Visual shape perception as Bayesian inference of 3D object-centered shape representations

Goker Erdogan

Robert A. Jacobs

Department of Brain & Cognitive Sciences

University of Rochester

Rochester, NY 14627

E-mail: {gerdogan, robbie}@bcs.rochester.edu

December 13, 2016

A preliminary version of a portion of this work was presented at the 38th Annual Conference of the Cognitive Science Society, and a conference paper appears in the proceedings of that conference. This work was supported by research grants from the Air Force Office of Scientific Research (FA9550-12-1-0303) and the National Science Foundation (BCS-1400784).

Abstract

Despite decades of research, little is known about how people visually perceive object shape. We hypothesize that a promising approach to shape perception is provided by a “visual perception as Bayesian inference” framework which augments an emphasis on visual representation with an emphasis on the idea that shape perception is a form of statistical inference. Our hypothesis claims that shape perception of unfamiliar objects can be characterized as statistical inference of 3D shape in an object-centered coordinate system. The unusual (possibly novel) combination of a probabilistic-inference approach and a 3D object-centered approach is shown to lead to surprising results, including the fact that probabilistic object-centered representations can underlie viewpoint-dependency. This result suggests that the distinction between view-based and view-independent representations is less useful than commonly believed when applied to the study of viewpoint invariance. We describe a computational model based on our theoretical framework, and provide evidence for the model along two lines. First, we show that, counterintuitively, the model accounts for viewpoint-dependency of object recognition, traditionally regarded as evidence against people’s use of 3D object-centered shape representations. Second, we report the results of an experiment using a shape similarity task, and present an extensive evaluation of existing models’ abilities to account for the experimental data. We find that our shape inference model captures subjects’ behaviors better than competing models. Taken as a whole, our experimental and computational results illustrate the promise of our approach and suggest that people’s shape representations of unfamiliar objects are probabilistic, 3D, and object-centered.

Keywords: visual perception; object perception; shape perception; experimentation; computational modeling

1 Introduction

Consider the objects in Figure 1. Even though you have not previously encountered these objects, you can readily perceive that the object in Figure 1c is more similar to the object in Figure 1a than the object in Figure 1b. However, the ease with which people make this judgment belies the complexity of the mental operations involved in this task. People's visual systems need to extract a representation of these objects from 2D images, and compare these representations to make a similarity judgment. This task illustrates the essence of the computational problem of object shape perception.

How people perceive object shape is one of the most fundamental questions about human visual perception. However, as evidenced by decades of research, this simple question is surprisingly difficult to answer. Researchers have proposed numerous hypotheses about shape perception, and much research has focused on proving or disproving particular hypotheses. These efforts have led the field toward theoretical dichotomies such as whether people's shape representations are "view-based" or "structural", or whether these representations code two-dimensional or three-dimensional information. To date, investigations into such dichotomies have rarely produced clear outcomes. For example, after a long line of research on whether people's shape representations are view-based or structural, Peissig and Tarr (2007) summarized the state of the debate as follows: "In the end, it is unclear whether the large body of work focused on view-based models is compatible with, incompatible with, or just orthogonal to structural models of object representation". Which approach, if either, properly characterizes human shape perception is still a matter of fierce debate.

Here, we argue that existing models of shape perception are inadequate in important respects, and we propose a new model based on the hypothesis that shape perception of unfamiliar objects can be best understood as Bayesian inference of 3D shape in an object-centered coordinate system. This hypothesis includes four important components: (i) Our hypothesis is a hypothesis about shape representations of **unfamiliar** objects. Shape representations of familiar objects might be best understood in other ways. Coverage of this topic is deferred until the “Discussion” section. (ii) Shape perception for unfamiliar objects is a form of statistical inference which can be characterized as Bayesian inference. This implies that people’s shape representations are probabilistic, and thus contain information about certainty or confidence. For example, the shape properties of one portion of an object (e.g., the portion of an object facing a viewer) might be represented with high certainty, whereas the shape properties of another portion of the same object (e.g., a portion seen in peripheral vision, or a portion that is partially or fully occluded) might be represented with low certainty. It also implies that shape representations are influenced by a person’s prior beliefs about shape properties. (iii) Shape representations code information about an object’s three-dimensional structure, not the two-dimensional structure of its retinal image. (iv) Shape representations code shape properties in an object-centered coordinate system, not a viewer-centered coordinate system.

Although each of these components has been studied previously in the scientific literature, their combination has not. Indeed, as demonstrated below, their combination gives rise to interesting and unexpected results. For example, we have found that probabilistic object-centered representations can underlie viewpoint-dependency, suggesting that the distinction between view-based and view-independent representations is less useful than commonly believed when applied to the study of viewpoint invariance.

This article provides support for our hypothesis along two lines. First, we show that the use of 3D object-centered shape representations does **not** imply viewpoint-invariant object recognition. As demonstrated below, a person may, for example, attempt to infer a 3D object-centered shape representation from a 2D image in which one portion of a viewed object is clearly visible whereas another portion is not. If shape representations are treated in a probabilistic manner, the person’s shape representation will have high certainty about shape properties in the former portion and low certainty about shape properties in the latter portion, thereby leading to viewpoint-dependent object recognition. We find that a computational model based on our hypothesis successfully accounts for the finding

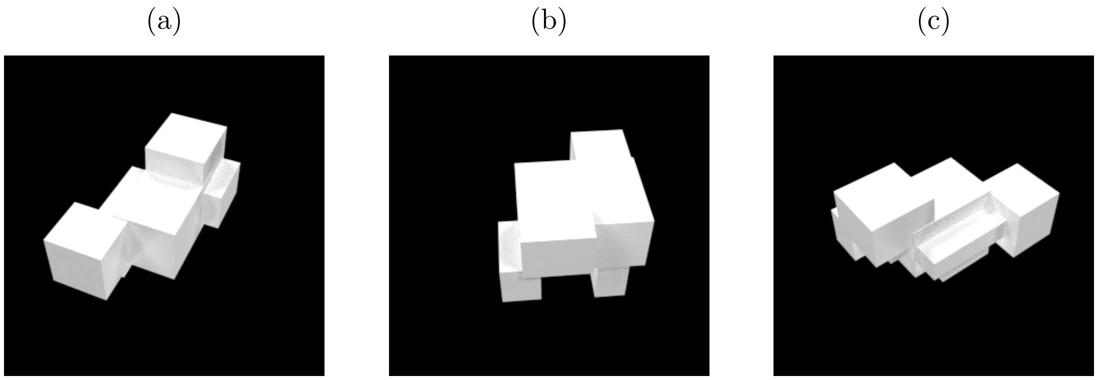


Figure 1: Is the shape of the middle or rightmost object more similar to the shape of the leftmost object?

that people's object recognition performances can be viewpoint-dependent. Consequently, viewpoint-dependency should not be regarded as evidence for a view-based account of object recognition, as is typically done in the scientific literature.

Second, we report the results of an experiment using a shape similarity task, and evaluate a broad array of existing models of shape perception for their abilities to account for the experimental data. This evaluation provides compelling empirical support for our 3D object-centered shape inference model. Because the model captures subjects' judgments better than its competitors, our results support the hypothesis that people's object shape representations for unfamiliar objects are probabilistic, 3D, and object-centered. We conclude that our hypothesis is unique in its explanatory power and scope, and provides a promising approach for future investigations of object shape perception.

2 Theoretical Background

It is frustratingly difficult to present a clear and well-organized analysis of hypotheses on shape perception. This is mostly because research on shape perception has revolved around dichotomies that are rarely rigorously defined, such as whether shape representations code 2D or 3D information, whether these representations are view-based or view-independent, or whether these representations are holistic or structural. These poorly defined dichotomies make the boundaries between different hypotheses hard to discern. In this section, we follow the analysis provided by Palmer (1999) and discuss three classes of shape perception hypotheses: feature-based, view-based, and structural description hypotheses. We present a critical review of each class, highlighting a class's strengths and weaknesses. For each class, we first present its main claims and then discuss computational models based on that class.

2.1 Feature-based hypotheses

Feature-based hypotheses claim that object shape is represented by a list of feature values extracted from 2D input images. These values are calculated by feature extractors through multiple layers of processing in the visual system. To compare the shapes of objects, one needs to specify a procedure for evaluating the similarity between two feature-based representations. In concrete models using a feature-based approach, feature values are usually real-valued and dissimilarity is quantified as Euclidean distance between representations. Feature-based hypotheses take their inspiration directly from what we know about biological visual systems, and this class of hypotheses represents the dominant perspective in the field of neuroscience. Building on the early work of Hubel and Wiesel (1962), neuroscientists have investigated visual perception by seeking to understand the neural feature detectors implemented by our visual systems. To date, this project faces major challenges in understanding cortical regions beyond primary visual cortex (Kourtzi & Connor, 2011).

To be meaningful, a feature-based hypothesis needs to specify the particular features that the hypothesis claims to be involved in shape perception. One popular proposal claims that what characterizes these features is that they are invariant to shape-preserving transformations such as translation and rotation (Palmer, 1999). Previous research has shown that some neurons in inferotemporal cortex (IT) are significantly position and scale invariant (Riesenhuber & Poggio, 2002). However, recent research suggests that the extent of the invariance exhibited by these neurons is significantly less than previously believed (Lehky & Tanaka, 2016). Moreover, the naive invariance hypothesis cannot be the whole story because features that are fully invariant to shape-preserving transformations are inadequate for visual object recognition. For example, features that are fully position-invariant cannot distinguish between two objects that consist of the same features but in different spatial arrangements.

2.1.1 Feature-based models

In the field of computational neuroscience, an influential example of a feature-based model is Riesenhuber and Poggio (1999)'s HMAX (hierarchical MAX) model. HMAX extends Hubel and Wiesel (1962)'s ideas about simple and complex cells to higher level visual areas by proposing a sequence of template matching and pooling operations that build position and scale invariant features. HMAX consists of alternating layers of what are called S and C layers. Units in an S layer implement template matching. These templates can be simple Gabor filters (as in early layers) or more complex features (as in later layers) that are either specified by hand or learned. C layers play a key role in building invariant features since these pool over multiple units in the previous S layer and apply "max-pooling" (i.e., select the maximum input activation). By pooling over units tuned to different positions and scales, HMAX builds position and scale invariant features. Riesenhuber and Poggio (1999) showed that HMAX captures tuning and invariance properties of IT neurons, and later work provided further evidence that HMAX is a good model of higher level processing in biological visual systems (Cadieu et al., 2007; Riesenhuber & Poggio, 2000, 2002; Serre, Oliva, & Poggio, 2007; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007).

Feature-based hypotheses are also popular in the study of computer vision. Recently, multi-layer artificial neural networks known as convolutional neural networks (CNNs) have achieved state-of-the-art object categorization performances (Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, & Hinton, 2015). These models are similar to HMAX in the sense that they implement a sequence of feature extraction and pooling operations. However, these models are much deeper (containing tens to hundreds of layers), and features are learned from large amounts of labeled image data to maximize performance. Given their successes in computer vision and their similarity to hierarchical processing in biological visual systems, recent work in cognitive science and neuroscience has started to investigate the extent to which these models provide insights into biological vision (Kriegeskorte, 2015). Khaligh-Razavi and Kriegeskorte (2014) compared a large set of models from computer vision and computational neuroscience (including HMAX) on how well they account for human fMRI and monkey neural data from cortical area IT. Results showed that AlexNet (Krizhevsky et al., 2012), a popular CNN trained on 1.2 million images, captured the most variance in IT activities. In a related study, Cadieu et al. (2014) showed that CNNs rival the representational performance of IT, matching the object categorization performance of neural responses from IT.

Feature-based hypotheses are appealing in multiple respects. From a neuroscience perspective, they build object representations hierarchically through multiple layers of processing, and thus resemble biological visual systems. They have been found to provide useful models of neural processing at all levels of the visual cortical hierarchy. From an engineering perspective, CNN implementations of feature-based hypotheses provide state-of-the-art performances, sometimes achieving object recognition and categorization performances comparable to those of people. Additionally, these implementations are appealing because they do not require preprocessing of the input image, and they can work directly on natural images.

The main weakness of feature-based hypotheses is that they are too unconstrained. Many feature-based models, such as CNNs, use adaptive features that are learned from data to maximize performance on a specified task. The shape perception procedure acquired by a feature-based model is determined by its training, including its training data and adaptation procedure (e.g., loss function and optimization procedure). Therefore, a feature-based model needs to specify not only its structural architecture (e.g., how many layers of units, how are units in one layer connected to units in the next layer, etc.), but also its training procedure in detail. Even when these details are specified, there is reason to doubt whether current feature-based models provide good scientific models of biological shape perception. These models usually have large numbers of parameters (e.g., 60 million in Krizhevsky et al., 2012) that adapt with nonlinear dynamics, meaning that the models are complex. To date, it is nearly impossible to know how and why these models achieve what they achieve. Understanding why feature-based models work so well is the focus of much current research (Anselmi, Rosasco, Tan, & Poggio, 2015; Mehta & Schwab, 2014; Patel, Nguyen, & Baraniuk, 2015; Yuille & Mottaghi, 2016).

2.2 View-based hypotheses

View-based hypotheses claim that people's shape representation for an object consists of a collection of memorized "views" of the object from different viewpoints. Recognition is achieved by comparing the observed view of an object to these stored views. View-based hypotheses focus on this comparison procedure rather than on how each view of an object is mentally represented. Indeed, view-based hypotheses are agnostic with respect to how views are represented (referred to as the "view encoding scheme"; see Tarr & Bülthoff, 1995, and Edelman, 1997). Different instantiations of view-based hypotheses have proposed different view comparison procedures (see below).

View-based hypotheses are motivated primarily by experimental findings demonstrating that visual object recognition performance can depend on the viewpoint from which an object is observed (Edelman, Bülthoff, & Weinshall, 1989; Edelman & Bülthoff, 1992; Rock & DiVita, 1987; Tarr & Bülthoff, 1995; Tarr, Williams, Hayward, & Gauthier, 1998). These studies have shown that it becomes harder to recognize an object as it is rotated away from its training view. View-dependent recognition has been presented as evidence for view-based hypotheses, and proponents of view-based hypotheses have argued that their findings provide strong evidence against approaches that use 3D, object-centered shape representations. However, as we discuss below in more detail, this view has been challenged by various researchers, and we demonstrate below with our simulation study that 3D, object-centered shape representations can in fact give rise to viewpoint dependency.

2.2.1 View-based models

Although view-based hypotheses do not make representational commitments, most view-based models have assumed that views are stored as lists of 2D features. These models have focused on how a test image can be compared with the stored 2D views in order to recognize objects. The "alignment-based" approach (Ullman, 1989) claims that the similarity between two view-based representations is calculated by first aligning the views and then comparing them. The alignment step aims to achieve robustness to shape-preserving transformations (e.g., scaling, translation, rotation), thereby enabling recognition despite such variation. Ullman (1989) has presented simple examples of how the alignment-based approach can be used to recognize objects but this model has not been evaluated for its ability to account for people's recognition performances.

Another approach is recognition by linear combination of views (Ullman & Basri, 1991). Ullman and Basri (1991) showed that under orthographic projection, views of an object span a linear subspace. Therefore, one can evaluate whether a test view depicts an object simply by checking if the test view can be represented as a linear combination of stored views of the object. Since this process requires multiple views of an object, this model cannot explain object recognition when relatively few views of an object reside in memory. For example, this model cannot recognize objects that are seen from a single view.

Another influential view-based model is that of Poggio and Edelman (1990). The model is an artificial neural network that is trained to map the input image of an object to an image depicting what the object would look like from a canonical viewpoint. The network is a "radial basis function" network in which the basis functions are centered around the stored views. The model has been used to replicate the experimental findings in Bülthoff and Edelman (1992) demonstrating that people's object recognition performances can be viewpoint-dependent. Despite its strengths, the model can be regarded as unsatisfactory in multiple respects. First, one needs hundreds of views of an object to train the network (Longuet-Higgins, 1990). Even if this might be possible for objects we encounter daily, it does not explain how people recognize objects that are seen only a few times or perhaps only once. Second, the model requires a separate network to be trained for each object. Even if this is plausible, training separate networks for each object ignores generalization across objects.

All view-based models suffer from a common problem—they all assume that the same set of features can be extracted from all views. This requires determining the same set of features in all views, and also the correspondences between features across different views. Ullman (1989) argued that our visual systems can achieve this feature extraction easily. However, Poggio and Edelman (1990) admitted that this is a non-trivial task. It might be easy to extract and match features in the case of simple images, but it is unclear whether feature extraction and matching can be so easily achieved in natural settings.

2.3 Structural description hypotheses

Structural description hypotheses claim that object shape can be analyzed using a finite set of simple shape primitives. The structural description of an object consists of a list of the primitives making up that object and the spatial relations among them. A structural description model needs to specify three components: the structural description format (i.e., the set of primitives and possible spatial relations between primitives); the shape extraction procedure (i.e., how structural descriptions are extracted from 2D images); and the shape comparison procedure (i.e., how similarity between structural descriptions is measured). In principle, the structural description of an object can characterize either 2D or 3D information in either viewer-centered or object-centered coordinate systems. However, structural description hypotheses have almost always used 3D, object-centered shape representations. Structural description hypotheses, along with the opposing view-based hypotheses, were the subject of fierce debate during the 1980s and 1990s (Biederman & Gerhardstein, 1993, 1995;

Tarr & Bülthoff, 1995). The main point of contention was the viewpoint dependence of object recognition. Structural description hypotheses were interpreted as implying that recognition would be viewpoint invariant since a full 3D, object-centered shape representation is used in the recognition process. However, as we have remarked above and will discuss in detail below, this conclusion is mistaken. 3D, object-centered representations can, in fact, account for viewpoint-dependency.

2.3.1 Structural description models

Structural description models have a long history starting with the early works of Binford (1971) and Marr and Nishihara (1978). Arguably the most famous and detailed proposal is Biederman's recognition-by-components (RBC) theory (Biederman, 1987, 2007). RBC claims that objects are represented as collections of 3D volumetric primitives called geons and the spatial relations among them. Crucially, structural descriptions in RBC represent shape only qualitatively. Geons do not encode metric properties such as the exact values of a part's width, height, depth, or aspect ratio. Similarly, relations between geons are encoded in coarse terms such as above, below, left-of, and right-of. Biederman (1987) presented a detailed account of the structural description format and a sketch of how these representations might be extracted from 2D images on the basis of "non-accidental" features. Similarity between two structural representations was assumed to depend on the degree of match between representations, but the similarity measure was not specified in detail.

RBC has been at the center of the debate between structural description and view-based hypotheses. It has been criticized because it fails to explain viewpoint-dependency. RBC predicts view-invariant recognition in Bülthoff and Edelman (1992)'s study because all stimuli used in the experiment have the same structural description. In response to this criticism, Biederman and Gerhardstein (1993) argued that RBC did not apply to the set of objects used in these experiments because RBC was intended as a model of "entry-level" categorization in which different objects have different structural descriptions and where all geons are visible in all images. Thus, in Bülthoff and Edelman (1992)'s experiment, subjects must be relying on a different shape perception mechanism.

The argument provided by Biederman and Gerhardstein (1993) is an instance of a two-process account of shape perception (Foster & Gilson, 2002; Marsolek, 1999; Palmeri & Gauthier, 2004). According to such an account, shape perception consists of two distinct processes. One is responsible for what is usually called "metric" recognition (mainly concerned with within-category discrimination, such as discrimination of objects that differ in metric properties such as length, size, and aspect ratio). The second process is responsible for discriminating between objects that are qualitatively different (e.g., across category discrimination). Biederman and Gerhardstein (1993) argued that RBC concerns this non-metric, qualitative recognition process. For this process, one should expect view-invariant recognition given that all geons of an object are visible in an image. However, although acknowledging that such a two-process system is possible, Tarr and Bülthoff (1995) argued that Biederman's theory failed to explain what it purported to explain. There are examples of objects (e.g., cow and horse) that have the same geon structural description but nonetheless belong to different categories. Additionally, Biederman's two-process account, although plausible, is far from elegant. It is unclear why there should be two processes in the first place, apart from the fact that RBC fails to adequately acount for the data from some experiments. Obviously, a far more satisfactory theory would capture both metric and non-metric recognition, and explain under which circumstances viewpoint-dependency is or is not obtained.

Overall, the strength of structural description hypotheses lies in the richness of their representations. Experimental data indicates that people seem to think of many natural objects as composed of parts (Tversky & Hemenway, 1984), some of which may be considered objects in their own right. For example, people think of bodies as consisting of parts such as limbs, torso, and head. Structural descriptions capture the compositionality of many objects in a natural manner. Compositionality is also crucial for efficiency since object representations can refer to other object representations, and object parts can be shared across objects. Additionally, structural descriptions make information about shape explicit. For example, a structural description model can discriminate objects and also explain why they are different. However, the power of structural description hypotheses can also be considered their weakness. The shape extraction problem is very difficult when the goal is to extract rich shape representations from realistic 2D images, and this might explain why there have been so few implementations of structural description hypotheses (Hummel & Biederman, 1992). Perhaps more importantly, it is unclear whether such powerful representations are needed for shape perception. One might argue that structural description hypotheses make the shape perception problem more difficult than is necessary in many circumstances, and people could do well enough at object recogntion with simpler representations.

This section has presented a critical analysis of existing hypotheses on shape perception. We believe that the above exposition shows that existing hypotheses are inadequate in important respects. This conclusion will be reinforced in Section 5 where we present an empirical evaluation of a broad array of models using data from an experiment on people's judgments of shape similarity. In the next section, we outline our own hypothesis claiming that shape perception for unfamiliar objects should be characterized as Bayesian inference of 3D object-centered shape representations.

3 Shape Perception as Bayesian Inference of 3D Object-Centered Shape Representations

Many researchers have argued that a frutiful approach to understanding biological visual perception is provided by the vision-as-inference hypothesis (Von Helmholtz, 1867). This hypothesis characterizes the task facing our visual systems as the inference problem of extracting a description of (the task-relevant portions of) the external world from the visual stimulations on our retina. Using tools from the calculus of probability, modern research has implemented and transformed this idea into the "visual perception as Bayesian inference" hypothesis (Jacobs & Kruschke, 2011; Kersten & Yuille, 2003; Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996; Yuille & Kersten, 2006). According to this hypothesis, perception is understood as the inversion of a generative model of how events in the visual environment give rise to retinal stimulations. Visual-perception-as-Bayesian-inference has been fruitfully applied to various aspects of visual perception, and past studies have shown that many perceptual phenomena can be understood from a probabilistic perspective as Bayesian inference under different probability models (Kersten & Yuille, 2003; Kersten et al., 2004; Knill & Richards, 1996). We believe that the visual-perception-as-Bayesian-inference hypothesis provides a promising approach to shape perception as well. We argue that shape perception can be best understood as the inference problem of extracting a description of object shape from 2D retinal stimulations.

The combination of this hypothesis with computational modeling provides natural cures for many of the problems we identified in our discussion of existing hypotheses in the previous section. We have seen that many models often leave important details unspecified. For example, RBC does not present an account of how two structural descriptions are compared, or view-based models do not specify how views are encoded. Building computational models forces researchers to specify their theories clearly and rigorously, and the visual-perception-as-Bayesian-inference hypothesis makes it especially easy to do so. All that is required is to specify the generative model of how causes (e.g., objects) in the world give rise to visual stimulations (i.e., images) on the retina. Once a generative model is specified, the calculus of probability provides equations for inferring the values of task-relevant variables. For instance, one can categorize or identify objects, one can judge the similarity between two shapes, and one can study the conditions under which recognition should be viewpoint-dependent versus viewpoint-invariant.

Here, we argue that shape representations for unfamiliar objects can be characterized as coding 3D shape properties in an object-centered coordinate system. An unusual feature of our approach is that these are probabilistic representations, inferred using a statistical—specifically Bayesian—inference mechanism. As a result, shape properties are random variables, meaning that their values have distributions. The variances of these distributions carry information about the certainty of knowledge regarding these properties. For instance, a shape property for the portion of an object that is clearly visible may be inferred to have a distribution with a small variance, indicating relative certainty of knowledge about this property. At the same time, a property for a portion that is less visible (e.g., it may be visible in peripheral vision, or it may be partially or fully occluded) may be inferred to have a distribution with a large variance, suggesting a lack of certainty of knowledge about this property. As discussed below, this aspect of our theory allows us to account for viewpoint-dependent object recognition (despite our theory's use of an object-centered coordinate system). In addition, our Bayesian approach implies that an observer's prior beliefs about shape properties influence his or her inferences about these properties.

To our knowledge, there are few previous articles in the psychology literature with an approach to shape perception that is closely similar to our own. In fact, the only one that we are aware of is the work of Feldman, Singh, and colleagues (Feldman & Singh, 2006; Feldman et al., 2013). These authors also treat shape perception as a form of Bayesian inference. In their model, observers infer 2D skeletal shape representations from 2D silhouettes of objects. These representations are based on medial-axis representations first introduced by Blum and Nagel (1978). Feldman et al. (2013) showed that their model is able to capture coarse shape similarity, and can also account for how some objects are decomposed into parts. While we have great admiration for this work (indeed, it has inspired our own efforts), it also has important shortcomings. To date, this model has not been tested as a general theory of object shape perception. Although Feldman et al. (2013) argued that their model can (eventually) be extended to handle 3D shape, their model is currently limited to inferring 2D shape representations. Section 5 presents an evaluation of their shape skeleton model on a shape similarity task.

4 Viewpoint-Dependency with Probabilistic 3D Object-Centered Representations

In this section, we show that a 3D object-centered shape inference model can account for the viewpoint-dependency of visual object recognition. We first discuss why 3D object-centered shape representations do not necessarily imply viewpoint-invariant recognition. Then we replicate an influential experimental finding regarding viewpoint-dependency with our shape inference model, and show that viewpoint-dependency of visual object recognition does not rule out probabilistic 3D object-centered shape representations.

Experiments showing that people's object recognition can be viewpoint dependent are often presented as evidence against shape perception models that use 3D object-centered representations. The reasoning underlying this claim is as follows. Because the 3D object-centered model of an object can be mentally rotated, recognition performance will not depend on viewpoint as long as a test object's true 3D shape representation can be extracted from the test viewpoint (Bülthoff & Edelman, 1992). In other words, differences between the viewpoint of an object at the time of study and the viewpoint of an object at the time of test can always be compensated for via mental rotation.

To us, this claim is poorly conceived. The claim assumes that the same 3D shape representation is extracted regardless of viewpoint. This is not necessarily the case and, in fact, is not perceptually (or computationally) plausible. Different views of an object are not equally informative about the object's shape. Some properties of an object's shape may be easy to infer (i.e., can be inferred with low variance or high confidence) from a particular viewpoint, but difficult to infer (i.e., are inferred with high variance) from other viewpoints. Importantly, shape properties for one portion of an object might be easy to infer from an image of the object at a particular viewpoint, whereas the properties for another portion of the object are difficult to infer from the same image. A good illustration of this point is the canonical view effect. Previous research shows that even if all views of an object

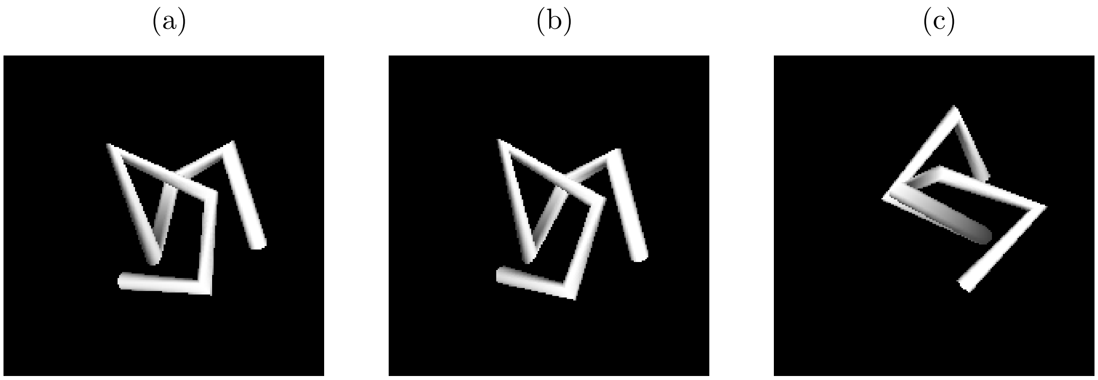


Figure 2: Three views of a paperclip object. Viewpoint differences between (a)-(b), (a)-(c), (b)-(c) are 10°, 70°, 80° respectively.

are presented an equal number of times during training, recognition performance depends significantly on viewpoint (Edelman & Bülthoff, 1992; Bülthoff, Edelman, & Tarr, 1995). These findings suggest that not all views of an object are equally informative. Therefore, one should generally expect that an observer will infer different 3D shape representations from different views of the same object. If so, one should expect object recognition to be viewpoint dependent. Furthermore, as long as the 3D shape inference procedure extracts more similar representations for closer views, one should expect object recognition to fall off gradually with viewpoint. That is, object recognition should be best when study and test viewpoints are most similar, should be moderate when these viewpoints are moderately similar, and should be worst when these viewpoints are least similar.

To illustrate these points, consider the three views of a paperclip object in Figure 2. To us, it seems intuitive that an observer's 3D shape representations for the first and second views will be more similar than the representations for the first and third views, and hence recognition will be viewpoint-dependent. We show below that is, in fact, the case for a shape inference model that infers 3D object-centered shape representations. Therefore, the use of probabilistic 3D object-centered shape representations does not imply viewpoint-invariant object recognition.

To our knowledge, similar points have been made by a few researchers in the past. Z. Liu, Kersten, and Knill (1999) and Tjan and Legge (1998) argued that not only the internal representation of shape but also the information available in the stimuli mattered for viewpoint-dependency of recognition. They presented ideal observer analyses and experimental findings that suggest, depending on the complexity of a stimulus set, one would expect object recognition to be more or less viewpoint-dependent. Even though such findings can explain why recognition performance for stimuli like paperclips are much worse than say objects made up of Biederman's geons, they do not speak to the issue of why recognition performance for a given object should get worse as the difference in viewpoint between the training and test views increases. Similarly, in a study investigating whether object representations are viewpoint-dependent, Z. Liu (1996) argued that a viewpoint-independent representation can also give rise to viewpoint-dependent performance. This point was repeated in a more recent article (Ghose & Liu, 2013). Unfortunately, neither of these articles provided an account of how this might happen. Bar (2001) also argued that viewpoint-dependency is not necessarily an indication of view-based representations. Bar (2001) presented an argument based on neural priming to show how object-centered representations can lead to viewpoint-dependent recognition. Although neural priming might be a plausible explanation for viewpoint-dependency, here we argue for an inference-based account where viewpoint-dependency follows from probabilistic inference of shape.

We show how our shape inference model accounts for viewpoint-dependency by replicating the main experimental findings from an influential study by Bülthoff and Edelman (1992). During training, subjects viewed two animations of a paperclip object. In one animation, the viewpoint of the object oscillated between $-15°$ and $15°$ around the vertical axis. In the other animation, the viewpoint oscillated between $-60°$ and $-90°$. During the test phase, subjects were presented with static test images in three conditions, and judged whether each test image depicted the same object as observed during training. In the *interpolation* condition, test viewpoints spanned the range between the two training viewpoints in 15° increments (i.e., $0°, -15°, \dots, -90°$ around the vertical axis) . In the *extrapolation* condition, test viewpoints spanned the range outside the training viewpoints in 15° increments (i.e., $0°, 15°, \dots, 90°$

around the vertical axis). Finally, in the *orthogonal* condition, test viewpoints differed from training viewpoints because they were rotations around the horizontal axis ($0°, 15°, \ldots, 90°$ around the horizontal axis). Bülthoff and Edelman (1992) argued that a view-based model predicts slower and less accurate recognition as the object is rotated away from its training views, but a recognition scheme using 3D object-centered models would predict no effect of viewpoint as long as subjects were able to extract the true 3D model from training images. They used paperclip objects comprised of multiple tubular segments to make sure that the true 3D model can, in principle, be extracted from any viewpoint (similar to the objects shown in Figures 2 and 3).

4.1 Computational model

For our simulations, we generated ten paperclip objects similar to the stimuli used by Bülthoff and Edelman (1992). Each object consisted of seven segments, and each segment's length was sampled from a normal distribution around a mean segment length. We started by placing one segment at the origin. Two new segments pointing in randomly selected directions were joined to this center segment, one on each side. These directions were selected such that the angles between segments were neither too small nor too large. We continued in this fashion by adding two segments to each end of the object until an object had seven segments. An object depicted from all simulated viewpoints is shown in Figure 3.[1]

Given the image of an object, our computational model infers the object's 3D structure in an object-centered coordinate system. In the model, an object is represented as a list of segment endpoint positions. For example, a 5-segment object shape S is represented as a list of six endpoint positions, $S = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_6\}$, with $|S|$ denoting the number of endpoints. (Although objects in our simulations always contained 7 segments, this information was not provided to the model. Instead, the model infers a posterior distribution over object shapes, meaning that shapes with, for example, 6, 7, or 8 segments might all be assigned non-zero

[1] The full set of stimuli can be seen online at `http://gokererdogan.github.io/ShapePerceptionAsBayesianInference/`.

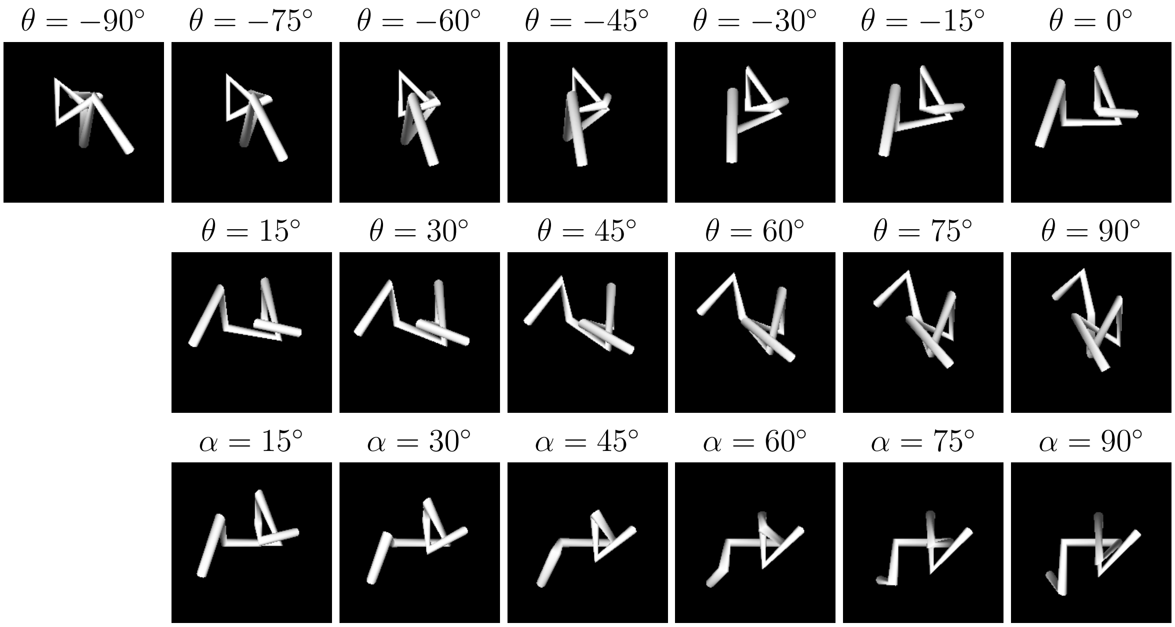


Figure 3: All views of an object used in our viewpoint-dependency simulations. θ refers to the angle around the vertical axis, and α refers to the angle around the horizontal axis.

probabilities.)

Prior distribution: In general, the model assumes that the number of segments comprising an object is sampled from a uniform distribution over integers in the interval $[2, 12]$, and that the coordinates of endpoint positions (i.e., the components of each vector $\vec{p_i}$) are sampled from a uniform distribution over $[-0.5, 0.5]$. However, without loss of generality, the model assigns the middle segment of an object to lie along the horizontal axis and to be centered at the origin. This enables the model to represent an object in a viewpoint-independent manner—that is, in an object-centered coordinate frame—and to easily "mentally" rotate the object to a canonical view if necessary. These assumptions define a prior probability distribution over possible object shapes:

$$P(S) \propto \frac{1}{|S| - 1}. \tag{1}$$

Likelihood function: To produce an image of shape S, we need to specify the viewpoint from which it is viewed. We denote viewpoint with $\vec{\phi} = (r, \theta, \alpha)$ using polar coordinates, and assume that the distance to the origin r is fixed. The prior probability distribution over viewpoint is assumed to be independent of shape S, and uniform over the sphere with radius r. The visual "forward model" $\mathcal{F} : (S, \vec{\phi}) \rightarrow I$ renders images by mapping a shape S and viewpoint $\vec{\phi}$ to image I. We implemented the forward model using the Visualization Toolkit (VTK; `http://www.vtk.org`), a software package for 3D computer graphics, image processing, and visualization. Assuming an observed image is corrupted by Gaussian pixel noise with variance σ^2, the likelihood of shape S and viewpoint $\vec{\phi}$ is:

$$P(I|(S, \vec{\phi})) \propto \exp\left(-\frac{||\mathcal{F}(S, \vec{\phi}) - I||_F^2}{\sigma^2}\right) \tag{2}$$

where $|| \cdot ||_F$ denotes the Frobenius norm.

Posterior distribution: Combining the prior distribution and likelihood function via Bayes' rule, the posterior distribution of S and $\vec{\phi}$ is:

$$P((S, \vec{\phi})|I) \propto P(S)P(\vec{\phi})P(I|(S, \vec{\phi})) \tag{3}$$

where $P(S)$ and $P(I|(S, \vec{\phi}))$ are given by Equations 1 and 2, respectively, and $P(\vec{\phi})$ is uniform. Samples from this distribution were obtained using Markov chain Monte Carlo techniques (see Appendix A for details of the sampling procedure).[2] Figure 4 provides examples of samples for three objects.

4.2 Modeling results

We evaluated the model as if it was a subject in Bülthoff and Edelman (1992)'s experiment. During the training stage of the experiment, the model inferred the posterior distribution $P((S, \vec{\phi})|I_{\text{train}})$ over 3D shapes from the set of training images. The training images I_{train} consisted of six images at views $\theta \in \{-90°, -75°, -60°, -15°, 0°, 15°\}$. On a test trial, the

[2] Implementation of our 3D shape inference model is available online at `https://github.com/gokererdogan/Infer3DShape/releases/tag/ro3Dpaper`.

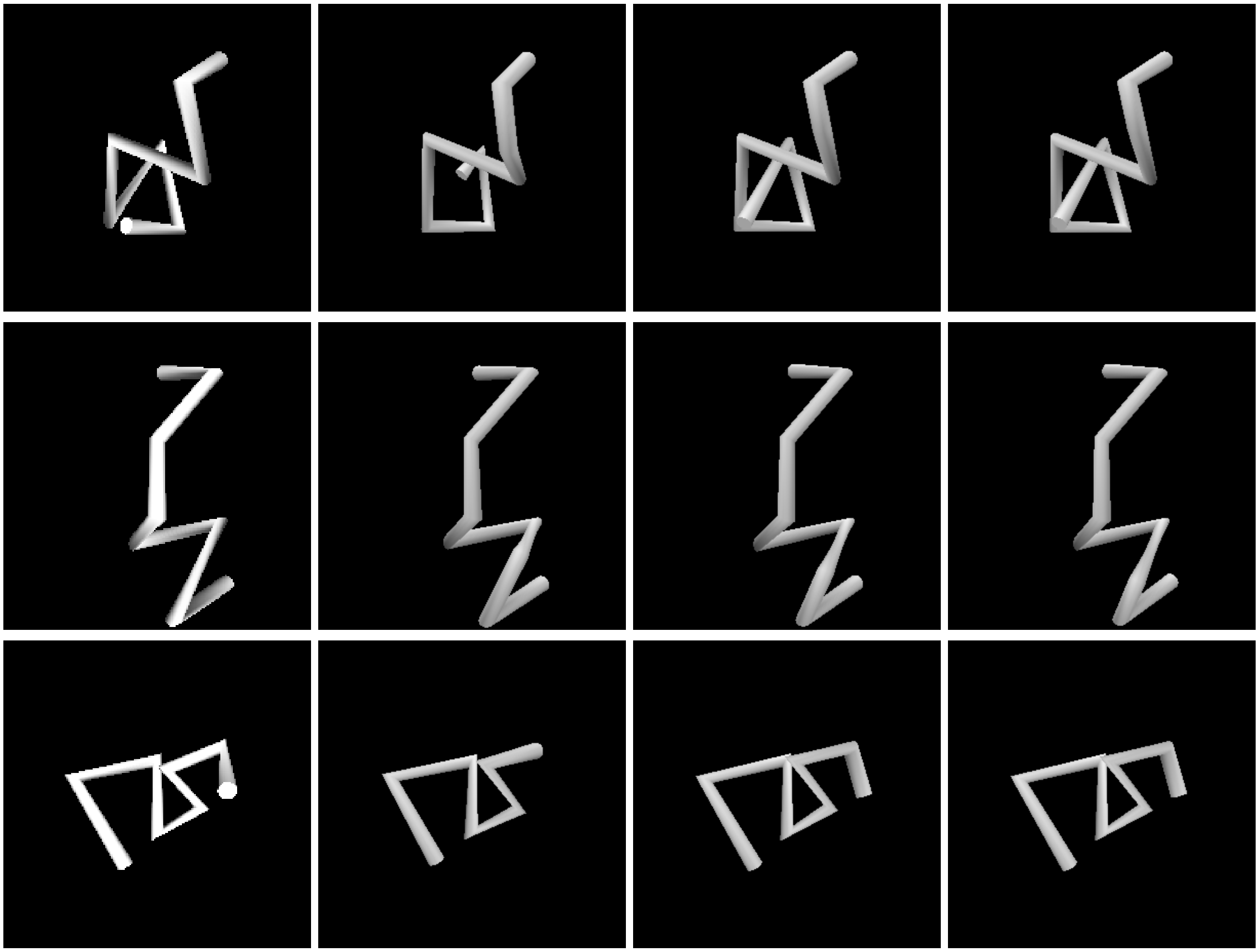


Figure 4: Examples of samples from the inferred posterior distribtuion $P(S|I_{\text{train}})$ for three objects. Each row depicts one object and three samples. The leftmost column shows the object from viewpoint $\theta = 0°$. Here, I_{train} consists of six views of an object from $\theta \in \{-90°, -75°, -60°, -15°, 0°, 15°\}$.

model was presented with test image I_{test}, and it judged whether the image depicted the same object as observed during training.

We implemented this decision process as a comparison between two probabilities: (i) the probability that the test image depicted the same object as depicted in the training images, versus (ii) the probability that the test image depicted any other object. These probabilities were formalized as $P(I_{\text{test}}|I = I_{\text{train}})$ and $P(I_{\text{test}}|I \neq I_{\text{train}})$, respectively. We estimated $P(I_{\text{test}}|I = I_{\text{train}})$ as follows:

$$\begin{aligned} P(I_{\text{test}}|I_{\text{train}}) &= \int P(I_{\text{test}}|S)\; P(S|I_{\text{train}})\; dS \\ &\approx \frac{1}{N}\sum_{i=1}^{N} P(I_{\text{test}}|S_i) \end{aligned} \tag{4}$$

where S_i is a sample from the posterior $P(S|I = I_{\text{train}})$.[3] Because an object can be depicted from any viewpoint on a test trial, viewpoint needs to be taken into account when calculating $P(I_{\text{test}}|S)$. In our simulations, we found the viewpoint that best aligned the object with the observed image (i.e., we used $P(I_{\text{test}}|S) = \max_{\vec{\phi}} P(I_{\text{test}}|S, \vec{\phi})$). To find the best viewpoint, we carried out a search over the whole viewing sphere (θ and α were each discretized into 5° bins).

We calculated $P(I_{\text{test}}|I \neq I_{\text{train}})$ in a similar manner:

$$P(I_{\text{test}}|I \neq I_{\text{train}}) = \int P(I_{\text{test}}|S)\; P(S|I \neq I_{\text{train}})\; dS. \tag{5}$$

To approximate this integral, samples from the posterior $P(S|I \neq I_{\text{train}})$ are needed. Because it is unlikely that any shape except the true shape was depicted in the training images, $P(S|I \neq I_{\text{train}})$ is close to the prior $P(S)$. Using this approximation, $P(S|I \neq I_{\text{train}})$ can be

[3] We sampled from the posterior $P((S, \vec{\phi})|I = I_{\text{train}})$ but we ignored viewpoint $\vec{\phi}$ and treated S as a sample from $P(S|I = I_{\text{train}})$. This is equivalent to approximating $P(S|I)$ with $P((S, \vec{\phi}_{\text{MAP}})|I)$. Since $P((S, \vec{\phi})|I)$ is highly peaked around the MAP sample, this is a very good approximation. Our results do not change if we integrate out $\vec{\phi}$ to get $P(S|I) = \int p(S, \vec{\phi}|I) d\vec{\phi}$ instead of using the approximation.

estimated as follows:

$$\begin{aligned} P(I_{\text{test}}|I \neq I_{\text{train}}) &\approx \int P(I_{\text{test}}|S)\ P(S)\ dS \\ &\approx \frac{1}{M}\sum_{i=1}^{M} P(I_{\text{test}}|S_i) \end{aligned} \tag{6}$$

where S_i is a sample from prior $P(S)$.[4]

Bülthoff and Edelman (1992) reported error rates in their experiment. As shown in Figure 5a, subjects' performances were excellent in the *interpolation* condition, but these rates were significantly higher in the *extrapolation* and *orthogonal* conditions. Importantly, performances in the *interpolation* condition were relatively unaffected by viewpoint. However, error rates rose with the difference in viewpoint between training and test in the other conditions. Performance was worst in the *orthogonal* condition. At first, this might seem to be due to the fact that subjects observed two sets of views varying along the horizontal axis during training, hence receiving more information about side views of objects. However, Bülthoff and Edelman (1992) ran a variant of their experiment where the training views varied along the vertical axis, and subjects still performed worse for test views varying along this axis. This finding suggests that people find it harder to generalize to top/bottom views than to side views. To account for this finding, Bülthoff and Edelman (1992) restricted their model's generalization capability along the vertical axis to be significantly less than what it is along the horizontal axis.

To compare our model's performances with those of the subjects in Bülthoff and Edelman (1992)'s experiment, we need to calculate an error measure for our model. Because an observer is expected to make more errors as the observer becomes less confident about whether a test image depicts the training object, we used the posterior ratio $\frac{P(I_{\text{test}}|I\neq I_{\text{train}})}{P(I_{\text{test}}|I=I_{\text{train}})}$ as

[4] In our simulations, we used an additional approximation based on the fact that for a random shape S_i, $P(I_{\text{test}}|S_i)$ is nearly proportional to $\exp\left(-||I_{\text{test}}||_F^2/2\sigma^2\right)$ since there will be little overlap between the image of a random shape and a test image (i.e., $P(I_{\text{test}}|S_i)$ is nearly independent of S_i). Simulations confirm that this approximation is in general quite good—the results are virtually the same as when we approximate $P(I_{\text{test}}|I \neq I_{\text{train}})$ with samples from $P(S)$.

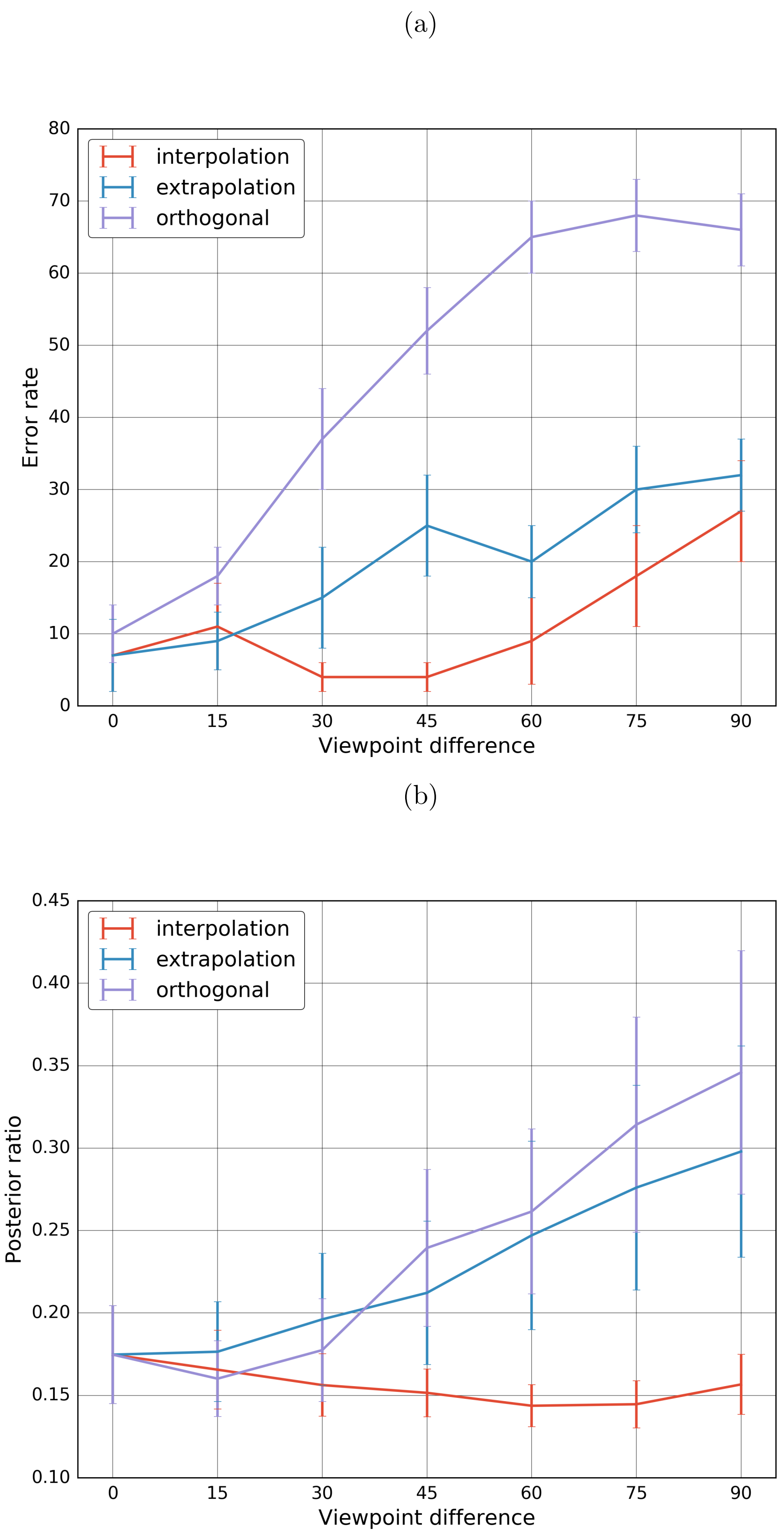
an error measure. For each test image in the three experimental conditions, this error measure was calculated. The results are summarized in Figure 5b. Overall, our model provides a good qualitative account of the experimental data. Its performance is best in the *interpolation* condition and markedly worse in the *extrapolation* and *orthogonal* conditions.[5] Moreover, its performance in the *interpolation* condition was relatively unaffected by viewpoint. However, its error measure rose with the difference in viewpoint between training and test in the other conditions.

Overall, these results show that viewpoint-dependency does not imply that an observer is using 2D or viewpoint-dependent object representations. Our model, using probabilistic 3D object-centered representations, accounts for viewpoint-dependency of visual object recognition. Contrary to received wisdom in the field, viewpoint-dependency does not provide compelling evidence about whether object shape representations are 2D versus 3D, nor does it provide evidence about whether these representations are view-dependent or view-independent.

5 Behavioral Experiment and Model Comparisons

The previous section reported results indicating that it is erroneous to claim that viewpoint-dependent visual object recognition suggests the use of view-based shape representations. Indeed, either view-based or probabilistic 3D object-centered representations can underlie viewpoint-dependency, particularly when such a representation is inferred from an image. The goal of this section is to report results strengthening our hypothesis that people's shape representations of unfamiliar objects are probabilistic, 3D, and object-centered. We present a behavioral experiment, along with an extensive evaluation of a diverse array of computational

[5] Given that we used a uniform prior over viewpoint, the lack of difference in performances between the *extrapolation* and *orthogonal* conditions is unsurprising. We could have captured this difference by assuming a non-uniform prior over viewpoint (like Bülthoff and Edelman (1992) do in their view-approximation model). However, we chose not to do so because our primary aim here is not to capture this difference but to account for view-dependency (i.e., increases in error rate with increases in viewpoint difference between training and test views).


(a)

(b)

Figure 5: (a) Experimental results from Bülthoff and Edelman (1992). (b) Simulation results from our model.

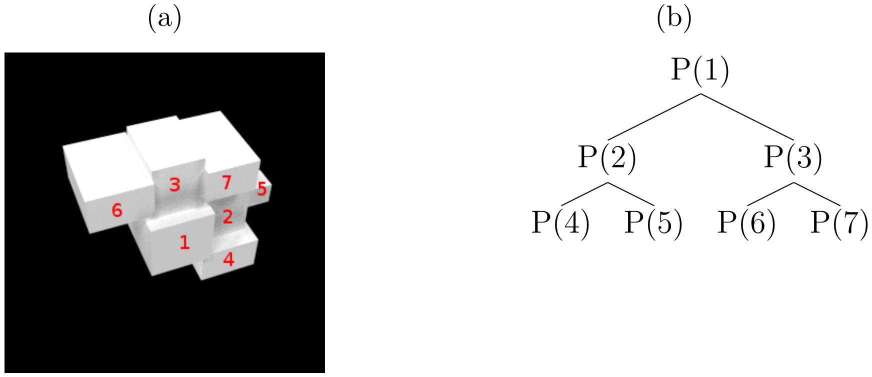


Figure 6: (a) Example of an object. The numbers on parts refer to the part numbers in its shape tree. (b) The shape tree representing the object in (a).

models based on how well the models account for the experimental data. We show that our probabilistic, 3D, object-centered inference model captures subjects' performances better than all other models.

5.1 Behavioral experiment: Stimuli and procedure

Experimental stimuli were objects built from rectangular blocks. They were generated as follows. Each object started with a single fixed-size block centered at the origin. Then, one or more faces of this root block were randomly selected, and one or more new blocks with randomly sampled sizes were connected to the selected faces. This procedure was applied recursively—after child blocks were connected to a parent block, each child became a parent and had one or more child blocks connected to it. In practice, a parent block was restricted to have at most three child blocks. We also restricted the depth of each object to three (i.e., an object consisted of its root block, the root block's child blocks, and the root block's grandchild blocks). A sample object and its corresponding shape tree representation are shown in Figure 6.

We generated 10 target objects in this manner. Comparison objects with shapes similar, but not identical, to target objects were also created. They were generated by applying the following four manipulations to each target object. Each manipulation was applied at levels two and three in the shape trees, resulting in 8 comparison objects generated from

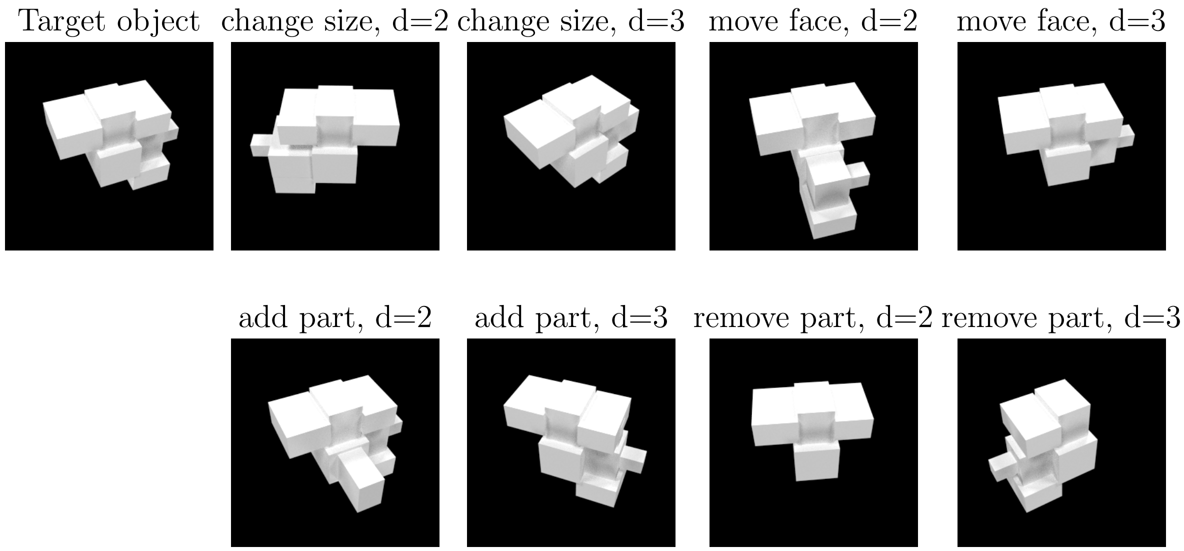


Figure 7: Target object (upper left) and its 8 comparison objects. The comparison objects were created using the four manipulations applied at levels two and three of the target object's shape tree. For example, "add part, d=2" refers to the object created by adding a new part to depth 2 in the shape tree.

each target object. When using the *change part size* manipulation, one object part was randomly selected, and its size was set to a random value. This operation might change the positions of the selected part's descendants. When using the *change connecting face of part* manipulation, we again picked one part randomly, picked a new connecting face for it from the unoccupied faces of its parent part, and moved the part to this new location. Again, this manipulation moves all descendants of the selected part. The *add part* manipulation added one part randomly to the desired level in the tree. For example, to add a new part to level 3, we picked one of the parts at level 2 randomly, picked one of its unoccupied faces randomly, and connected a new part with a random size to the chosen face. When using the *remove part* manipulation, we picked one part randomly and removed it and all of its descendants. Figure 7 illustrates a target object and examples of its 8 comparison objects.[6]

The experiment used a shape similarity judgment task. On each trial, a subject viewed images of three objects, one target and two comparisons. Subjects judged which comparison

[6] The full set of stimuli can be seen online at `http://gokererdogan.github.io/ShapePerceptionAsBayesianInference/`.

was most similar in shape to the target. Images were rendered from a random viewpoint on the 45° parallel ($\alpha = 45°$) along the viewing sphere using Blender (`http://www.blender.org`), a 3D graphics and animation software package. Each subject performed 100 trials, 16 of which were catch trials where one of the comparison objects was identical to the target. Forty-one subjects participated in the experiment, but data from five subjects were discarded because they failed to achieve 85% accuracy on catch trials. Subjects participated in this web-based experiment via Amazon Mechanical Turk.

5.2 Competing computational models

A diverse array of models of shape perception was simulated, and each model was evaluated based on how well it accounts for subjects' responses in our experiment. We include the following models in our evaluation.

Pixel-based model: The pixel-based model compares two objects by calculating the Euclidean distance between the pixel values in their images. This model can be regarded as an implementation of a view-based hypothesis that stores images as views. Subjects in our experiment saw each object only from a single viewpoint, and thus the shape representation for an object in the pixel-based model consists of a single image. Because there is only one image stored for each object, the pixel-based model is also an implementation of a particular version of Poggio and Edelman (1990)'s view-approximation model that works directly on raw images.

Alignment-based models: Another set of models that use 2D representations is motivated by the recognition-by-alignment approach (Ullman, 1989). Here, images of objects are aligned before they are compared. This alignment process requires a set of image features to be labeled in the images. The best alignment is calculated on the basis of these features. The dissimilarity between two images is taken to be the Euclidean distance in pixel space between the images after alignment. In our simulations, we used the corners of the root block as features for alignment since these corners are present in every image. One can imagine allowing various types of transformations in the alignment process. Here we tried two transformations: one allowing only scaling and translation, and another allowing any affine transformation.

We also tried a third method that does not do any alignment. Instead, this no-alignment model simply calculates the Euclidean distances between feature lists (i.e., the coordinates of corners of root blocks). The model is an implementation of a view-based hypothesis that uses a simple feature-based representation for views. For this reason, the model is also referred to as a naive feature-based model. Since Poggio and Edelman (1990)'s original view approximation model worked on similar feature-based representations, the no-alignment model also provides a test of the view approximation model.

HMAX: An influential example of a feature-based model is HMAX (Riesenhuber & Poggio, 1999).[7] The model is a type of artificial neural network consisting of four layers of units: S1, C1, S2, C2. We used the outputs from the C1 and C2 layers (as is generally done in previous work evaluating HMAX models). The particular implementations we used applied feature extraction at the C1 layer at eight different spatial scales. We treated each scale as a separate model, and also combined all eight scales into a single C1 layer representation. Our HMAX implementation also used eight different patch (i.e., feature) sizes at the C2 layer. Again, we treated activations for each of these patch sizes as a separate model, but also combined all of these models to form a single C2 layer representation. Therefore, in total, there are 18 versions of HMAX (8 scales for C1, all scales combined, 8 patch sizes for C2, and all patch sizes combined). We used the feature dictionary provided with the HMAX implementation. These features were extracted from random natural images, and are intended as a universal set of features. To get feature-based representations for each object in our experiment, we fed each image of an object to an HMAX model and calculated the responses of the C1 and C2 layers. These responses constitute objects' shape representations. We used Euclidean distance to compute dissimilarities between two such

[7] We used the implementation provided by the authors at `http://maxlab.neuro.georgetown.edu/docs/hmax/hmaxMatlab.tar`

shape representations.

Convolutional neural networks: We evaluated two convolutional neural networks (CNNs) that are regarded as state-of-the-art computer vision systems: AlexNet (Krizhevsky et al., 2012) and GoogLeNet (Szegedy et al., 2014).[8] AlexNet is an eight-layer (five convolutional, three fully connected layers) CNN trained on 1.2 million images in the ImageNet dataset. AlexNet achieved the best performance on the 2012 ImageNet Large Scale Visual Recognition Challenge, and was in large part responsible for the recent surge of interest in deep neural networks. We treated each of its 14 layers (making the three max-pooling and two normalization layers explicit) as a separate model. Using the standard terminology in the deep neural network literature, these layers are: conv1, pool1, norm1, conv2, pool2, norm2, conv3, conv4, conv5, pool5, fc6, fc7, fc8, and prob. The set of unit activations in the last layer, prob, is a 1000-dimensional vector encoding the probability of belonging to each of 1000 object categories in ImageNet. The second CNN that we tested was GoogLeNet by Szegedy et al. (2014). This model set the state-of-the-art performance on the 2014 ImageNet Large Scale Visual Recognition Challenge. GoogLeNet has 22 layers (with an additional five pooling layers). Our simulations used 16 layers: pool1, conv2, inception3a-b, pool3, inception4a-e, pool5, inception5a-b, pool5, loss3 and prob. To make predictions from AlexNet and GoogLeNet, we input each image to the networks and performed a feedforward pass to calculate each layer's responses. The dissimilarity between two objects is computed as the Euclidean distance between vectors of these responses.

Structural distance-based model: We implemented a structural distance-based model that calculates object similarity using the structural descriptions of objects. Unfortunately, there are no concrete proposals in the literature for how this should be done. Because the objects in our experiment can be represented as shape trees (see Figure 6), one plausible way is to use the distance between these trees as a measure of dissimilarity. We used one such measure referred to as tree-edit distance (Zhang & Shasha, 1989). Using this measure, the

[8] We use the pre-trained networks provided by the Caffe framework (Jia et al., 2014).

distance between two shape trees is the total cost of operations needed to turn one tree into the other.[9] Tree-edit distance allows add-node, remove-node and change-node operations, and we assumed that each operation has equal cost.

Shape skeleton model: As discussed above, Feldman et al. (2013) proposed to represent the 2D shape of a 2D object as a shape skeleton. This skeleton is inferred from an image silhouette using Bayesian inference. To calculate similarities between shapes, we first extracted the boundaries of objects in images to create 2D silhouettes. Then we used Feldman et al. (2013)'s model[10] to find the maximum-a-posteriori (MAP) shape skeleton for each silhouette. The similarity between two shapes can be formalized as the probability of observing the image for one shape given the image for the other shape. For example, the similarity between the target I_t and a comparison I_c can be evaluated by calculating either $P(I_t|I_c)$ or $P(I_c|I_t)$. $P(I_t|I_c)$ (and similarly $P(I_c|I_t)$) can be approximated on the basis of an estimated MAP shape skeleton for each shape as follows:

$$P(I_t|I_c) \approx P(I_t|Sk_{\text{MAP}})P(Sk_{\text{MAP}}|I_c) \tag{7}$$

where Sk_{MAP} denotes the MAP skeleton for I_c. We tried three similarity measures based on these probabilities: $P(I_t|I_c)$, $P(I_c|I_t)$, and their average $\frac{1}{2}[P(I_t|I_c) + P(I_c|I_t)]$.

3D shape inference model: Lastly, we describe our proposed model that treats shape perception as Bayesian inference of 3D shape in an object-centered coordinate system. To specify our model, we need to describe the representation for object shape as well as the generative process or forward model mapping these representations to images. We assume that shape representations consist of the positions and sizes of a collection of rectangular blocks. Each object S is represented by a tuple (T, M) where T is a string from a probabilistic shape grammar with production rules: $P \to P \mid PP \mid PPP \mid \epsilon$. In these rules, P is a non-

[9] Tree-edit distance considers two nodes to be equal if their labels are the same. In the case of our shape trees, this means that two P nodes need to have the same connection face to be considered equal.

[10] We used the implementation provided by the authors at `http://ruccs.rutgers.edu/images/ShapeToolbox1.0.zip`

terminal symbol and ϵ is a terminal null symbol. In a string T generated by this grammar, each P symbol corresponds to an object part (i.e., a rectangular block). Hence, the string T characterizes the parent-child relations between parts in an object. The grammar follows closely our stimulus generation procedure, with each part being constrained to have at most three children. The sizes and positions of each part are specified in spatial model M. The spatial model associates a size $s \in \mathbf{R}^3$ and a connecting face of a block $f_i \in \{1, 2, 3, 4, 5, 6\}$ with each P node in T (see Figure 8 for an example object and its associated (T, M) shape representation).

The prior probability of shape S is:

$$P(S) = P(T)P(M|T). \tag{8}$$

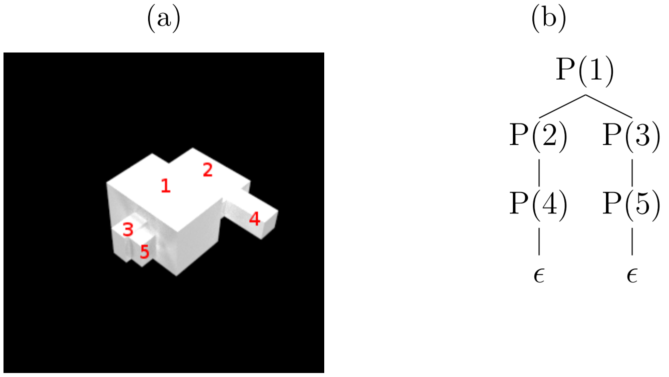
The probability of producing T from the shape grammar, $P(T)$, is calculated as follows:

$$P(T) = \prod_{n \in \mathcal{P}} P(n \to ch(n)) \tag{9}$$

where $\mathcal{P}$ is the set of P nodes in tree T, $ch(n)$ are the children of node n, and $p(n \to ch(n))$ is the probability of the production rule $n \to ch(n)$. We assume production probabilities to be uniform (i.e., each of the four production rules has a probability of 0.25) which simplifies $P(T)$ to:

$$P(T) = \frac{1}{4^{|\mathcal{P}|}}. \tag{10}$$

The probability for spatial model M, $P(M|T)$, consists of the probabilities of picking part sizes and connecting faces. Because we assumed part sizes to be uniform over the interval [0, 1], we only need to focus on the probabilities for connecting faces. For a part with k available faces and c children, there are $\binom{k}{c}$ possible combinations of face assignments to its children. Since we have six empty faces for the root P node and five empty faces for the remaining P nodes (because one face is occupied by the parent), the probability of spatial



(c)

Part Id	Size	Conn. face
P1	[0.75, 0.50, 0.50]	n/a
P2	[0.64, 0.18, 0.45]	3
P3	[0.14, 0.28, 0.23]	2
P4	[0.37, 0.19, 0.18]	1
P5	[0.13, 0.24, 0.15]	1

Figure 8: (a) An example object. The numbers on parts refer to the part numbers in its parse tree. (b) Parse tree T associated with the object in (a). (c) Spatial model M associated with the object in (a). "Conn. face" is shorthand for "connection face" (i.e., the parent's face to which a part is connected).

model M is

$$P(M|T) = \frac{1}{\binom{6}{|\mathcal{O}_{\text{root}}|} \prod_{n \in \{\mathcal{P} \setminus \text{root}\}} \binom{5}{(|\mathcal{O}_n|-1)}} \tag{11}$$

where $\mathcal{O}_i$ refers to the set of occupied faces of node i.

Given a shape S and a viewpoint $\vec{\phi}$, forward model $\mathcal{F} : (S, \vec{\phi}) \to I$ maps 3D shape representations to 2D images. As above, we used the Visualization Toolkit software package to implement the forward model. Assuming Gaussian noise on images, the likelihood function $\mathcal{L}(H, \theta; I)$ is:

$$\mathcal{L}(S, \vec{\phi}; I) = P(I|S, \vec{\phi}) \propto \exp\left(\frac{1}{\sigma^2}||I - \mathcal{F}(S, \vec{\phi})||_F^2\right) \tag{12}$$

where σ^2 denotes the variance of the noise on I and $|| \cdot ||_F$ is the Frobenius norm.

The posterior distribution over shapes given an image can be calculated via Bayes' rule:

$$P(S, \vec{\phi}|I) \propto P(I|S, \vec{\phi}) P(S) P(\vec{\phi}). \tag{13}$$

We assumed that $P(\vec{\phi})$ is a uniform distribution, and that viewpoint $\vec{\phi}$ is independent of shape S. We sampled from this posterior using MCMC techniques (see Appendix B for details). Figure 9 shows samples from the posterior over shapes for various objects in our experiment.

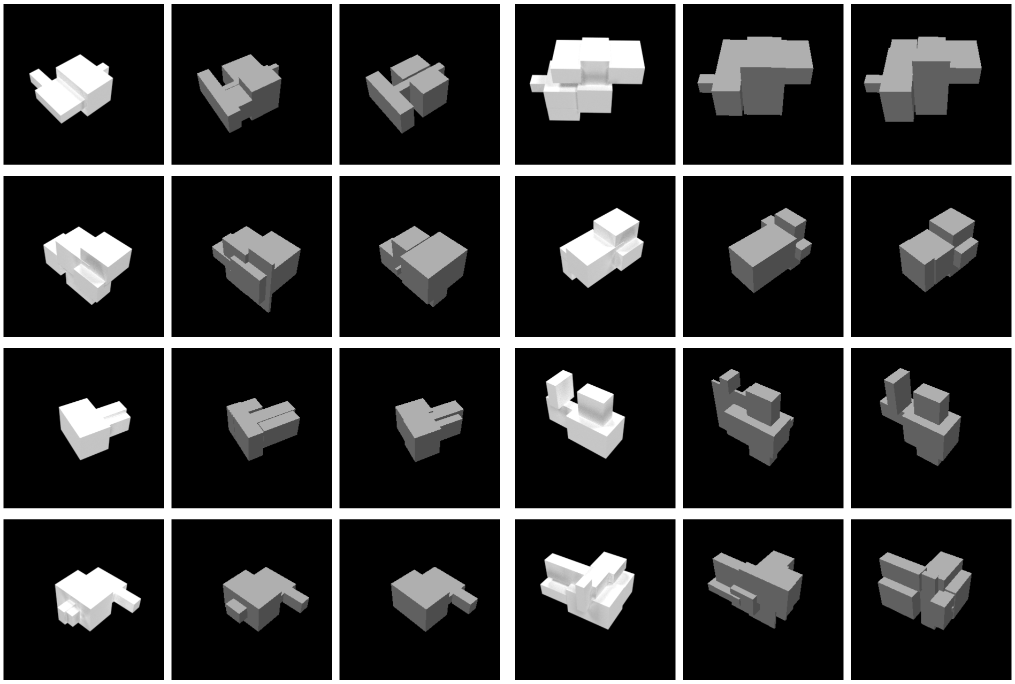


Figure 9: Samples from the posterior over shapes for various objects in our experiment. Each row contains two sets of one object followed by two samples.

To calculate the similarity between target and comparison objects, we evaluated how likely it is to observe the image for one object given the image of the other object. Denoting the images for target and comparison by I_t and I_c, respectively, we calculated three similarity measures: $P(I_\mathrm{t}|I_\mathrm{c})$, $P(I_\mathrm{c}|I_\mathrm{t})$, and their average. We calculated $P(I_\mathrm{c}|I_\mathrm{t})$ as follows (and similarly for $P(I_\mathrm{t}|I_\mathrm{c})$):

$$P(I_\mathrm{c}|I_\mathrm{t}) = \int P(I_\mathrm{c}|S,\vec{\phi})P(S|I_\mathrm{t})P(\vec{\phi})dSd\vec{\phi}. \tag{14}$$

In a similar vein to Equation 4, the value of this integral was approximated using samples from $P(S|I_\mathrm{t})$.

5.3 Simulation results

For each computational model described above, we calculated its predictions as follows. For each simulated trial, we computed the similarities between a target object and each comparison object, and used the most similar comparison as a model's prediction. We evaluated the performance of each model by calculating the percentage of trials in which a model and our experimental subjects made the same judgment (i.e., they picked the same comparison object as most similar to the target). The results are shown in Figure 10.

Clearly, our proposed computational model significantly outperformed all other models (particularly the version whose similarity measure averaged $p(I_\mathrm{t}|I_\mathrm{c})$ and $p(I_\mathrm{t}|I_\mathrm{c})$; binomial test, $p < 0.005$ for all comparisons). The pixel-based (i.e., view-based) model performed at 58%. Even though this is significantly better than chance, it still lags far behind our model's performance of 72%. Similarly, the best alignment-based model only reached an accuracy of 59%.[11] The structural distance-based model lagged even the pixel-based model at 54% accuracy, which is not significantly better than chance. Similarly, the best version of the shape skeleton model performed worse than the pixel-based model with 56% accuracy (with similarity measure based on $P(I_c|I_t)$). However, this performance is significantly better than chance ($p = 0.035$). The best version of HMAX also performed worse than the pixel-based model and naive feature-based model (i.e., no alignment model) with an accuracy of 57% (with layer C1, s=5). Convolutional neural networks (CNNs) performed slightly better than pixel-based and alignment-based models. The best version of AlexNet reached an accuracy of 62% using its output layer prob, and the best version of GoogLeNet achieved 64% using layer inception5a. However, neither of these accuracies are significantly better than the pixel-based model's performance (binomial test, $p > 0.05$).

We also looked at the performance of each model on trials with high between-subject agreement. Even though average agreement between subjects was high (75%), it might be unfair to expect models to predict subjects' judgments on trials when subjects did not clearly prefer either comparison object significantly more than the other. The following

[11] Interestingly, allowing only translation and scaling transformations led to better performance than allowing any affine transformation. This might seem implausible because translation and scaling transformations are special cases of affine transformations. However, the alignment-based method simply finds the transformation that aligns two images as well as possible. This is not necessarily the alignment that makes the Euclidean distances between images reflect subjects' judgments.

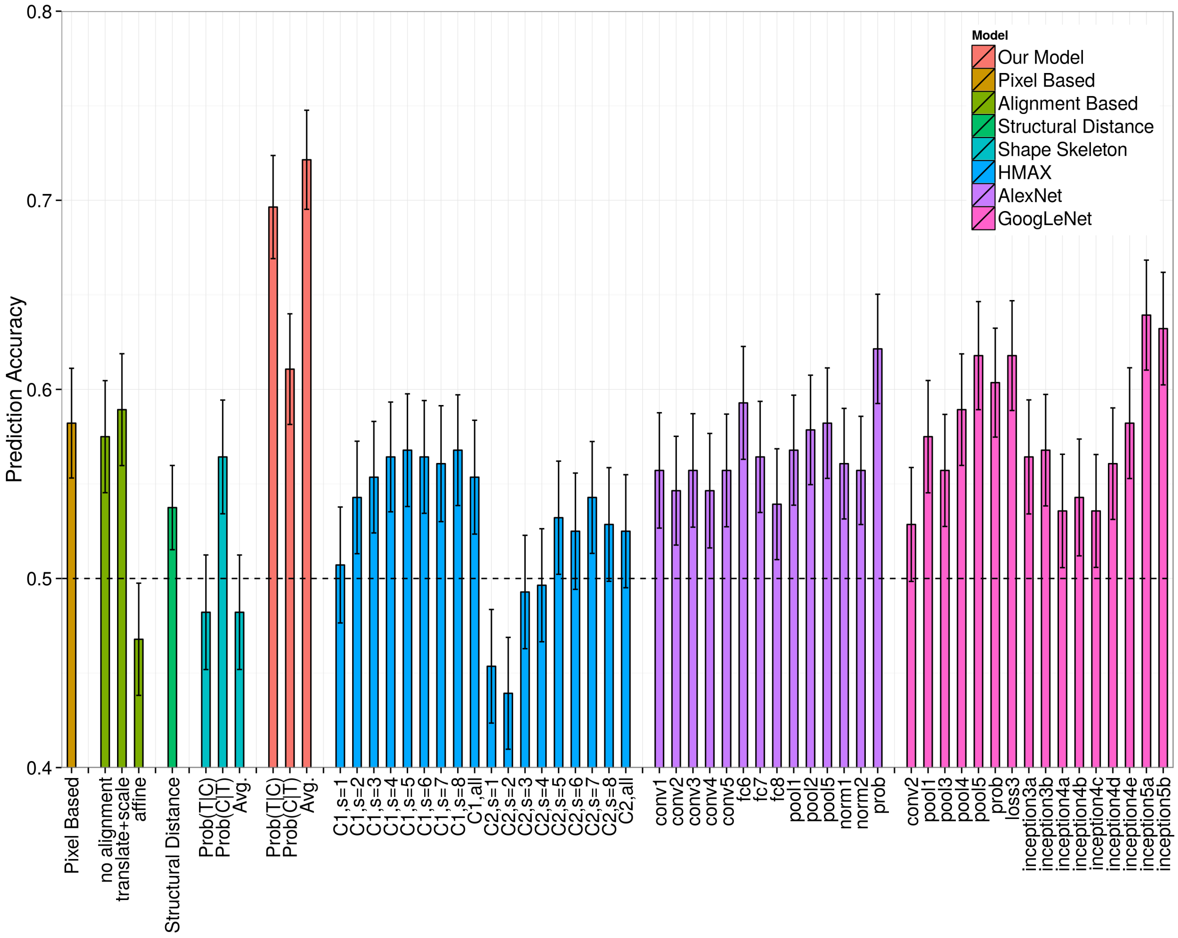


Figure 10: Predictions accuracies for each model on all trials. Error bars show SEMs estimated by a bootstrap procedure with 1000 replications. Note that the y-axis starts from 0.4.

analysis focuses on "high confidence" trials where at least 80% of subjects picked the same comparison object. Model accuracies on these high-confidence trials are shown in Figure 11. Our model significantly outperformed all other models with an accuracy of 87% ($p < 0.001$ for all comparisons). Pixel-based and alignment-based models achieved accuracies of 62% and 64%, respectively. Both of these values are significantly better than chance ($p = 0.01$ for pixel-based; $p = 0.002$ for alignment-based). Similarly, the structural distance-based model and shape skeleton model achieved an accuracy equal to that of the pixel-based model at 62%. The best version of HMAX performed at 57% which is not significantly different from the performance of either the pixel-based or alignment-based model. The best version of AlexNet reached an accuracy of 73% (with layer prob) which is significantly better than both pixel-based and alignment-based models ($p = 0.005$ for comparison with pixel-based; $p = 0.017$ for comparison with alignment-based). However, the best version of GoogLeNet reached an accuracy of 68% (with layer inception5b) which is not significantly better than the performance of either pixel-based or alignment-based models ($p = 0.11$ for comparison with pixel-based; $p = 0.24$ for comparison with alignment-based).

In the evaluations presented so far, object similarity was computed using the Euclidean similarity metric for several models. What would happen, however, if these models used a more powerful metric such as the Mahalanobis similarity metric? Would their performances significantly improve? The Euclidean metric is a special case of the Mahalanobis metric. Let $\vec{r}_M(I_i)$ denote a vector coding model M's shape representation based on image I_i. The Mahalanobis metric for the similarity of shape representations based on images I_i and I_j is:

$$[\vec{r}_M(I_i) - \vec{r}_M(I_j)]^T \, \Sigma^{-1} \, [\vec{r}_M(I_i) - \vec{r}_M(I_j)] \tag{15}$$

where Σ is a covariance matrix. The Euclidean metric is obtained by setting Σ to the identity matrix.

In the next analysis, we re-evaluated those models that previously used a Euclidean

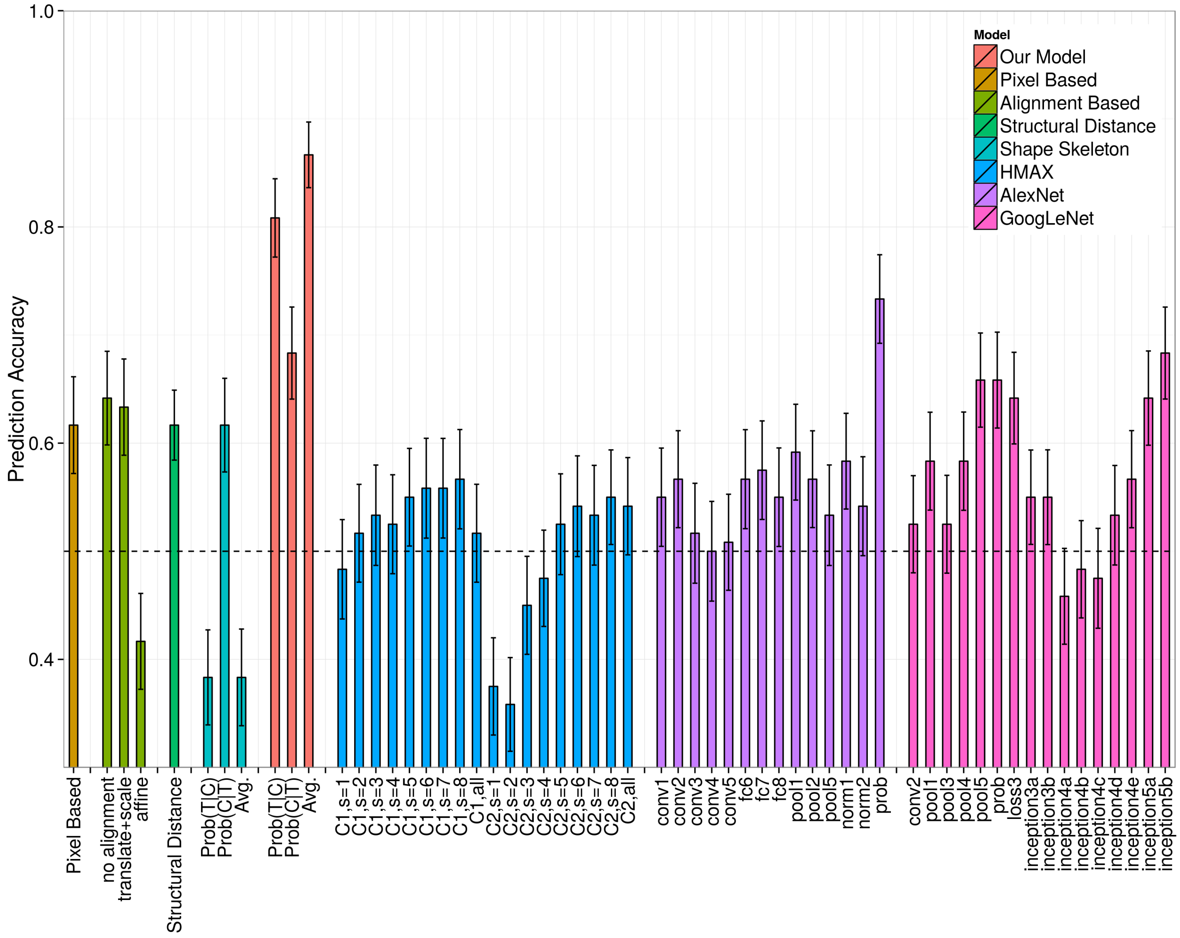


Figure 11: Predictions accuracies for each model on only high conidence trials. Error bars show SEMs estimated by a bootstrap procedure with 1000 replications. Note that the y-axis starts from 0.3.

metric by allowing the models to use a Mahalanobis metric. For each model, the covariance matrix Σ was obtained as follows. Subjects' judgments in our experiment can be thought of as relative similarity constraints. For example, if subjects picked object O_j to be more similar to O_i than O_k is, this can be characterized by a constraint of the form $s(O_i, O_j) > s(O_i, O_k)$, where s measures similarity between two objects. Using these constraints, it is possible to learn a Mahalanobis metric (i.e., learn a covariance matrix Σ) that satisfies as many of these constraints as possible. This problem is known as "metric learning" in the literature on Machine Learning (Kulis, 2013) where it is treated as an optimization problem that can be solved by iterative methods (see Appendix C for details on how we solved this problem).

To evaluate each model, 70% of subjects' similarity judgments selected at random were placed in a training set and the remaining judgments formed a test set. Using the training set, a model learned a Mahalanobis metric, and then this model was evaluated using the test set. This procedure was repeated 50 times to get a performance estimate for each model. We tried both diagonal and low-rank Σ matrices with varying rank values and report the best results. Table 1 shows the performances on all trials for the pixel-based model, the no-alignement (naive feature-based) model, HMAX, AlexNet, and GoogLeNet. (Recall that metric learning cannot be applied to the shape skeleton models, to the structural distance-based models, and to our proposed model because these models do not use vectors to represent shapes.)

Metric learning seems to help only AlexNet and, to a lesser extent, HMAX. However, neither of these increases in performance are statistically significant ($p = 0.18$ and $p = 0.37$ respectively). Importantly, our proposed computational model still outperforms all other models significantly ($p = 0.03$ for comparison with AlexNet). If we focus on only high confidence trials (see Table 2), metric learning improves the performances of all models, albeit not significantly for any model except HMAX ($p > 0.05$ for all other comparisons). Again, our 3D shape inference model is still significantly better than all other models ($p = 0.003$ for comparison with AlexNet). These results show that—even if we fit the similarity metric used by competing models to subject data—our shape inference model still provides a better

Model	Metric type	Accuracy	Best accuracy w/o metric learning
Pixel-based	low rank, r=10	0.566	0.582
Naive feature-based	diagonal	0.568	0.575
HMAX (C2, s=3)	diagonal	0.595	0.568
AlexNet (prob)	low rank, r=20	0.660	0.621
GoogLeNet (inception5b)	low rank, r= 20	0.633	0.639

Table 1: Best metric learning prediction accuracies on all trials.

Model	Metric type	Accuracy	Best accuracy w/o metric learning
Pixel-based	low rank, r=10	0.698	0.616
Naive feature-based	diagonal	0.648	0.642
HMAX (C1, s=6)	diagonal	0.714	0.567
AlexNet (prob)	low rank, r=5	0.752	0.733
GoogLeNet (inception5b)	diagonal	0.715	0.683

Table 2: Best metric learning prediction accuracies on high-confidence trials.

account of subjects' judgments.

We believe that our results are significant in multiple respects. First, our results suggest that people's shape representations for unfamiliar objects code 3D, rather than 2D, shape properties. Models that use 2D representations (i.e., pixel-based, alignment-based, and shape skeleton models)[12] were far inferior to our 3D shape inference model. Even if we allowed these models to fit their similarity metrics to subjects' data, our model still significantly outperformed them. These results strongly suggest that people do not represent shape for unfamiliar stimuli using 2D representations.

Second, our results raise doubts as to the promise of feature-based models. Even though these models tended to perform better than other models, they were still significantly behind our 3D shape inference model. This result is especially interesting for CNNs, which have attracted interest in the cognitive science and neuroscience communities as good models of biological visual systems. Their poor performance at accounting for our experimental

[12] It is worth emphasizing here that we base our claim on the 3D nature of shape representations, *not* on a comparison between our model and deep neural networks because it is unclear whether deep neural network models of shape perception use 2D or 3D representations. We touched upon this difficulty of knowing how and why these models achieve what they achieve in our review of feature-based models in Section 2. In fact, one line of research (Patel et al., 2015) suggests that deep neural networks are implementing an approximate version of probabilistic inference in a hierarchical probabilistic rendering model, similar to our proposed approach.

data suggests that these models might be representing visual objects in a manner that is different from how people represent visual objects. Further evidence for this claim is provided by studies showing that CNNs are easily fooled by images that seem indistinguishable or unrecognizable to the human eye (Nguyen, Yosinski, & Clune, 2014; Szegedy, Zaremba, & Sutskever, 2013).

Third, the structural-description based model's poor performance suggests that it is not adequate to represent objects as lists of parts and the coarse spatial relations among parts. Subjects' similarity judgments in our experiment seem to be based on finer-scale information than encoded in these structural descriptions, including the probabilistic information inferred by our proposed model.

Finally, our results have implications for the view-based hypothesis. Here we tested several view-based models. Alignment-based models tested Ullman (1989)'s approach, and our pixel-based model and no-alignment models tested two versions of Poggio and Edelman (1990)'s influential view approximation model. Our results show that none of the view-based models can account for subjects' judgments, and strongly suggest that view-based models do not provide good models of human shape perception.

6 Discussion

In summary, we have pursued an approach to investigating shape perception based on the "visual perception as Bayesian inference" framework. We hypothesized that shape perception of unfamiliar objects is well characterized as statistical inference of 3D shape in an object-centered coordinate system. The article provided evidence for this hypothesis along two lines. It first showed that a shape inference model that uses probabilistic, 3D, object-centered shape representations can account for view-dependency. This is a surprising result because previous researchers have interpreted view-dependency as incompatible with 3D, object-centered representations. Based on this result, we argued that view-dependency is not diagnostic of whether shape representations are 2D versus 3D, nor is it diagnostic of whether these representations are view-based versus view-independent. In addition, the article reported the results of a behavioral experiment using a shape similarity task, and compared the predictions of a diverse array of computational models to the experimental data. We found that our proposed shape inference model captures subjects' behaviors better than competing models. In conjunction, our experimental and computational results illustrate the promise of our approach and suggest that people's shape representations of unfamiliar objects are probabilistic, 3D, and object-centered.

Research on the visual perception of object shape has a long history. However, in terms of understanding the representations and algorithms involved in shape perception, it often seems as if we have made little progress (Peissig & Tarr, 2007; Gauthier & Tarr, 2016). We believe this is largely due to a lack of rigorous and quantitative approaches addressing the whole shape perception process from images to behavior. For example, view-based hypotheses rarely made commitments on the representation of individual views, or structural description hypotheses never completely specified how structural descriptions can be extracted from 2D images or how such descriptions can be compared. Hence, it became difficult to test these hypotheses, since without a clear specification of the whole perception process, their predictions were subject to interpretation. We believe progress is possible only if we build rigorous computational models, and our study is significant because it presents one such rigorous model of shape perception. As argued by Gauthier and Tarr (2016), we need to move away from unproductive dichotomies such as view-dependent versus view-invariant representations towards understanding the nature of the representation and algorithms involved in shape perception, which ultimately will explain when and why view-invariant or view-dependent performance is obtained. Our rigorous and quantitative approach here enables us to do exactly that.

We believe our work here is also significant because it presents a conceptual framework for understanding shape perception in its totality, rather than one aspect of it such as view-dependency or behavior on some single task such as object recognition. For example, view-based models focused almost exclusively on view-dependency of object recognition. Similarly, popular feature-based models are all models of object recognition. However, there is much more to shape perception than view-dependency or mapping images of objects to labels. We believe our approach is significant because it addresses shape perception in its totality, not just one aspect of it. By treating shape perception as inference of 3D, object-centered representations, we can explain not only view-dependency but also capture perceived similarities between unfamiliar objects. This is possible because our framework presents a generative model of shape perception, capturing how causes in the world give rise to retinal stimulations. Such models are often contrasted with discriminative approaches (such as popular feature-based models like AlexNet and GoogLeNet) that are built for individual tasks (such as object recognition) and cannot be easily adapted to new tasks (Lake, Ullman, Tenenbaum, & Gershman, 2016).

Our work directly or indirectly addresses or raises a large number of questions about the representation of object shape. Here we address several of these questions.

Previous research in the psychology literature has focused on how people might represent object shape, but has largely ignored the question of how people might acquire these representations. Why does the hypothesis proposed here emphasize that shape perception is a form of statistical inference? We believe that focusing on visual representations without also focusing on the acquisition of these representations is misguided. For example, it led researchers to develop theories of shape perception based on complete and accurate 3D, object-centered shape representations despite the fact that the acquisition of such representations is perceptually (and computationally) implausible, especially from a small number of viewpoints. If one augments an emphasis on representation with an emphasis on inference, one quickly realizes that people's shape representations will rarely be complete and accurate. For example, when a person views an object from a single viewpoint, the person is likely to infer a relatively accurate representation of some portions of the object but an inaccurate representation of other portions (e.g., portions seen in the periphery, or portions that are partially or fully occluded). We claim that this shape-inference problem underlies view-dependency.

The proposed computational model uses a specific approach, namely one based on probabilistic shape grammars. Why adopt this approach? Our proposed model uses a probabilistic shape grammar for several reasons. First, a shape grammar characterizes knowledge of possible object parts and of how parts might be combined to form objects. Part-based shape representations have previously received considerable theoretical and empirical support in the psychology literature (Biederman, 1987; Hoffman & Richards, 1984; Marr & Nishihara, 1978; Saiki & Hummel, 1998; Yildirim & Jacobs, 2013). Second, we represent shape in a probabilistic manner because probabilistic approaches are robust in noisy and uncertain environments, and because probabilistic inference algorithms often show excellent performances (as evidenced by the tremendous progress in the fields of Machine Learning and Statistics over the past few decades). Third, we are reasonably optimistic that the proposed model (or, rather, appropriately extended versions of the model) will scale well to larger-scale settings. Although important challenges obviously remain (too many to be mentioned here), our optimism stems from the fact that probabilistic shape grammars (much more complex than the one reported here) are regularly used in the Computer Vision and Computer Graphics literatures to address large-scale problems (Amit & Trouvé, 2007; Bienenstock, Geman, & Potter, 1997; Fu, 1986; Grenander & Miller, 2007; Talton et al., 2012; Tu, Chen, Yuille, & Zhu, 2005; Zhu, Chen, & Yuille, 2007, 2009).

The proposed computational model seems restricted to part-based objects. Is this a significant shortcoming? Can this model be scaled up to handle natural objects? Our main focus in this study was to argue for probabilistic, 3D, and object-centered shape representations. We have chosen the particular part-based shape representations used in this work because these are both powerful enough to capture 3D geometry of the stimuli we used and simple enough to make inference computationally feasible. Our mental shape representations are no doubt much richer than the representations we used here. A comprehensive understanding of object shape perception will require future work on shape representations that are rich enough to represent natural objects.

It is notoriously hard to predict the future[13] but we are hopeful that our approach can be scaled up to deal with the full complexity of natural objects. 3D volumetric representations similar to ours are being scaled to larger and larger settings by computer vision researchers (Rezende et al., 2016; Qi et al., 2016; Wu, Zhang, Xue, Freeman, & Tenenbaum, 2016). Moreover, recent research in Machine Learning and Statistics is leading to exciting advances in efficient inference in generative models. For example, fast, discriminative models can be trained to speed up inference dramatically in generative models (Kingma & Welling, 2014; Kulkarni, Yildirim, Kohli, Freiwald, & Tenenbaum, 2014; Yildirim, Kulkarni, Freiwald, & Tenenbaum, 2015).

The proposed computational model makes use of a powerful "forward model" that maps shape representations and viewpoints to visual images. Is this realistic? We believe that it is. Our results show that people discount viewpoint to a large extent when judging similarities which suggests such a forward model is implemented by our visual systems. In other settings, this mapping is referred to as visual imagery. Visual imagery is a type of mental simulation which researchers are increasingly hypothesizing as playing an essential role in human perception and cognition (Battaglia, Hamrick, & Tenenbaum, 2013).

The hypothesis proposed here is restricted to unfamiliar objects. Why? There are at least two reasons for this choice. First, our focus on unfamiliar objects provides a setting where potential confounding factors are controlled. Given past experience with familiar objects and their possible semantic significance, it is difficult (perhaps impossible) to dissociate the representation of shape from other possible relevant factors such as object category, object function, and developmental and evolutionary significance. Indeed, previous research clearly shows that conceptual knowledge affects visual perception (Dixon, Bub, & Arguin,

[13] Minsky and Selfridge (1961) famously predicted that hill-climbing approaches will never scale beyond the simple neural networks of the time. The current ubiquitous use of the backpropagation algorithm for training deep neural networks illustrates how wrong well-intentioned predictions can be.

1997; Gauthier, James, Curby, & Tarr, 2003; Goldstone, Lippa, & Shiffrin, 2001; Wiseman, MacLeod, & Lootsteen, 1985). Second, and perhaps more important, we believe that it is unrealistic to expect that people's visual systems use a single shape representation for all objects. For example, given the significance of some familiar objects—such as faces—and the difficulty of the associated visual recognition problem, it seems likely that people have specialized mechanisms and representations for these highly significant and familiar objects.

The hypothesis proposed here does not take into account an observer's task or goal. Is this a significant shortcoming? Yes and no. Consistent with the "active vision" approach to the study of perception (Findlay & Gilchrist, 2003; Hayhoe & Ballard, 2005), we believe that visual perception is often task-based. At the same time, we also believe that people use multiple representations of object shape, including representations that are not strongly dependent on task. Among other sources, evidence for this claim comes from our own recent brain-imaging research showing that cortical region LOC forms similar (and part-based) object shape representations when people visually or haptically perceive an object's shape in the absence of a task (Erdogan, Chen, Garcea, Mahon, & Jacobs, 2016).

Object shape can be perceived visually but it can also be perceived haptically. What is the relationship between visually-based and haptically-based shape representations? We believe that behavioral and computational studies (Erdogan, Yildirim, & Jacobs, 2015; Yildirim & Jacobs, 2013) as well as brain imaging studies (Erdogan et al., 2016) suggest that people acquire and use modality-independent object shape representations. These representations underlie behavioral phenomenon, such as cross-modal transfer of shape knowledge (Lacey & Sathian, 2011; Newell, 2010; Wallraven, Bülthoff, Waterkamp, van Dam, & Gaißert, 2014), and seem to reside in neural region LOC as well as other regions (Amedi, Malach, Hendler, Peled, & Zohary, 2001; Erdogan et al., 2016; Grill-Spector, Kourtzi, & Kanwisher, 2001; James et al., 2002). Our own previous work has shown that a computational model related to the one proposed here can infer shape representations from visual information, from haptic information, or both, and can account for an array of experimental data on cross-modal transfer of shape knowledge (Erdogan et al., 2015; Yildirim & Jacobs, 2013).

Is the proposed computational model psychologically plausible? Is it neurally plausible? Cognitive scientists often make a distinction between rational models and process models. Rational models are models of optimal or normative behavior, characterizing the problems that need to be solved in order to generate the behavior as well as their optimal solutions. In contrast, process models are models of people's behaviors, characterizing the mental representations and operations that people use when generating their behavior. Because our model's inference algorithm is optimal according to Bayesian criteria, and because this algorithm is not psychologically plausible, the model should be regarded as a rational model, not as a process model. Nonetheless, we believe that there are benefits to regarding the model as a rational/process hybrid. Like rational models, our model is based on optimality considerations. However, like process models, it uses psychologically plausible representations and operations (e.g., grammars, forward models).

For readers solely interested in process models, we claim that our model is a good starting point. As pointed out by others (Griffiths, Vul, & Sanborn, 2012; Sanborn, Griffiths, & Navarro, 2010), the MCMC inference algorithm used by our model can be replaced by approximate inference algorithms (known as particle filter or sequential Monte Carlo algorithms) that are psychologically plausible. Doing so would lead to a so-called "rational process model", a type of model that is psychologically plausible and also possesses many of the advantages of rational models.

In regard to neural plausibility, an important trend in computational neuroscience is to interpret neural activity in terms of probabilistic representations and operations (Pouget, Beck, Ma, & Latham, 2013). We, therefore, regard our model as at least potentially neurally plausible.

What are some important areas for future studies? We have emphasized the need to augment an emphasis on visual representation with an emphasis on the idea that shape perception is a form of statistical inference. This perspective leads to at least two areas for future research. First, any statistical inference mechanism needs to contain inductive biases in order to be effective. Future research needs to study the biases that play a role when people infer shape. These biases might take the form of "generic view" assumptions (Freeman, 1996) or "simplicity" assumptions (Feldman, 2000; Feldman et al., 2013; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). Second, the fact that our shape representations are the product of an inference process means that these representations may be inaccurate or incomplete (highlighting an advantage of probabilistic representations which directly code uncertainty). Here, we showed that an important consequence of this fact is that our percepts are view-dependent. Future research will need to study other perceptual consequences of our visual inference mechanisms.

A Details of MCMC algorithm for viewpoint-dependency simulations

In this appendix, we present the details of our Markov chain Monte Carlo (MCMC) procedure for inferring posterior probability distributions over the shapes of paperclip stimuli used in our viewpoint dependency simulations.[14] To sample from the posterior distribution $P(S, \vec{\phi}|I)$ over shape representations given a 2D image, we use MCMC techniques (J. S. Liu, 2004). These techniques produce samples from a desired probability distribution by constructing a Markov chain whose stationary distribution is the distribution of interest. In our inference procedure, we use the Metropolis-Hastings (MH) algorithm, a popular algorithm for constructing such Markov chains (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970).

An MH algorithm proposes a new hypothesis H' based on the current hypothesis H at each iteration, and accepts or rejects the proposed hypothesis with some probability. This accept/reject probability, called the acceptance ratio, is designed in such a way as to ensure that the stationary distribution of the Markov chain is the distribution of interest. Denote the probability of proposing hypothesis H' given the current hypothesis H with $q(H'|H)$ and the distribution of interest with $\pi(H)$. The MH acceptance ratio is:

$$a(H \to H') = \min\left(1, \frac{\pi(H')q(H|H')}{\pi(H)q(H'|H)}\right). \tag{16}$$

In our case, the target distribution $\pi(H)$ is the posterior $P(S, \vec{\phi}|I)$, and we need to design a proposal function q to move efficiently in the space of hypotheses. We use a mixture proposal (Tierney, 1994; Brooks, 1998) that consists of multiple proposals where one proposal is picked randomly at each iteration. Below we discuss each proposal function and its associated acceptance ratio.

[14] Implementations of the inference procedure for both paperclip and block stimuli are available online at `https://github.com/gokererdogan/Infer3DShape/releases/tag/ro3Dpaper`

Add/remove endpoint proposal: Given a shape $S = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_{|S|}\}$ consisting of $|S|$ endpoints, the add/remove endpoint proposal adds or removes a single endpoint. We allow only the "free" endpoints (i.e., $\vec{p}_1$ or $\vec{p}_{|S|}$) to be removed, and a new endpoint can only attach to one of these free endpoints. We calculate the probability for this proposal by considering the probability of each step in the procedure for adding/removing an endpoint. The add/remove endpoint proposal first randomly picks whether an add or remove endpoint manipulation should be carried out. We set each manipulation to be equally likely (i.e., $P(\text{add}|H) = P(\text{remove}|H) = 0.5$).[15] For a remove endpoint manipulation, the next step is to pick the endpoint to remove. Since there are two free endpoints, one of these is picked at random. For an add endpoint manipulation, again we first need to pick the free endpoint. In addition, we need to pick the position (x', y', z') of the new endpoint. A random vector on the unit sphere is picked randomly and added to the picked free endpoint to determine the position of the new endpoint. The proposal probabilities for add and remove endpoint manipulations are:

$$q_{\text{add}}(H'|H) = P(\text{add}|H)\ P(\text{pick endpoint}|H, \text{add})\ P(x', y', z'|H, \text{add}, \text{pick endpoint}) \quad (17)$$

$$q_{\text{remove}}(H'|H) = P(\text{remove}|H)\ P(\text{pick endpoint}|H, \text{remove}). \quad (18)$$

However, we cannot simply plug these into the MH acceptance ratio formula because the add/remove endpoint proposal manipulations move between spaces with different numbers of dimensions—shapes with different numbers of endpoints live in spaces with different number of dimensions. Therefore, we use a variant of the MH algorithm called "reversible jump MCMC" that can move between such spaces (Green, 1995). To see how it is applied for

[15] For some shapes, it might not be possible to add or remove a endpoint. For example, we never allow shapes with no endpoints. Therefore, we cannot apply a remove endpoint manipulation to a shape with only a single segment. In such cases, add and remove manipulation probabilities need to be modified accordingly. Similar modifications may be required for other steps in the add/remove endpoint proposal as well. See the implementation of our model for details.

our add/remove endpoint proposal, assume that we have a shape $S = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_{|S|}\}$ that consists of $|S|$ endpoints, and we add a new endpoint to get the proposed hypothesis $S' = \{\vec{p'}_1, \vec{p'}_2, \ldots, \vec{p'}_{|S|}, \vec{p'}_{|S|+1}\}$. Reversible jump MCMC assumes that we have sampled random variable $\vec{u}$ to make the number of dimensions equal in both hypotheses. In our case, we sampled $\vec{u} = (x', y', z') \in \mathbf{R}^3$ and added it to shape S (i.e., $S = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_{|S|}, \vec{u}\}$). We define a function $h : \vec{p}_1, \vec{p}_2, \ldots, \vec{u} \rightarrow \vec{p'}_1, \vec{p'}_2, \ldots, \vec{p'}_{|S|+1}$ that maps shape S to shape S'. Then, the reversible jump acceptance ratio is:

$$a(S \rightarrow S') = \min\left(1, \frac{\pi(S')q(S|S')}{\pi(S)q(S'|S)} \left|\det\left(\frac{\partial(\vec{p'}_1, \vec{p'}_2, \ldots, \vec{p'}_{|S|+1})}{\partial(\vec{p}_1, \vec{p}_2, \ldots, \vec{u})}\right)\right|\right) \tag{19}$$

where the rightmost term in this equation is the absolute value of the determinant of the Jacobian of the mapping h. Since in our case h is the identity function, its Jacobian is 1. Therefore, the acceptance ratio for the add endpoint manipulation is:

$$a(H = (S, \vec{\phi}) \rightarrow H' = (S', \vec{\phi})) = \min\left(1, \frac{\pi(H')q_{\text{remove}}(H|H')}{\pi(H)q_{\text{add}}(H'|H)}\right) \tag{20}$$

where q_{add} and q_{remove} are given by Equations 17 and 18, respectively. The acceptance ratio for the remove endpoint manipulation from H' to H is the inverse of the above expression.

Move endpoint proposal: This proposal picks one endpoint randomly and moves it a random amount $\vec{m} \in \mathbf{R}^3$ sampled from a normal distribution $\mathcal{N}(0, \sigma^2\mathbf{I})$. Hence, the proposal probability $q(H'|H)$ is:

$$q(H'|H) \propto \exp\left(-\frac{\sum_{i=1}^{3} m_i^2}{2\sigma^2}\right). \tag{21}$$

Since this proposal is symmetric (i.e., $q(H'|H) = q(H|H')$), the MH acceptance ratio is:

$$a(H \rightarrow H') = \min\left(1, \frac{\pi(H')}{\pi(H)}\right). \tag{22}$$

Rotate viewpoint proposal: This proposal changes the viewpoint $\vec{\phi} = (r, \theta, \alpha)$ from which a shape is viewed. We sample two random angles from a von Mises distribution with mean zero and variance κ, and add these to the polar coordinates θ and α. Since this proposal is symmetric, the acceptance ratio is again given by Equation 22.

B Details of MCMC algorithm for shape similarity task

In this appendix, we present the details of our MCMC inference procedure for block stimuli used in our behavioral experiment. See Appendix A for a short discussion of the Metropolis-Hastings algorithm. Similar to our inference procedure for paperclip stimuli, we use a mixture proposal that consists of multiple proposals, one of which is picked randomly at each iteration. Below we provide the details for each proposal procedure.

Add/remove part proposal: Let $S = (T, M)$ denote a shape where T refers to the parse tree associated with the shape, and M is the spatial model that consists of one size vector $\vec{s}_i \in \mathbf{R^3}$ and connecting face $f_i \in \{1, 2, 3, 4, 5, 6\}$ for each P node in parse tree T. The add/remove part proposal first randomly picks whether an add or remove manipulation will be carried out. We assume each manipulation is equally likely. For a remove part manipulation, a P node is picked randomly from the set $\mathcal{R}$ of P nodes with no children, and this part is removed. For an add part manipulation, a P node is picked randomly from the set $\mathcal{A}$ of P nodes that have fewer than three child P nodes. Then, a new child P node is added to the picked P node. This requires randomly sampling a size $\vec{s}$ for the new part and a connecting face f from the unoccupied connecting faces of its parent. The proposal probabilities for add and remove manipulations are:

$$
\begin{aligned}
q_{\text{add}}(H'|H) &= P(\text{add}|H)\ P(\text{pick part}|H,\text{add})\ P(\vec{s})\ P(f|H,\text{add},\text{pick part}) \\
&= \frac{1}{2}\frac{1}{|\mathcal{A}|}\frac{1}{(6-|O_P|)}
\end{aligned}
\tag{23}
$$

$$
\begin{aligned}
q_{\text{remove}}(H'|H) &= P(\text{remove}|H)\ P(\text{pick part}|H,\text{remove}) \\
&= \frac{1}{2}\frac{1}{|\mathcal{R}|}
\end{aligned}
\tag{24}
$$

where we assume $P(\vec{s})$ is uniform and use O_P to denote the set of occupied faces of the picked parent P part for add part manipulation.[16] Similar to the add/remove endpoint proposal for paperclip stimuli discussed above, we cannot simply plug these proposal probabilities into the MH acceptance ratio because hypotheses H and H' reside in spaces with different numbers of dimensions. Therefore, we use the reversible jump MCMC algorithm. A derivation similar to the one discussed for the add/remove endpoint proposal shows that the acceptance ratio for the add part manipulation is:

$$
a(H = (T, M, \vec{\phi}) \to H' = (T', M', \vec{\phi})) = \min\left(1, \frac{\pi(H')\ q_{\text{remove}}(H|H')}{\pi(H)\ q_{\text{add}}(H'|H)}\right) \tag{25}
$$

where q_{add} and q_{remove} are given by Equations 23 and 24, respectively. The acceptance ratio for the remove part manipulation from H' to H is the inverse of the above expression.

Change part size proposal: This proposal picks one P node randomly from shape $S = (T, M)$ and resamples its size $\vec{s}$ from a uniform distribution over $[0,1] \times [0,1] \times [0,1]$. Since this proposal is symmetric, the MH acceptance ratio is given by Equation 22.

[16] In some cases, it might not be possible to add or remove parts for a shape S. The proposal probabilities need to be modified accordingly in such cases.

Change connecting face of part proposal: This proposal picks one P node randomly from the set of P nodes whose parent P node has at least one empty face. A new connecting face is picked randomly from the set of empty faces of its parent, and the P node is connected to this new face. Again, because this proposal is symmetric, the MH acceptance ratio is given by Equation 22.

Rotate viewpoint proposal: This proposal changes the viewpoint $\vec{\phi} = (r, \theta, \alpha)$ from which a shape is viewed. In contrast to the proposal we used for paperclip stimuli, here we allow rotations only around the vertical direction. We sample a random angle from a von Mises distribution with mean zero and variance κ and add this to the polar coordinate θ. Since this proposal is symmetric, the acceptance ratio is given by Equation 22.

C Details of metric learning

In our evaluation of shape perception models, we use metric learning to fit the representations learned by models to behavioral data. Metric learning (Kulis, 2013) aims to learn a linear transformation of input data such that the distances between data points in the transformed space capture similarity/dissimilarity relations as well as possible. More formally, denote the representation for stimuli i with $\vec{r}_i$, and the distance between stimuli i and j with $d(\vec{r}_i, \vec{r}_j)$. Assume that we are given a set of relative similarity constraints of the form $d(\vec{r}_i, \vec{r}_j) < d(\vec{r}_i, \vec{r}_k)$. Our aim is to learn a linear mapping A such that the distances $d_A(\vec{r}_i, \vec{r}_j), d_A(\vec{r}_i, \vec{r}_k)$, etc., in this new space will satisfy as many of these relative similarity constraints as possible. Here $d_A(r_i, r_j)$ is the Mahalonobis distance between $\vec{r}_i$ and $\vec{r}_j$ which is given by $(\vec{r}_i - \vec{r}_j)^T A(\vec{r}_i - \vec{r}_j)$. Because there might not be a linear mapping A satisfying all constraints, we introduce slack variables ξ_{ijk} to express the metric learning problem as the following optimization problem (Schultz & Joachims, 2003):

$$\begin{aligned} \min_{A,\{\xi_{ijk}\}} \quad & \frac{1}{2}||A||_F^2 + C \sum_{ijk \in R} \xi_{ijk} \qquad (26) \\ \text{s.t.} \quad & d_A(\vec{r}_i, \vec{r}_k) - d_A(\vec{r}_i, \vec{r}_j) \leq 1 - \xi_{ijk} \\ & \xi_{ijk} \geq 0 \end{aligned}$$

where $|| \cdot ||_F$ denotes the Frobenius norm and C is a cost parameter controlling how much we care about violations of the relative similarity constraints. We consider two variants of this problem. In the first variant, we constrain A to be a diagonal matrix. In that case, this problem becomes equivalent to the one treated in Schultz and Joachims (2003). We find the optimal diagonal A by solving the dual of the optimization problem using the L-BFGS-B algorithm (Byrd, Lu, Nocedal, & Zhu, 1995) provided in the "scipy" open-source package of scientific tools (Jones, Oliphant, Peterson, & others, 2001). The second variant constrains A to be a low rank matrix. This can be achieved by writing A as $G^T G$ where G has fewer rows than columns. To solve this problem, we rewrite it in the following unconstrained form:

$$\min_{G} \quad \frac{1}{2}||G^T G||_F^2 + C \sum_{ijk \in R} \max(0, d_A(\vec{r}_i, \vec{r}_j) - d_A(\vec{r}_i, \vec{r}_k) + 1) \qquad (27)$$

and again use L-BFGS-B to find the optimal A matrix. Implementations of these metric learning methods are provided online at `https://github.com/gokererdogan/gmllib`.

References

Amedi, A., Malach, R., Hendler, T., Peled, S., & Zohary, E. (2001). Visuo-haptic Object-Related Activation in the Ventral Visual Pathway. *Nature Neuroscience*, *4*(3), 324–330.

Amit, Y., & Trouvé, A. (2007). POP: Patchwork of Parts Models for Object Recognition. *International Journal of Computer Vision*, *75*(2), 267–282.

Anselmi, F., Rosasco, L., Tan, C., & Poggio, T. (2015). Deep Convolutional Networks are Hierarchical Kernel Machines. *arXiv:1508.01084*.

Bar, M. (2001). Viewpoint Dependency in Visual Object Recognition Does Not Necessarily Imply Viewer-Centered Representation. *Journal of Cognitive Neuroscience*, *13*(6), 793–799.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an Engine of Physical Scene Understanding. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(45), 18327–32.

Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, *94*(2), 115–147.

Biederman, I. (2007). Recent Psychophysical and Neural Research in Shape Recognition. In N. Osaka, I. Rentschler, & I. Biederman (Eds.), *Object Recognition, Attention, and Action.*

Biederman, I., & Gerhardstein, P. (1993). Recognizing Depth-Rotated Objects: Evidence and Conditions for Three-Dimensional Viewpoint In variance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(6), 1162–1182.

Biederman, I., & Gerhardstein, P. C. (1995). Viewpoint-dependent Mechanisms in Visual Object Recognition: Reply to Tarr and Bülthoff (1995). *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1506–1514.

Bienenstock, E., Geman, S., & Potter, D. (1997). Compositionality, MDL Priors, and Object Recognition. In *Advances in Neural Information Processing Systems* (pp. 838–844).

MIT Press.

Binford, T. O. (1971). Visual Perception by Computer. In *Proceedings of IEEE Conference on Systems and Control.* Miami.

Blum, H., & Nagel, R. N. (1978). Shape Description Using Weighted Symmetric Axis Features. *Pattern Recognition*, *10*(3), 167–180.

Brooks, S. (1998). Markov Chain Monte Carlo Method and its Application. *Journal of the Royal Statistical Society: Series D*, *47*(1), 69–100.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical Support for a Two-dimensional View Interpolation Theory of Object Recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(1), 60–4.

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How Are Three-Dimensional Objects Represented in the Brain? *Cerebral Cortex*, *5*(3), 247–260.

Byrd, R., Lu, P., Nocedal, J., & Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Comput Biol*, *10*(12), e1003963.

Cadieu, C. F., Kouh, M., Pasupathy, A., Connor, C. E., Riesenhuber, M., & Poggio, T. (2007). A Model of V4 Shape Selectivity and Invariance. *Journal of Neurophysiology*, *98*(3), 1733–1750.

Dixon, M., Bub, D. N., & Arguin, M. (1997). The Interaction of Object Form and Object Meaning in the Identification Performance of a Patient with Category-specific Visual Agnosia. *Cognitive Neuropsychology*, *14*(8), 1085–1130.

Edelman, S. (1997). Computational Theories of Object Recognition. *Trends in Cognitive Sciences*, *1*(8), 296–304.

Edelman, S., & Bülthoff, H. H. (1992). Orientation Dependence in the Recognition of

Familiar and Novel Views of Three-dimensional Objects. *Vision Research*, *32*(12), 2385–2400.

Edelman, S., Bülthoff, H. H., & Weinshall, D. (1989). *Stimulus Familiarity Determines Recognition Strategy for Novel 3d Objects* (Tech. Rep. No. AI Memo No 1138). Boston, MA: MIT.

Erdogan, G., Chen, Q., Garcea, F. E., Mahon, B. Z., & Jacobs, R. A. (2016). Multisensory Part-based Representations of Objects in Human Lateral Occipital Cortex. *Journal of Cognitive Neuroscience*, *28*(6), 869–881.

Erdogan, G., Yildirim, I., & Jacobs, R. A. (2015). From Sensory Signals to Modality-Independent Conceptual Representations: A Probabilistic Language of Thought Approach. *PLOS Comput Biol*, *11*(11), e1004610.

Feldman, J. (2000). Minimization of Boolean Complexity in Human Concept Learning. *Nature*, *407*(6804), 630–3.

Feldman, J., & Singh, M. (2006). Bayesian Estimation of The Shape Skeleton. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(47), 18014–9.

Feldman, J., Singh, M., Briscoe, E., Froyen, V., Kim, S., & Wilder, J. (2013). An Integrated Bayesian Approach To Shape Representation and Perceptual Organization. In S. J. Dickinson & Z. Pizlo (Eds.), *Shape Perception in Human and Computer Vision* (pp. 55–70). London: Springer London.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active Vision: The Psychology of Looking and Seeing.* New York, NY: Oxford University Press.

Foster, D. H., & Gilson, S. J. (2002). Recognizing Novel Three-dimensional Objects by Summing Signals from Parts and Views. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1503), 1939–1947.

Freeman, W. T. (1996). The Generic Viewpoint Assumption in a Bayesian Framework. In D. C. Knill & W. Richards (Eds.), *Perception As Bayesian Inference* (pp. 365–389). New York, NY, USA: Cambridge University Press.

Fu, K. S. (1986). A Step Towards Unification of Syntactic and Statistical Pattern Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8*(3), 398–404.

Gauthier, I., James, T. W., Curby, K. M., & Tarr, M. J. (2003). The Influence of Conceptual Knowledge on Visual Discrimination. *Cognitive Neuropsychology*, *20*(3-6), 507–523.

Gauthier, I., & Tarr, M. J. (2016). Visual Object Recognition: Do We (Finally) Know More Now Than We Did? *Annual Review of Vision Science*, *2*(1), 377–396.

Ghose, T., & Liu, Z. (2013). Generalization Between Canonical and Non-canonical Views in Object Recognition. *Journal of Vision*, *13*(1), 1–1.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering Object Representations through Category Learning. *Cognition*, *78*(1), 27–43.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, *82*(4), 711–732.

Grenander, U., & Miller, M. (2007). *Pattern Theory: From Representation to Inference.* New York, NY: Oxford University Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic Models of Cognition: Exploring Representations and Inductive Biases. *Trends in Cognitive Sciences*, *14*(8), 357–64.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging Levels of Analysis for Probabilistic Models of Cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The Lateral Occipital Complex and its Role in Object Recognition. *Vision Research*, *41*(10-11), 1409–1422.

Hastings, W. (1970). Monte Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika*, *57*(1), 97–109.

Hayhoe, M., & Ballard, D. (2005). Eye Movements in Natural Behavior. *Trends in Cognitive Sciences*, *9*(4), 188–94.

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

Hubel, D. H., & Wiesel, T. N. (1962). Receptive Fields, Binocular Interaction and Functional

Architecture in the Cat's Visual Cortex. *The Journal of Physiology*, *160*(1), 106–154.2.

Hummel, J. E., & Biederman, I. (1992). Dynamic Binding in a Neural Network for Shape Recognition. *Psychological Review*, *99*(3), 480.

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian Learning Theory Applied to Human Cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 8–21.

James, T. W., Humphrey, G. K., Gati, J. S., Servos, P., Menon, R. S., & Goodale, M. A. (2002). Haptic Study of Three-dimensional Objects Activates Extrastriate Visual Areas. *Neuropsychologia*, *40*(10), 1706–1714.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv:1408.5093*.

Jones, E., Oliphant, T., Peterson, P., & others. (2001). *SciPy: Open Source Scientific Tools for Python.* Retrieved from `http://www.scipy.org/`

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object Perception as Bayesian Inference. *Annual Review of Psychology*(55), 271–304.

Kersten, D., & Yuille, A. (2003). Bayesian Models of Object Perception. *Current Opinion in Neurobiology*, *13*(2), 150–158.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput Biol*, *10*(11), e1003915.

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception As Bayesian Inference.* New York, NY, USA: Cambridge University Press.

Kourtzi, Z., & Connor, C. E. (2011). Neural Representations for Object Perception: Structure, Category, and Adaptive Coding. *Annual Review of Neuroscience*, *34*(1), 45–67.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modelling Biological

Vision and Brain Information Processing. *Annual Reviews of Vision Science*, *1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25* (pp. 1097–1105).

Kulis, B. (2013). Metric Learning: A Survey. *Foundations and Trends® in Machine Learning*, *5*(4), 287–364.

Kulkarni, T., Yildirim, I., Kohli, P., Freiwald, W., & Tenenbaum, J. (2014). Deep Generative Vision as Approximate Bayesian Computation. In *NIPS 2014 ABC Workshop.*

Lacey, S., & Sathian, K. (2011). Multisensory Object Representation: Insights from Studies of Vision and Touch. In *Progress in Brain Research* (1st ed., Vol. 191, pp. 165–76). Elsevier B.V.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building Machines That Learn and Think Like People. *arXiv:1604.00289 [cs, stat]*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*(7553), 436–444.

Lehky, S., & Tanaka, K. (2016). Neural Representation for Object Recognition in Inferetemporal Cortex. *Current Opinion in Neurobiology*, *37*, 23–35.

Liu, J. S. (2004). *Monte Carlo Strategies in Scientific Computing.* New York, NY: Springer New York.

Liu, Z. (1996). Viewpoint Dependency in Object Representation and Recognition. *Spatial Vision*, *9*(4), 491–521.

Liu, Z., Kersten, D., & Knill, D. C. (1999). Dissociating Stimulus Information from Internal Representation—A Case Study in Object Recognition. *Vision Research*, *39*(3), 603–612.

Longuet-Higgins, H. C. (1990). Recognizing Three Dimensions. *Nature*, *343*(6255), 214–215.

Marr, D., & Nishihara, H. K. (1978). Representation and Recognition of the Spatial Organization of Three-dimensional Shapes. *Proceedings of the Royal Society of London.*

Series B, *200*(1140), 269–94.

Marsolek, C. J. (1999). Dissociable Neural Subsystems Underlie Abstract and Specific Object Recognition. *Psychological Science*, *10*(2), 111–118.

Mehta, P., & Schwab, D. J. (2014). An Exact Mapping between the Variational Renormalization Group and Deep Learning. *arXiv:1410.3831*.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, *21*(6), 1087.

Minsky, M. L., & Selfridge, O. G. (1961). Learning in Random Nets. In *Fourth London Symposium on Information Theory.* London: Butterworth Ltd.

Newell, F. N. (2010). Visuo-Haptic Perception of Objects and Scenes. In J. Kaiser & M. J. Naumer (Eds.), *Multisensory Object Perception in the Primate Brain* (pp. 251–271). New York, NY: Springer New York.

Nguyen, A., Yosinski, J., & Clune, J. (2014). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *arXiv:1412.1897*.

Palmer, S. (1999). *Vision Science: Photons to Phenomenology.* Bradford Book.

Palmeri, T. J., & Gauthier, I. (2004). Visual Object Understanding. *Nature Reviews Neuroscience*, *5*(4), 291–303.

Patel, A. B., Nguyen, T., & Baraniuk, R. G. (2015). A Probabilistic Theory of Deep Learning. *arXiv:1504.00641*.

Peissig, J. J., & Tarr, M. J. (2007). Visual Object Recognition: Do We Know More Now Than We Did 20 Years Ago? *Annual Review of Psychology*, *58*, 75–96.

Poggio, T., & Edelman, S. (1990). A Network that Learns to Recognize Three-dimensional Objects. *Nature*, *343*(6255), 263–6.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic Brains: Knowns and Unknowns. *Nature Neuroscience*, *16*, 1170–8.

Qi, C. R., Su, H., Niessner, M., Dai, A., Yan, M., & Guibas, L. J. (2016). Volumetric and

Multi-View CNNs for Object Classification on 3d Data. *arXiv:1604.03265 [cs]*.

Rezende, D. J., Eslami, S. M. A., Mohamed, S., Battaglia, P., Jaderberg, M., & Heess, N. (2016). Unsupervised Learning of 3d Structure from Images. *arXiv:1607.00662 [cs, stat]*.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience*, *2*(11), 1019–25.

Riesenhuber, M., & Poggio, T. (2000). Models of Object Recognition. *Nature Neuroscience*, *3*, 1199–204.

Riesenhuber, M., & Poggio, T. (2002). Neural Mechanisms of Object Recognition. *Current Opinion in Neurobiology*, *12*(2), 162–8.

Rock, I., & DiVita, J. (1987). A Case of Viewer-centered Object Perception. *Cognitive Psychology*, *19*(2), 280–293.

Saiki, J., & Hummel, J. E. (1998). Connectedness and the Integration of Parts with Relations in Shape Perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(1), 227–51.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational Approximations to Rational Models: Alternative Algorithms for Category Learning. *Psychological Review*, *117*(4), 1144–67.

Schultz, M., & Joachims, T. (2003). Learning a Distance Metric from Relative Comparisons. In *Advances in Neural Information Processing Systems.*

Serre, T., Oliva, A., & Poggio, T. (2007). A Feedforward Architecture Accounts for Rapid Categorization. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(15), 6424–9.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust Object Recognition with Cortex-like Mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–26.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A.

(2014). Going Deeper with Convolutions. *arXiv:1409.4842*.

Szegedy, C., Zaremba, W., & Sutskever, I. (2013). Intriguing Properties of Neural Networks. *arXiv:1312.61*.

Talton, J., Yang, L., Kumar, R., Lim, M., Goodman, N., & Měch, R. (2012). Learning Design Patterns with Bayesian Grammar Induction. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology - UIST '12* (pp. 63–74). ACM Press.

Tarr, M. J., & Bülthoff, H. H. (1995). Is Human Object Recognition Better Described by Geon Structural Descriptions or by Multiple Views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1494–1505.

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional Object Recognition is Viewpoint Dependent. *Nature Neuroscience*, *1*(4), 275–277.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, *22*(4), 1701–1728.

Tjan, B. S., & Legge, G. E. (1998). The Viewpoint Complexity of an Object-recognition Task. *Vision Research*, *38*(15–16), 2335–2350.

Tu, Z., Chen, X., Yuille, A. L., & Zhu, S.-c. (2005). Image Parsing : Unifying Segmentation, Detection, and Recognition. *International Journal of Computer Vision*, *63*(2), 113–140.

Tversky, B., & Hemenway, K. (1984). Objects, Parts, and Categories. *Journal of Experimental Psychology: General*, *113*(2).

Ullman, S. (1989). Aligning Pictorial Descriptions: An Approach to Object Recognition. *Cognition*, *32*(3), 193–254.

Ullman, S., & Basri, R. (1991). Recognition by Linear Combinations of Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*(10), 992–1006.

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik.* Leipzig : Leopold Voss.

Wallraven, C., Bülthoff, H. H., Waterkamp, S., van Dam, L., & Gaißert, N. (2014). The Eyes Grasp, the Hands See: Metric Category Knowledge Transfers between Vision and Touch. *Psychonomic Bulletin & Review*, *21*(4), 976–85.

Wiseman, S., MacLeod, C. M., & Lootsteen, P. J. (1985). Picture Recognition Improves with Subsequent Verbal Information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(3), 588–595.

Wu, J., Zhang, C., Xue, T., Freeman, W. T., & Tenenbaum, J. B. (2016). Learning a Probabilistic Latent Space of Object Shapes via 3d Generative-Adversarial Modeling. *arXiv:1610.07584 [cs]*.

Yildirim, I., & Jacobs, R. (2013). Transfer of Object Category Knowledge Across Visual and Haptic Modalities: Experimental and Computational Studies. *Cognition*, *126*, 135–148.

Yildirim, I., Kulkarni, T. D., Freiwald, W. a., & Tenenbaum, J. B. (2015). Efficient and Robust Analysis-by-synthesis in Vision: A Computational Framework, Behavioral Tests, and Modeling Neuronal Representations. In *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian Inference: Analysis by Synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–8.

Yuille, A., & Mottaghi, R. (2016). Complexity of Representation and Inference in Compositional Models with Part Sharing. *Journal of Machine Learning Research*, *17*(11), 1–28.

Zhang, K., & Shasha, D. (1989). Simple Fast Algorithms for the Editing Distance between Trees and Related Problems. *SIAM Journal on Computing*, *18*(6), 1245–1262.

Zhu, L., Chen, Y., & Yuille, A. (2007). Unsupervised Learning of a Probabilistic Grammar for Object Detection and Parsing. In *Advances in Neural Information Processing Systems 19* (pp. 1617–1624).

Zhu, L., Chen, Y., & Yuille, A. (2009). Unsupervised Learning of Probabilistic Grammar-

Markov Models for Object Categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 114–128.