), 397–402.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Macario, J. F., Shipley, E. F., & Billman, D. O. (1990). Induction from a single instance: Formation of a novel category. *Journal of experimental child psychology*, *50*(2), 179–199.

Macrini, D., Dickinson, S., Fleet, D., & Siddiqi, K. (2011). Object categorization using bone graphs. *Computer Vision and Image Understanding*, *115*(8), 1187–1206.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision

bound models in large, ill-defined category structures. *J. Experimental Psychology: Human perception and performance*, *21*(1), 128–148.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775.

Mustafa, W., Pugeault, N., & Krüger, N. (2013). Multi-view object recognition using view-point invariant shape relations and appearance information. In *Robotics and automation (icra), 2013 ieee international conference on* (pp. 4230–4237).

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (pp. 427–436).

Norman, J. F., Phillips, F., & Ross, H. E. (2001). Information concentration along the boundary contours of naturally shaped solid objects. *Perception*, *30*(11), 1285–1294.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT press.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*(3p1), 353.

Radenovic, F., Tolias, G., & Chum, O. (2018). Deep shape matching. In *Proceedings of the european conference on computer vision (eccv)* (pp. 751–767).

Rezanejad, M., & Siddiqi, K. (2015). Flux graphs for 2d shape analysis. In S. Dickinson & Z. Pizlo (Eds.), *Shape perception in human and computer vision* (p. 41-54). Springer.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current opinion in neurobiology*, *12*(2), 162–168.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*, 328–350.

Sebastian, T. B., & Kimia, B. B. (2005). Curves vs. skeletons in object recognition.

Signal Processing, *85*, 247–263.

Sebastian, T. B., Klein, P. N., & Kimia, B. B. (2001). Recognition of shapes by editing shock graphs. In *Iccv* (Vol. 1, pp. 755–762).

Shepard, R. N. (1957). Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Siddiqi, K., & Kimia, B. B. (1996). A shock grammar for recognition. In *Computer vision and pattern recognition, 1996. proceedings cvpr'96, 1996 ieee computer society conference on* (pp. 507–513).

Siddiqi, K., Kimia, B. B., Tannenbaum, A., & Zucker, S. W. (1999). Shapes, shocks and wiggles. *Image and vision computing*, *17*(5), 365–373.

Siddiqi, K., Shokoufandeh, A., Dickinson, S. J., & Zucker, S. W. (1998). Shock graphs and shape matching. In *Proceedings of the sixth international conference on computer vision* (p. 222). IEEE Computer Society.

Siddiqi, K., Tresness, K. J., & Kimia, B. B. (1996). Parts of visual form: psychophysical aspects. *Perception*, *25*, 399–424.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, R., Papanikolopoulos, N. P., & Cherkassky, V. (1998). Object skeletons from sparse shapes in industrial image settings. In *Proceedings. 1998 ieee international conference on robotics and automation (cat. no. 98ch36146)* (Vol. 4, pp. 3388–3393).

Tam, R., & Heidrich, W. (2003). Shape simplification based on the medial axis transform. In *Visualization, 2003. vis 2003. ieee* (pp. 481–488).

Telea, A., Sminchisescu, C., & Dickinson, S. (2004). Optimal inference for hierarchical skeleton abstraction. In *Pattern recognition, 2004. icpr 2004. proceedings of the 17th international conference on* (Vol. 4, pp. 19–22).

Torsello, A., & Hancock, E. R. (2004). A skeletal measure of 2d shape similarity.

Computer Vision and Image Understanding, *95*(1), 1–29.

Twarog, N. R., Tappen, M. F., & Adelson, E. H. (2012). Playing with puffball: simple scale-invariant inflation for use in vision and graphics. In *Proceedings of the acm symposium on applied perception* (pp. 47–54).

Wagemans, J. (2015). Two-dimensional shape as a mid-level vision gestalt. In S. Dickinson & Z. Pizlo (Eds.), *Shape perception in human and computer vision* (p. 85-102). Springer.

Wagenmakers, E.-J., & Farrell, S. (2004). Aic model selection using akaike weights. *Psychonomic bulletin & review*, *11*(1), 192–196.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic bulletin & review*, *8*(1), 168–176.

Wilder, J., Feldman, J., & Singh, M. (2011). Superordinate shape classification using natural shape statistics. *Cognition*, *119*(3), 325–340.

Wilder, J., Feldman, J., & Singh, M. (2015a). Contour complexity and contour detection. *Journal of Vision*, *15*(6), 1–16.

Wilder, J., Feldman, J., & Singh, M. (2015b). The role of shape complexity in the detection of closed contours. *Vision Research*.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.