Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Vladislav Ayzenberg Date

Unique contributions of skeletal structure to shape perception and object recognition

By

Vladislav Ayzenberg
Doctor of Philosophy

Psychology

_____
Stella F. Lourenco
Advisor

_____
Daniel D. Dilks
Committee Member

_____
Hillary R. Rodman
Committee Member

_____
Michael T. Treadway
Committee Member

_____
Jocelyn Bachevalier
Committee Member

_____
Phillip Wolff
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____
Date

Unique contributions of skeletal structure to shape perception and object recognition

By

Vladislav Ayzenberg
B.A. Temple University, 2012

Advisor: Stella F. Lourenco, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Psychology
2020

Abstract

Unique contributions of skeletal structure to shape perception and object recognition
By Vladislav Ayzenberg

With seemingly little effort, humans can both identify an object across large changes in orientation and extend category membership to novel exemplars. Most remarkably, humans can accomplish these feats with little experience, categorizing never-before-seen objects from as little as one training exemplar. Although researchers have long suggested that global shape information is crucial for robust object recognition, it is unknown how humans perceptually organize visual information to create global shape percepts and use these percepts to recognize objects. In the current dissertation, I used behavioral, neural, computational, and developmental methods to test the hypothesis that a model of structure known as the medial axis, or shape skeleton, can support both perceptual organization and object recognition. Moreover, I examined whether shape skeletons play a role in rapid object learning by testing whether they can support one-shot categorization in infants. Consistent with these hypotheses, I found that a skeletal model was predictive of adult participants' object similarity and category judgments (Study 1). Moreover, neuroimaging of the adult visual system revealed that a skeletal model was predictive of the multivariate response in V3 and lateral occipital cortex (LO), regions implicated in perceptual organization and object recognition, respectively (Study 2). Finally, I found that 6- to 12-month-old infants, a population with little object experience, could categorize never-before-seen objects by their skeleton after seeing just one exemplar (Study 3). In all studies, the skeletal model best fit participants' responses (behavioral and neural) across changes in image-level properties, contour, and when controlling for other state-of-the-art artificial neural networks. Taken together, these studies highlight the unique and privileged role of shape skeletons in perceptual organization, object recognition, and one-shot categorization.

Unique contributions of skeletal structure to shape perception and object recognition

By

Vladislav Ayzenberg
B.A. Temple University, 2012

Advisor: Stella F. Lourenco, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Psychology
2020

Acknowledgments

There is an extremely long list of people without whom I would have never succeeded in graduate school.

First and foremost, I want to thank my advisor Stella Lourenco. Stella, you are an endless source of support and inspiration to me. You have literally taught me everything I know, and I could not imagine having spent my graduate career anywhere else. To Danny Dilks and the many other Psychology faculty I've had to pleasure of working with, thank you for constantly challenging me to be a better scientist and really just making the whole experience a lot of fun. Thanks also to my undergraduate mentors: Justin Harris, Nora Newcombe, and Kathy Hirsh-Pasek. I would have never started this journey without you.

To my fellow graduate students: Lauren Aulet, Jenny Merritt, Cassie Hendrix, Freddy Kamps, Ryan Brady, Sara Botto. You are like family to me and I could not imagine a better group of people to go through graduate school with. I'm excited to spend the rest of my life arguing about science with you all. To all the talented, hard-working research assistants I have had over the course of my graduate career: Sami Yousif, Samoni Nag, Meghan Hickey, Adi Rosenthal, Bahar Sener, Amy Krivoshik, Jessica Kubert. None of this would have been accomplished without your help. It was a privilege to watch many of you grow into successful scientists yourselves.

Last, but not least, thank you to my non-academic friends and family for all of their love and support.

Table of Contents

Abstract 1

Chapter 1: General Introduction 2

Chapter 2: Skeletal descriptions of shape provide unique perceptual information for object recognition
- Introduction 9
- Experiment 1 11
- Experiment 2 16
- Experiment 3 19
- Discussion 24
- Supplementary Materials 26

Chapter 3: A dual role for shape skeletons in human vision: Perceptual organization and object recognition
- Introduction 40
- Methods 41
- Results 48
- Discussion 54
- Supplementary Materials 59

Study 4: The shape skeleton supports single exemplar categorization in infants
- Introduction 64
- Experiment 1 64
- Experiment 2 68
- Discussion 71
- Supplementary Materials 72

Chapter 5: General Discussion 76

References 83

List of Figures

Figure 1	An illustration of the shape skeleton for a 2D airplane	5
Figure 2	A subset of results from Ayzenberg, Chen, Yousif, and Lourenco (2019).	7
Figure 3	Stimuli used in Study 1, Experiment 1	12
Figure 4	Results from Study 1, Experiment 1	15
Figure 5	Example stimuli and results from Study 1,Experiment 2	18
Figure 6	Examples of the three trial types used in Study 1, Experiment 3	22
Figure 7	Results from the match-to-sample task of Study 1, Experiment 3	23
Figure 8	Stimuli used in the Study 2	43
Figure 9	Regions of interest in a sample participant from Study 2	47
Figure 10	Correlation results from Study 2	49
Figure 11	Variance partitioning results from Study 2	50
Figure 12	Experimental design and results for Study 3 Experiment 1	66
Figure 13	Experimental design and results for Study 3 Experiment 2	70
Supplemental Figure 1	Discrimination accuracies for all non-skeletal models from Study 1	37
Supplemental Figure 2	All stimuli used in Study 1 Experiment 2	38
Supplemental Figure 3	The stimulus set used in Study 1 Experiment 3	39
Supplemental Figure 4	Discrimination accuracies for all non-skeletal models from Study 2	60
Supplemental Figure 5	Stimulus set used in Study 3 Experiment 1	75
Supplemental Figure 6	Stimulus set used in Study 3 Experiment 2	75

List of Tables

Table 1	Results from Study 2	51
Supplemental Table 1	Linear regression results from Study 1	34
Supplemental Table 2	Variance partitioning results from Study 1	35
Supplemental Table 3	Mixed-effect model results from Study 1	36
Supplemental Table 4	Variance partitioning results from Study 2	61
Supplemental Table 5	Mixed-effect model results from Study 2	62

Abstract

With seemingly little effort, humans can both identify an object across large changes in orientation and extend category membership to novel exemplars. Most remarkably, humans can accomplish these feats with little experience, categorizing never-before-seen objects from as little as one training exemplar. Although researchers have long suggested that global shape information is crucial for robust object recognition, it is unknown how humans perceptually organize visual information to create global shape percepts and use these percepts to recognize objects. In the current dissertation, I used behavioral, neural, computational, and developmental methods to test the hypothesis that a model of structure known as the medial axis, or shape skeleton, can support both perceptual organization and object recognition. Moreover, I examined whether shape skeletons play a role in rapid object learning by testing whether they can support one-shot categorization in infants. Consistent with these hypotheses, I found that a skeletal model was predictive of adult participants' object similarity and category judgments (Study 1). Moreover, neuroimaging of the adult visual system revealed that a skeletal model was predictive of the multivariate response in V3 and lateral occipital cortex (LO), regions implicated in perceptual organization and object recognition, respectively (Study 2). Finally, I found that 6- to 12-month-old infants, a population with little object experience, could categorize never-before-seen objects by their skeleton after seeing just one exemplar (Study 3). In all studies, the skeletal model best fit participants' responses (behavioral and neural) across changes in image-level properties, contour, and when controlling for other state-of-the-art artificial neural networks. Taken together, these studies highlight the unique and privileged role of shape skeletons in perceptual organization, object recognition, and one-shot categorization.

Chapter 1 - General Introduction

The same object produces vastly different shapes on the retina across changes in orientation, and objects of the same category have vastly different shape contours across exemplars. Yet, with little experience, human adults and infants (Biederman & Bar, 1999; Mash, Arterberry, & Bornstein, 2007), as well as nonhuman animals (Wood, 2013; Zoccolan, Oertelt, DiCarlo, & Cox, 2009), recognize objects with ease across such variations. Research in the vision sciences suggests that shape is crucial for object recognition (Biederman & Ju, 1988; Elder, 2018; Marr & Nishihara, 1978). Humans readily use shape to recognize objects in the absence of other visual information (e.g., texture and shading; Biederman & Ju, 1988; Wagemans et al., 2008) and both adults and children preferentially categorize novel objects by their shape across conflicting color and texture cues (Elder & Velisavljević, 2009; Landau, Smith, & Jones, 1988). Moreover, human representations of shape are robust to changes in view (Biederman, 1987; Biederman & Bar, 1999), contour perturbation (Kanizsa, 1976; Spröte, Schmidt, & Fleming, 2016), and deformations from bending or stretching (e.g., hand poses; Barenholtz & Tarr, 2008; Leyton, 1989; Spröte & Fleming, 2016), suggesting a reliance on global shape properties over local contour information (Baker & Kellman, 2018; Sanocki, 1993). However, it remains unknown as to how humans form representations of global shape in order to recognize objects (Elder, 2018).

Models of Object Recognition

The vision sciences have proposed several theories to explain how humans form robust representations of shape in service of object recognition. One class of theories has proposed that object shape is represented via a series of diagnostic object viewpoints and recognized by comparing the degree of image-level similarity between the current view of an object and a representation stored in memory (Tarr & Bülthoff, 1995, 1998). Computational implementations of these theories, such as the Gabor-jet model (Margalit, Biederman, Herald, Yue, & von der Malsburg, 2016), successfully approximate human object discrimination judgments in some contexts (Yue,

Biederman, Mangini, Malsburg, & Amir, 2012), and match the representations of early- and mid-level visual cortical areas (Olshausen & Field, 1996; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007). However, these models have difficulty recognizing objects from novel viewpoints (Biederman & Bar, 1999; Biederman & Gerhardstein, 1993), or categorizing object exemplars with different image-level properties (Amir, Biederman, & Hayworth, 2012; Hummel, 2000).

Another class of theories proposes that humans recognize objects by diagnostic component parts and a coarse spatial structure (Biederman, 1987; Biederman & Gerhardstein, 1993). This theory is better able to explain how humans recognize objects in novel viewpoints (Biederman, 2000; Biederman & Gerhardstein, 1993) and is consistent with both behavioral and neural evidence that humans represent object shape via parts and their relations (Brincat & Connor, 2006; Pasupathy & Connor, 2002; Yamane, Carlson, Bowman, Wang, & Connor, 2008). However, it is unclear from this perspective how organisms identify category exemplars with different component parts (Barenholtz & Tarr, 2006; Tarr & Bülthoff, 1995), and there have been few successful computational implementations of component description models (Crouzet & Serre, 2011; Hummel & Biederman, 1992).

Most recently, researchers have proposed that humans represent objects via a learned set of diagnostic features, such as those used by artificial neural networks (ANNs; Krizhevsky, Sutskever, & Hinton, 2012; Ullman, Assif, Fetaya, & Harari, 2016). These models are thought to form human-like shape representations by extracting and combining object features over a series of hierarchically organized layers (DiCarlo, Zoccolan, & Rust, 2012; Kubilius, Bracci, & Op de Beeck, 2016). Compared to image similarity and component description models, ANNs provide the best approximation of human behavioral and neural object responses (Schrimpf et al., 2018). However, ANNs require thousands more supervised training examples than humans to achieve such performance (Zador, 2019), and they are unable to categorize objects on the basis of global shape information (Ayzenberg, Sener, & Lourenco, under review; Baker, Lu, Erlikhman, & Kellman, 2018).

Indeed, ANNs show catastrophic object recognition failures even in the presence of minor image distortions that are imperceptible to humans (Szegedy et al., 2013). Thus, these theories and models are unable to explain the robustness of human shape representations, nor the speed at which humans form shape representations to recognize objects.

Skeletons in computer and human vision

One class of models that can both explain how humans represent object shape and how objects are learned with little experience are known as shape skeletons, or medial axis models. Shape skeletons are a class of geometric models that describe shape via the set of symmetry axes that lie equidistant between two or more points along the boundary (Blum, 1967; 1973; see Figure 1). For most shapes, the axes are organized hierarchically, such that there may be a series of parent axes that describe the shape's coarse global geometry, as well as smaller 'off-shoot' axes that describe individual component parts. More specifically, they describe a shape's structure by providing a low-dimensional description of the spatial relations between contours, as well as component parts. The strength of such a description for human vision is that it can support perceptual organization by specifying how local visual features are integrated into a complete shape. Moreover, because skeletons are tolerant to variations in shape, they can also support recognition of objects across changes in viewpoint or exemplar. Importantly, because shape skeletons can be computed for any object, they can be used to categorize novel objects with minimal training.

Consistent with these possibilities, computer vision research has shown that such a description can be used to determine an object's shape from noisy or incomplete contour information (Feldman & Singh, 2006; Kimia, 2003; Wilder et al., 2019), and to identify objects across never-before-seen viewpoints and category exemplars (Sebastian, Klein, & Kimia, 2004; Trinh & Kimia, 2011). Indeed, modern skeletal algorithms (i.e., pruned medial axis models; Shaked & Bruckstein, 1998; Wieser, Seidl, & Zeppelzauer, 2017), are particularly good descriptors of an

object's global shape because their structure remains stable across contour variations typical of natural contexts (e.g., perturbations, bending; Feldman & Singh, 2006; Liu & Geiger, 1999; Trinh & Kimia, 2011). Moreover, incorporating pruned models into off-the-shelf ANNs significantly improves their performance on visual perception tasks (Rezanejad et al., 2019).

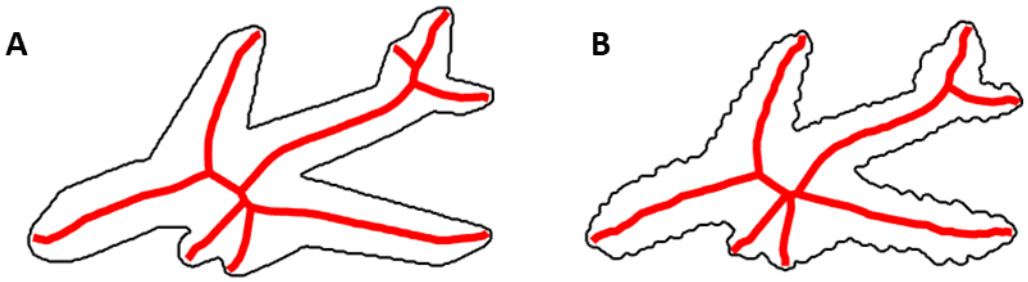


Figure 1. An illustration of the shape skeleton for a 2D airplane with (B) and without (A) perturbed contours. A strength of a skeletal model is that it can describe an object's shape structure across variations in contour. Skeletons computed using the ShapeToolbox (Feldman & Singh, 2006).

Increasingly, behavioral research with humans has suggested that participants represent the skeletons of shapes. Participants show increased contrast sensitivity for Gabor patches when they fall along the skeleton of a 2D shape (Kovács, Fehér, & Julesz, 1998; Kovacs & Julesz, 1994) and they are more likely to direct their attention to points within the shape that correspond to the skeleton (Firestone & Scholl, 2014; Psotka, 1978). Other research has shown that manipulating the similarity between object skeletons changes participants' abilities to discriminate those objects (Destler, Singh, & Feldman, 2019; Lowet, Firestone, & Scholl, 2018; Wilder, Feldman, & Singh, 2011). Although these studies provide preliminary evidence for the hypothesis that shape skeletons play a role in creating and comparing shape representations in human vision, no study has tested these hypotheses directly. Indeed, the extant studies did not compare shape skeletons against other plausible models of vision, leaving it unknown whether skeletons are the best fit to participants'

responses. Moreover, no study has tested whether shape skeletons support these functions in the absence of extensive visual experience.

To begin addressing these gaps in the literature, my masters research (Ayzenberg, Chen, et al., 2019) tested whether human skeletal representations were best described by a pruned medial axis model, which is tolerant to noisy or missing contours (Feldman & Singh, 2006; Shaked & Bruckstein, 1998; Wieser et al., 2017), and can support perceptual organization (Ardila, Mihalas, von der Heydt, & Niebur, 2012; Feldman et al., 2013). To this end, participants were shown a single 2D shape with either complete, perturbed, or illusory contours on a tablet computer, and were asked to tap the shape once anywhere they liked. Consistent with prior research (Firestone & Scholl, 2014; Psotka, 1978), we found that the collective pattern of participants' responses in complete shapes corresponded to the skeletons of the shapes, not other models (see Figure 2A). Importantly, in shapes with perturbed or illusory contours, responses were best fit by a pruned medial axis skeleton, rather than a skeleton that was sensitive to every edge (Figure 2B-C). That a pruned skeleton best described participants' responses in these conditions suggests that shape skeletons are a biologically plausible model of perceptual organization and can support the creation of shape percepts. However, it remains unknown whether, in addition to perceptual organization, shape skeletons also support object recognition, and, importantly, can do so with little visual experience.

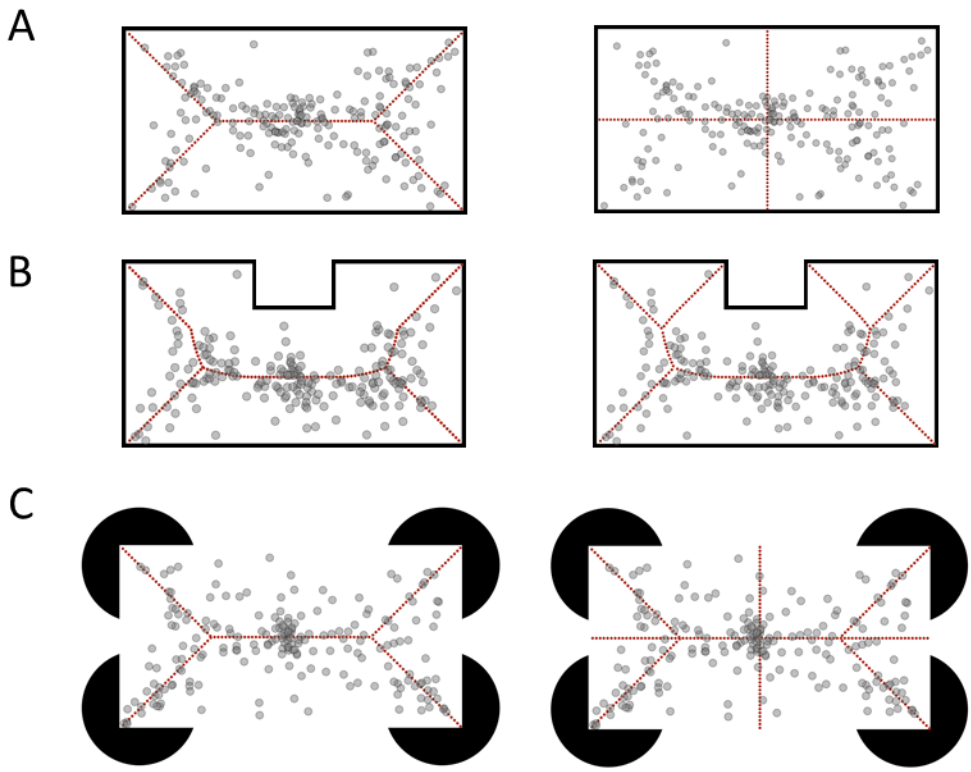


Figure 2. A subset of results from Ayzenberg, Chen, et al. (2019). Gray circles represent individual responses, and the red dotted lines represent the model they were fit with. (A) Responses and model comparison for a shape with complete contours. Participants responses were better fit by the shape skeleton (left) than another model of shape based on the principal axes (right). (B) Responses and model comparison for a shape with perturbed contours. Participants responses were better fit by a pruned skeleton (left) than a skeleton sensitive to each edge, known as the medial axis transform (MAT) (right). (C) Responses and model comparison for a shape with illusory contours. Participants responses were better fit by a pruned skeleton (left) than the MAT (right).

Current Dissertation

To test whether shape skeletons are a mechanism by which humans create shape percepts and recognize objects with little visual experience, three studies were conducted. In Study 1, we investigated whether shape skeletons play a role in object recognition by testing whether a skeletal model was uniquely predictive of human object similarity and category judgments, even when controlling for other models of vision. In Study 2, we used fMRI to provide converging neural

evidence for the hypothesis that shape skeletons are involved in both perceptual organization and object recognition. In particular, we examined whether a skeletal model was predictive of the multivariate patterns in V3 and LO, regions classically associated with perceptual organization and object recognition, respectively. Finally, in Study 3, we examined whether shape skeletons support one-shot object categorization by testing whether infants could categorize never-before-seen objects by their skeleton from just one training example. Importantly, in all three studies of the dissertation I examined the unique contributions shape skeletons to human perception by comparing them to image-similarity, component descriptions, and ANN models. *The studies in this dissertation are either published or under review and, therefore, are presented below in minimally altered form.* Together, these studies shed light on the mechanisms that support rapid and robust object recognition in humans and expand our understanding of the organization of the mind and brain more generally.

Chapter 2 - Skeletal descriptions of shape provide unique perceptual information for object recognition (Ayzenberg & Lourenco, 2019b)

Accumulating evidence suggests that the skeletal structure of objects is extracted by the primate visual system during shape perception. In particular, behavioral studies have shown that human participants extract the skeletons of different 2D shapes (Firestone & Scholl, 2014; Harrison & Feldman, 2009; Kovács et al., 1998; Kovacs & Julesz, 1994; Psotka, 1978) and those skeletal structures remain relatively stable across border disruptions resulting from perturbations or illusory contours (Ayzenberg, Chen, et al., 2019). Increasingly, studies have shown that skeletal structures may be represented in three dimensions (3D) within an object-centered reference frame. Indeed, human adults are better at discriminating 3D objects by skeletal differences than by differences in component parts (e.g., part orientation; Lowet et al., 2018). Moreover, studies using neural recording (i.e., fMRI and electrophysiology) with humans and monkeys have found sensitivity to 3D object skeletons in high-level visual cortical areas (e.g., IT), including those known to support object recognition (Hung, Carlson, & Connor, 2012; Lescroart & Biederman, 2012). Skeletal sensitivity in these regions was decoded across changes in orientation and variations in local shape properties, suggesting a 3D object-centered representation that is robust to changes in viewpoint and component parts.

Despite the success of skeletal descriptions in computer vision systems (Trinh & Kimia, 2011) and their biological plausibility in the primate visual system (Hung et al., 2012), shape skeletons are rarely incorporated into models of object recognition. Instead, modern computational approaches to object recognition emphasize image statistics (Oliva & Torralba, 2006) or hierarchical feature extraction operations such as those implemented by CNNs (Krizhevsky et al., 2012; Serre, Wolf, et al., 2007). Yet, without explicitly invoking any skeletal description, these models match human performance on object recognition tasks , and they are predictive of both human behavioral and neural responses (Jozwik, Kriegeskorte, Storrs, & Mur, 2017; Schrimpf et al.,

2018; Yamins et al., 2014). Even models that do emphasize global shape properties do so by describing the local properties of components parts (e.g., geons) and coarse, categorically-defined, spatial relations (Biederman, 1987; Hummel, 2001), not a skeletal structure. Given that these other models successfully approximate human object recognition, one might ask whether skeletal descriptions of shape are necessary for human object recognition at all. Thus, in the current study, we tested the degree to which skeletal descriptions of shape make unique, and possibly privileged, contributions to human object recognition in comparison to several other models of shape and object perception.

If the shape's skeletal structure provides unique contributions to object recognition, then humans should perceive objects with similar skeletons as more similar to one another, even when controlling for other models. Moreover, if skeletal structures are a privileged source of information for object recognition, then humans should favor the shape skeleton over both non-shape based models of visual similarity, as well as other descriptors of shape. To this end, we assessed whether participants' perceptual judgments of object similarity scaled with the skeletal similarity between novel 3D objects (Experiment 1), including objects whose coarse spatial relations could not be used for judging similarity (Experiment 2). We also tested how participants classified objects when the shape's skeletal structure was placed in conflict with the object's surface form, a manipulation that altered the shape's contours and non-accidental properties (NAPs) without changing its skeleton (Experiment 3). In all cases, we examined the unique contributions of skeletal structures in object recognition by contrasting the shape skeleton with models of vision that do not explicitly incorporate a skeletal structure, but are nevertheless predictive of human object recognition. These models included those that describe visual similarity by their image statistics, namely, the Gabor-Jet (GBJ) model (Margalit et al., 2016) and GIST model (Oliva & Torralba, 2001), as well as biologically plausible neural network models, namely, the HMAX model (Serre, Wolf, et al., 2007) and AlexNet, a CNN pre-trained to identify objects (Krizhevsky et al., 2012). To anticipate our findings, a model of

skeletal similarity was predictive of participants' perceptual similarity and classification judgments even when accounting for these other models, suggesting that skeletal descriptions of shape play a crucial role in human object recognition, independent of other models of shape and object perception.

Experiment 1 – Is perceived object similarity uniquely predicted by a model of skeletal similarity?

Here we tested one of the predictions outlined above: namely, as object skeletons become more similar, participants should judge the objects as being more alike. To test for a relation between human perceptual judgments and the shape skeleton, we generated a novel set of 3D objects that varied in their skeletal structures. Crucially, we compared the predictive power of skeletal descriptions to other models of visual similarity and tested the degree to which a model of skeletal similarity explained unique variance in human perceptual judgments.

Stimuli and experimental design. A total of 150 3D objects consisting of 30 skeletons were generated (see Figure 3A). All objects were comprised of three segments and were normalized for overall size (see Supplemental Methods). Each object was rendered with five surface forms, serving to change the visible shape of the object on the retina without altering the underlying skeleton (see Figure 3B and Supplemental Methods). Skeletal similarity between every object was calculated in 3D, object-centered, space as the mean Euclidean distance between each point on one skeleton and the closest point on the second skeleton following maximal alignment (see Supplemental Methods). We chose to test a 3D skeletal description because of behavioral (Erdogan & Jacobs, 2017) and neural (Yamane et al., 2008) evidence for 3D object-centered representations in the visual system.

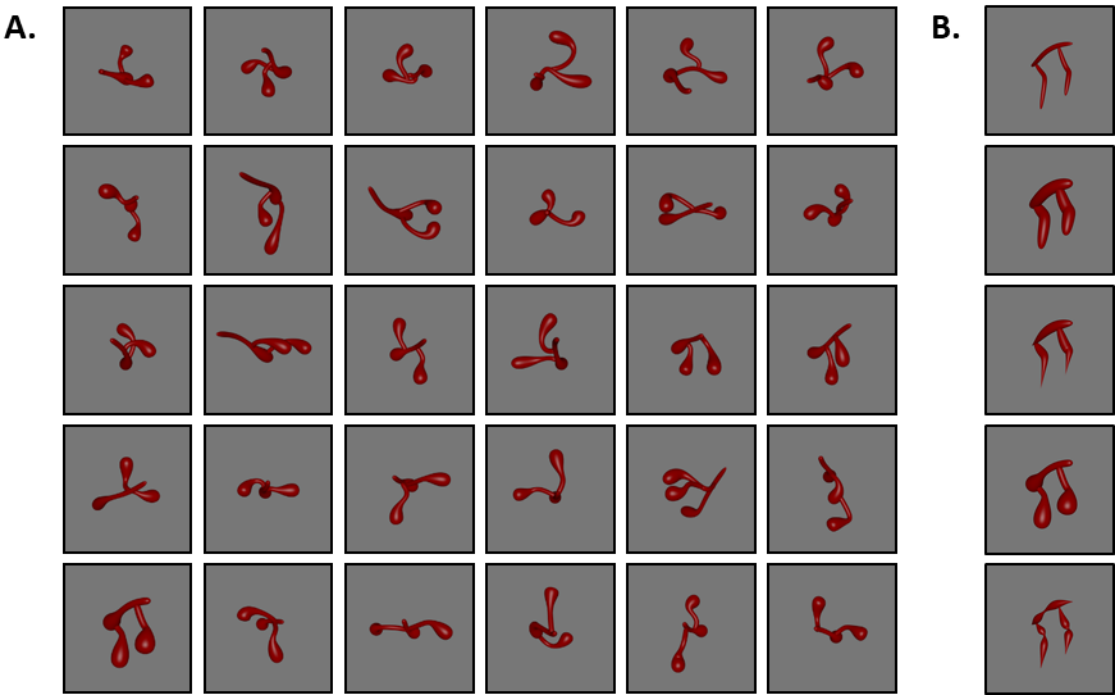


Figure 3. Stimuli used in Experiment 1. (A) Objects were procedurally generated to have different skeletal structures. (B) Each object was also rendered with five surface forms so as to vary in contour shape and non-accidental properties (NAPs) without disrupting the object's skeleton. A cluster analysis revealed that the first and second surface forms (from top to bottom) were comprised of the same NAPs (see Experiment 3 and Supplemental Methods for more stimulus details). Subsets of these stimuli were used in Experiments 2 and 3 (see Supplemental Methods).

Participants (n = 42) were administered a discrimination task in which they were shown images of two objects presented simultaneously in one of three depth orientations (-30°, 0°, +30°), with either the same or different skeletons. Participants were instructed to decide whether the two images showed the same or different object (independent of orientation). Participants were given unlimited time to respond but the instructions emphasized speed and accuracy.

We chose to use an untimed discrimination task where the objects were presented simultaneously in order to minimize task demands. However, we also confirmed that this task could be accomplished in a speeded context and found comparable performance to that reported below (see Supplemental Experiment 1).

Results and discussion. Participants discriminated the objects significantly above chance (0.50) $M_{accuracy}$ = 0.80, $t(41)$ = 17.64, $p < .001$, d = 2.72 (M_{RT} = 2129 ms). Thus, even though our stimulus set may be considered one class of object, and potentially difficult to discriminate, the objects differed sufficiently to support accurate discrimination (see also Supplemental Experiment 1 for comparable performance in a speeded version of the task).

To analyze whether a skeletal model was predictive of human object judgments, we converted participants' binary discrimination judgments for each object pair into a continuous dissimilarity score. Dissimilarity scores for each object pair were computed by taking the mean discrimination accuracy for each pair across all participants. Human judgments were compared to each model by regressing human dissimilarity scores on model dissimilarity scores (see Supplemental Methods).

Skeletal similarity was a significant predictor of participants' judgments, r = 0.30, $p < 0.001$, explaining 9% of the variance (significance determined via permutation test; see Figure 4). That is, as the similarity between skeletal structures increased, participants were more likely to judge the objects as the same. However, one might ask whether another model of vision, which does not incorporate skeletal information, would also correlate with human judgments. To answer this question, we compared participants' judgments to GBJ, GIST, HMAX, and AlexNet models. When compared independently, these models were all predictive of participants' judgments (rs = 0.25 – 0.32, r^2 = 6% – 11%; see Figure 4), with no significant differences between models (overlapping confidence intervals). For context, a noise ceiling representing a hypothetical true model (calculated by repeatedly splitting participants' data into two sets and correlating them to one another; 1000 iterations) was computed: r_{mean} = 0.50, SE = 0.03 (see Figure 4A).

Because the different models were predictive of participants' judgments to similar degrees, and because objects with similar skeletons might also have similar image-level properties, it was important to test whether the different models accounted for the same variance in participants'

judgments, or whether a model of skeletal similarity explained unique variance. To this end, we conducted a regression analysis wherein all of the models and the most predictive layer of AlexNet (Skeleton ∪ GBJ ∪ GIST ∪ HMAX ∪ AlexNet-fc6) were included as predictors of human dissimilarity judgments. Together, these models explained 20.5% of the variance in human judgments, with Skeletal and GBJ models each explaining significant unique variance (*ps* < .01; see Supplemental Table 1). To ensure that the predictive power of the skeletal model was not simply the result of a suppression effect, we tested the skeletal model individually against every other model (Skeleton ∪ GBJ; Skeleton ∪ GIST; Skeleton ∪ HMAX; Skeleton ∪ AlexNet-fc6). Skeletal similarity was predictive of human judgments in each case (r^2 = 14% –18%, *ps* < .001).

Finally, to better understand how the amount of variance explained by the skeletal model compared to the other models, we used variance partitioning analyses (Bonner & Epstein, 2018; Lescroart, Stansbury, & Gallant, 2015). These analyses allowed us to determine how much of the total explained variance was unique to the different models and how much was shared by a combination of models. These analyses revealed that the model of skeletal similarity accounted for the greatest amount of unique variance in participants' responses (6.6%) explaining 33.13% of the total explainable variance (see Figure 4B and Supplemental Table 2). A 4-predictor model consisting of the GBJ, GIST, HMAX, and AlexNet-fc6 models accounted for the next greatest amount of variance in participants' responses (3.1%) accounting for 15.49% of the total explainable variance (see Supplemental Table 2). Taken together, these analyses suggest that, although other models of visual similarity were predictive of participants' perceptual judgments, a model of skeletal similarity uniquely explained these judgments. These results suggest that skeletal structures may be an important source of information in making object identity.

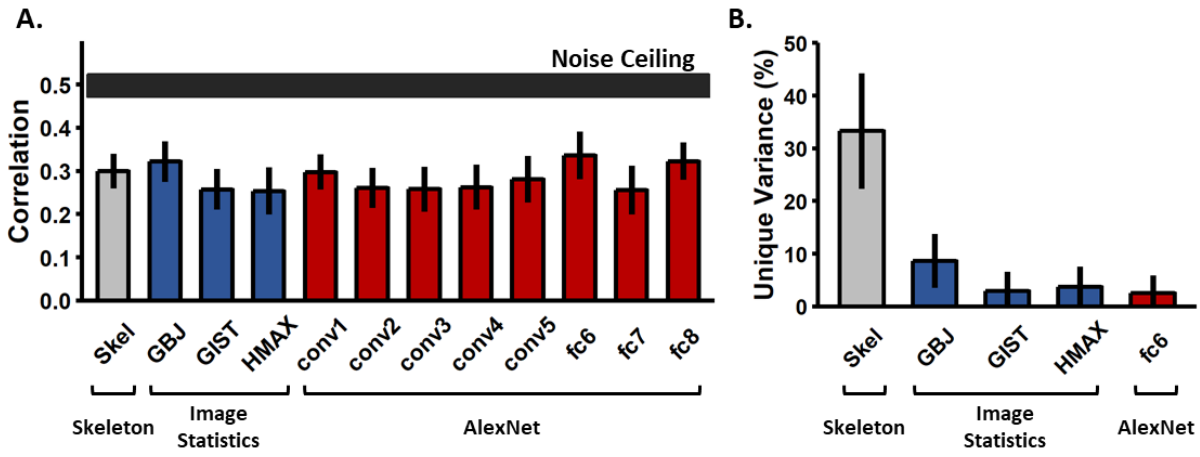


Figure 4. Results from Experiment 1. (A) Bar plot displaying the correlations (Pearson) between each model and human perceptual similarity judgments (error bars are bootstrapped SE). Models did not differ significantly from each other in the degree to which they predicted human judgments. The horizontal black bar represents the noise ceiling, which indicates the expected performance of the true model given the noise in the data (width represents SE). (B) Bar plot displaying the percentage of unique variance accounted for by each model. A model of skeletal similarity explained the most unique variance (33.13% of total explainable variance) when compared to any single model or combination of models (see Supplemental Table 2 for the unique and shared variance explained by all model combinations).

A potential concern with these findings is that, because we created objects that varied in skeletal similarity, it was inevitable that a model of skeletal similarity would predict participants' performance. We would emphasize, however, that other, non-skeletal models were also predictive of participants' judgments, suggesting that participants incorporated other visual properties into their judgments. Nevertheless, to address this concern more directly, we tested whether the objects differed sufficiently for non-skeletal models to discriminate between them. A feature vector was extracted for every image (30 skeletons × 5 surface forms × 3 orientations) from each of these models (GBJ, GIST, HMAX, AlexNet-fc8). Then, for each model and object pair (same surface form), a linear support vector machine (SVM) classifier was trained to label objects using two object orientations; its ability to label the objects was tested using the third orientation. This procedure

was repeated for every surface form and every combination of orientations between objects (0º × 0º; 0º × 30º; 0º × -30º; 30º × 30º; 30º × -30º; -30º × -30º). A final discrimination score was computed for each object pair by averaging the decoding accuracies across every surface form and combination of orientations. This analysis revealed that every model could discriminate between objects significantly above chance (0.50; *Ms* > 0.75), *ts* > 41.88, *ps* < .001, *ds* > 2.01 (see Supplemental Figure 1). Together, these findings demonstrate that the objects within our stimulus set were sufficiently different along other visual dimensions that non-skeletal models could accurately discriminate them.

Experiment 2 – Can perceived similarity be explained by another model of structure?

The results of Experiment 1 suggest that humans incorporate skeletal representations when making object similarity judgments. However, an alternative possibility is that participants' sensitivity reflected a different model of structure, namely one based on the coarse spatial relations between object parts (Biederman & Gerhardstein, 1993; Hummel, 2000; Hummel & Stankiewicz, 1996). A model based on coarse spatial relations suggests that the structure of an object is represented by the categorical relations between component parts (e.g., components above one another vs. components side-by-side). In contrast to skeletal descriptions of shape, which describe quantitative relations between component parts, a coarse spatial-relations model would predict that only qualitative changes to the overall spatial arrangement of the parts (e.g., changing component relations from 'side-by-side' to 'end-to-end') should influence object recognition. Yet objects with similar spatial relations also have more similar skeletal structures. Thus, the relation between skeletal similarity and human perceptual judgments in Experiment 1 could have reflected the co-variation between the shape skeleton and an object's coarse spatial relations.

Here we tested whether participants' judgments of perceptual similarity were influenced by an object's skeletal structure even when coarse spatial relations were held constant, and thus unable to be used as a similarity cue. If perception of object shape is based on a skeletal structure,

then a proportional change to the shape skeleton would elicit a proportional decrease in recognition, even when the coarse spatial relations are unchanged. Thus, as two skeletons become more dissimilar, participants should judge the objects as more different from one another.

Stimuli and experimental design. To test whether proportional changes to the shape skeleton led to proportional deficits in recognition, we adapted three objects from Experiment 1 to have six increments of skeletal dissimilarity (0%, 10%, 20%, 30%, 40%, 50% difference; see Supplemental Methods for additional details). The three objects consisted of distinct coarse spatial relations (see Figure 5A, Supplemental Methods, and Supplemental Figure 2). Changes to the skeleton were implemented by moving one component along the length of another component in 10% increments (see Figure 5A). This manipulation caused systematic changes to the shape skeleton without changing the coarse spatial relations between the object's component parts. Thus, if skeletal similarity affects participants' perceptual judgments, independent of coarse spatial relations, then performance should scale proportionally with changes to the skeleton.

Participants (n = 42) completed a discrimination task in which they were shown two simultaneously presented objects and were instructed to judge whether the objects were the same or different in their coarse spatial relations. Crucially, participants were instructed to ignore any changes to the precise positions of the object parts (i.e., exact skeleton) so as to make their decision on the basis of "overall shape." Participants were given a familiarization phase to ensure they understood that these instructions referred to objects with the same coarse spatial relations (e.g., in Figure 5A each column consists of objects with the same "overall shape", but different skeletons). Objects were presented from three orientations that maximized the visibility of the object's structure (30°, 60°, and 90°). Participants were given unlimited time to respond but were encouraged to respond quickly and accurately.

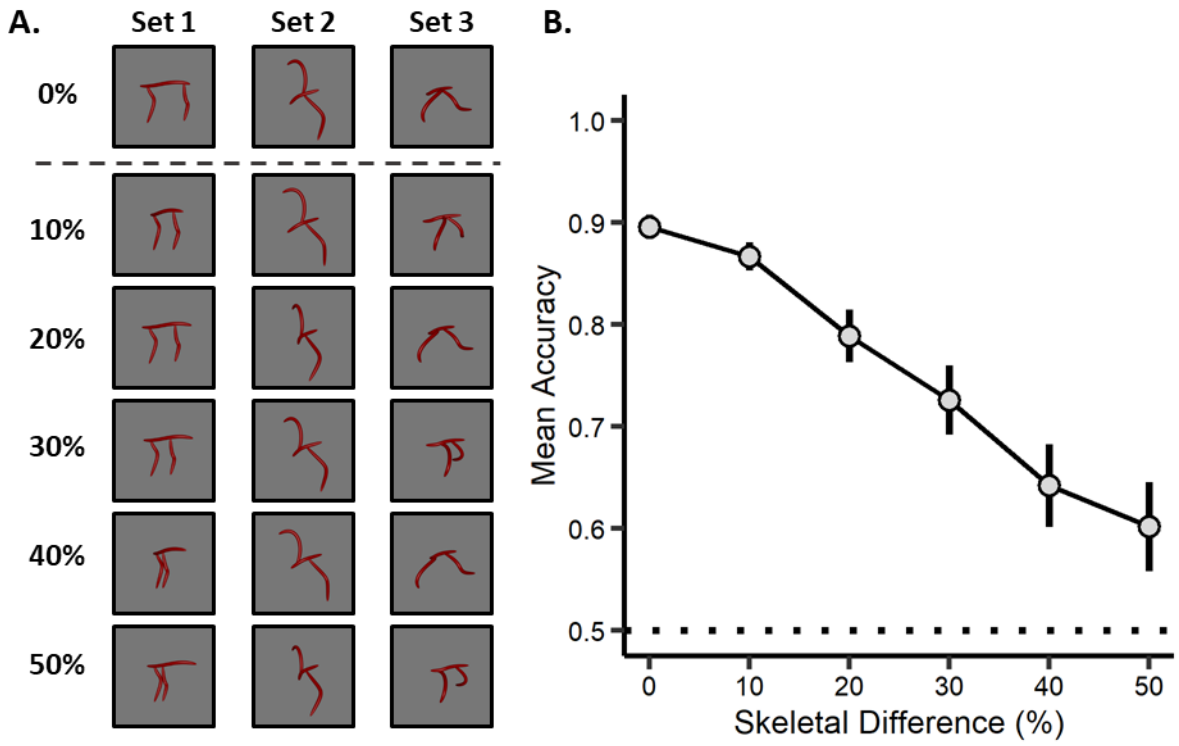


Figure 5. Example stimuli and results from Experiment 2. (A) Objects were comprised of three sets, each with distinct coarse spatial relations (separate columns). Crucially, objects with the same spatial relations varied in skeletal similarity by increments of 0%, 10%, 20%, 30%, 40%, or 50% (each row within a column). On the 'same' test trials (objects within the same column), participants were presented with a reference object (0%; top row) and an object with the same coarse spatial relations. On the 'different' test trials (objects across columns), participants were presented with objects that had different coarse spatial relations. Objects were presented in one of three orientations (30°, 60°, 90°; see Supplemental Figure 2 for full stimulus set). (B) Participants' recognition accuracy (proportion correct) for objects with the same coarse spatial relations decreased as a function of skeletal change, suggesting that humans represent object structure by their skeletons. The dotted line represents chance performance and the error bars represent SE.

Results and discussion. Participants performed significantly above chance (0.50) at every level of skeletal change, $Ms > 0.60$, $ts(34) > 2.32$, $ps < 0.026$, $ds > 0.39$, ($M_{RTs} < 1705$ ms), demonstrating that they followed the task instructions to identify objects by their coarse spatial relations. Crucially, however, participants' performance in discriminating between objects with the

same coarse spatial relations was less accurate as a function of skeletal change, $F(1, 34) = 51.77$, $p < 0.001$, $\eta_p^2 = 0.60$ (Figure 5B), suggesting that when participants make perceptual similarity judgments, they incorporate fine-grained structural information, as predicted by a skeletal model. Nevertheless, as in the previous experiment, a change to the object's shape skeleton also induced changes along other visual dimensions (e.g., image statistics). Thus, we tested whether participants' performance was better described by other models of vision. To this end, we used a random-effects regression analysis, with the skeletal similarity model and other models (i.e., GBJ, GIST, HMAX, and AlexNet-fc6) as predictors (subject and object as the random effects; see Supplemental Methods for additional details). Analyses revealed that the model of skeletal similarity remained a significant predictor of human performance, even when controlling for the other models, $\chi^2(1) = 22.30$, $p < 0.001$, and that it explained the greatest amount of variance in participants' responses, $\beta = -1.24$. The only other model to explain unique variance was GIST, $\chi^2(1) = 19.55$, $p < 0.001$, $\beta = 0.63$, suggesting a role for image-statistics in this process (see Supplemental Table 3 for the results of the other models). To ensure that the predictive power of the skeletal model was not simply the result of a suppression effect, we tested the skeletal model against every other model (Skeleton ∪ GBJ; Skeleton ∪ GIST; Skeleton ∪ HMAX; Skeleton ∪ AlexNet-fc6). Skeletal similarity was predictive of human judgments in each case, $\chi^2(1) > 9.27$, $ps < 0.002$. Taken together, these findings suggest that participants' judgments of perceived object similarity reflect the metric positions of object parts, consistent with an object representation based on skeletal structure, not the course spatial relations. Combined with Experiment 1, these results provide further support for the unique, and possibly privileged, role of skeletal descriptions of shape in object recognition.

Experiment 3 – Are skeletal structures a privileged source of information for object recognition?

A model of skeletal similarity was most predictive of human perception in Experiments 1 and 2, when compared to other, non-shape-based, models of visual similarity (i.e., GBJ, GIST, HMAX,

and AlexNet), as well as another descriptor of structure (i.e., coarse spatial relations). Together, these results suggest that shape information, particularly skeletal descriptors of shape, plays an important role in object recognition. However, there exist alternative descriptors of shape, which emphasize local contour information (Elder, 1999; Op de Beeck, Torfs, & Wagemans, 2008) or the non-accidental properties (NAPs) of component parts (Biederman, 1987), not skeletal structures. Thus, it remains unknown whether, for object recognition, skeletal structures offer a more informative descriptor of shape than local contour information and component parts.

To test this hypothesis, the skeleton of an object was pitted against its surface form in a match-to-sample task. Surface forms were designed to alter the object's contours without changing the object's underlying skeleton (Hung et al., 2012; see Figure 3B). As described in more detail below, surface form similarity was perceptually matched to skeletal similarity and surface forms were well characterized by other models of vision. Moreover, surface forms were created such that they differed in NAPs in order to compare the skeletal descriptions against a model of shape based on component parts (Amir et al., 2012; Biederman, 1987). NAPs, such as the degree to which a component tapers or bulges outward, are thought to play an essential role in models of shape perception because they serve as unique identifiers of component parts, allowing objects to be identified from a variety of viewpoints (Biederman, 1987, 2000). Thus, by pitting an object's skeleton against its surface form, we can better understand the degree to which different descriptors of shape are used for object recognition.

Surface form properties. To quantify the degree of visual similarity between surface forms, participants (n = 41) conducted a surface form discrimination task (see Supplemental Methods). In this task, participants judged whether two objects were the same or different in surface form (same skeletons). Surface form discrimination accuracy was compared to skeletal discrimination accuracy from Experiment 1 for the four skeletons used here (see Supplemental Methods and Supplemental Figure 3). This analysis revealed that surface form discrimination accuracy (M = 0.87, SD = 0.19) did

not differ from skeleton discrimination accuracy (M = 0.86, SD = 0.20), $t(77) = 0.76$, $p = 0.94$. Follow-up analyses revealed that surface form discrimination was well described by GBJ, GIST, HMAX, and AlexNet-fc6 models, rs = 0.63-0.77, with AlexNet-fc6 explaining unique variance when all four models were entered into a random-effects regression, $\chi^2(1) = 12.71$, $p < 0.001$.

To test whether surface forms were comprised of unique NAPs, a separate group of participants (n = 41) were taught to identify different NAPs and they then rated the degree to which each surface form exhibited a particular NAP (e.g., "To what extent do parts of this object exhibit taper?") on a 7-point Likert scale (1 "not at all"; 7 "a lot"; see Supplemental Methods for details). A k-means cluster analysis (Hartigan & Wong, 1979) revealed that participants' ratings were best described by four clusters, and a permutation test, in which cluster labels were shuffled 10,000 times, revealed that cluster labels were predictive of each surface form significantly better than chance ($ps < 0.002$). That the surface forms were better described by four, rather than five (one for each surface form), clusters is consistent with two of the surface forms having the same NAPs, but differing in metric properties such as circumference (see Figure 3B; Amir et al., 2012; Vogels, Biederman, Bar, & Lorincz, 2001).

Match-to-sample task: design. Having confirmed that the surface forms were perceptually matched to skeletal differences, and that they were comprised of unique NAPs, we were in a position to test whether skeletal descriptions of shape were a privileged source of information for object recognition relative to other descriptors of shape. In a match-to-sample task, a separate group of participants (n = 39) were presented with a sample object and two choice objects (i.e., target and distractor; see Figure 6A-C). They were instructed to judge which of the two choice objects was most likely to be from the same category as the sample object. The target object matched the sample object in either its skeleton or surface form. The distractor object differed from the sample object by both skeleton and surface form. These trials ensured that participants were able to match objects by either their skeleton or surface form when each property was presented in

isolation (see Supplemental Figure 5A-B). Other trials presented a conflict between the object's skeleton and surface form such that one of the choice objects matched the sample's skeleton, but not surface form, and the other object matched the object's surface form, but not skeleton (see Figure 6C). The conflict trials tested whether the skeletal descriptors served as a preferred cue for object recognition. The objects were presented as still images in one of three depth orientations (30°, 60°, 90°; see Supplemental Figure 3). Participants were instructed to ignore the orientations of the objects and, on each trial, to choose which of the two choice objects was from the same category as the sample object.

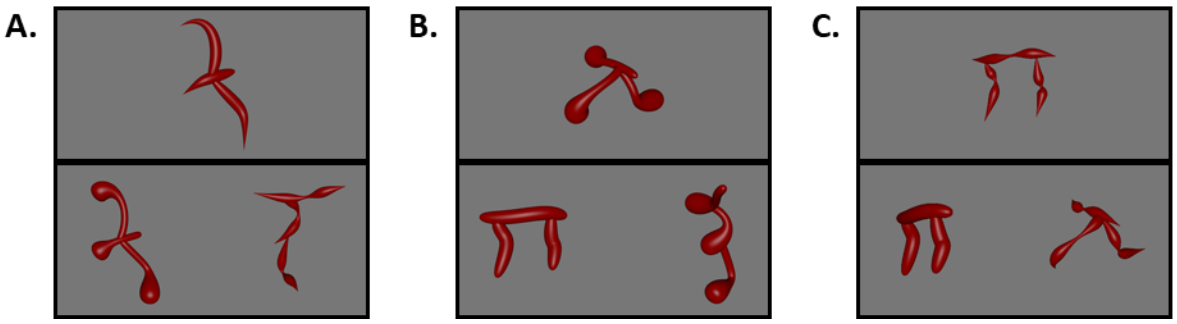


Figure 6. Examples of the three trial types used in Experiment 3. (A) A skeleton match trial wherein one choice object matched the sample's skeleton, but not surface form. The other choice object matched on neither skeleton nor surface form. (B) A surface form match trial wherein one choice object matched the sample's surface form, but not skeleton. The other choice object matched on neither skeleton nor surface form. (C) A conflict trial wherein one choice object matched the sample's skeleton, but not surface form, and the other choice object matched the sample's surface form, but not skeleton.

Match-to-sample task: results and discussion. Participants successfully categorized objects by either their skeletons, $M = 0.88$ ($M_{RT} = 1167$ ms), $t(38) = 27.01$, $p < 0.001$, $d = 4.32$, 95% CI [3.35, 5.41], or surface forms, $M = 0.78$ ($M_{RT} = 1419$ ms), $ts(38) = 15.51$, $p < 0.001$, $d = 2.48$, 95% CI [1.87, 3.16], when each cue was presented in isolation, as indicated by their above chance performance in these conditions (see Figure 7A). Crucially, however, on the conflict trials, participants categorized objects by their skeletons, not surface forms, $t(38) = 6.63$, $p < 0.001$, $d =$

1.06, 95% CI [0.66, 1.45] (see Figure 7B-C). Indeed, participants preferentially categorized objects by their skeletons when pitted against all, $M > 0.61$ ($M_{RT} < 1218$ ms), $ts[38] > 2.52$, $ps < .016$, $ds > 0.40$), but one, $M = 0.58$ ($M_{RT} = 1140$ ms) ($p = .059$, $d = 0.31$) surface form. Thus, although surface forms were perceptually matched to the objects' skeletons and were comprised of unique NAPs, participants relied more heavily on the shape skeleton when classifying objects, suggesting that skeletal structure may be a privileged source of shape information for object recognition.

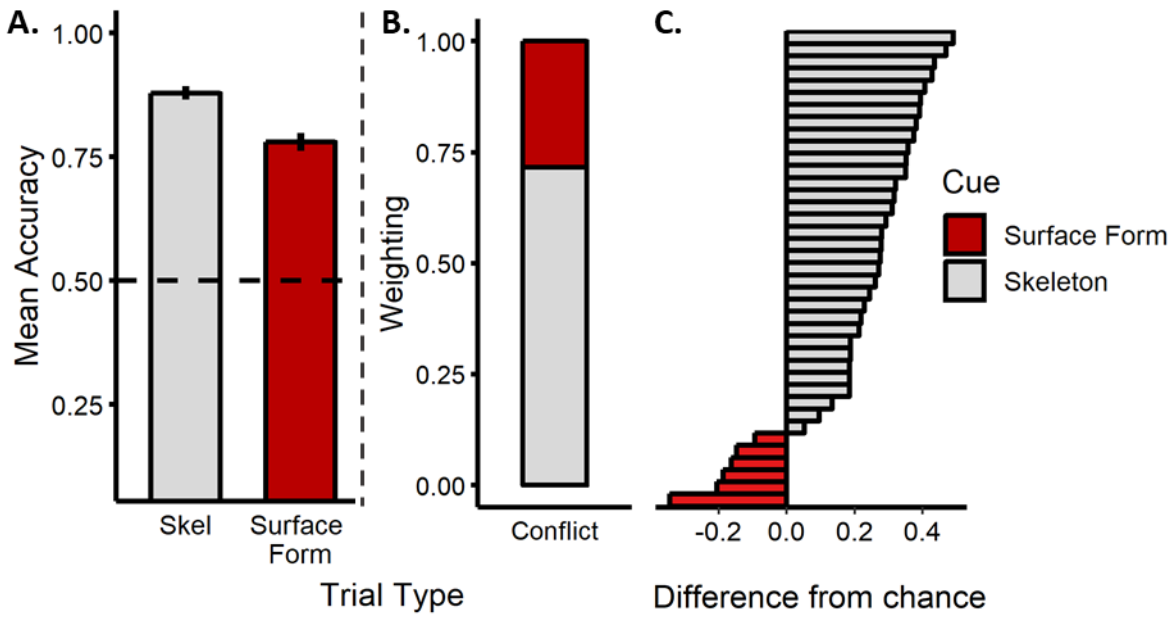


Figure 7. Results from the match-to-sample task of Experiment 3. (A) Participants' mean accuracy (error bars represent SE) on trials in which only a skeleton or surface form match was possible (dotted line indicates chance performance). (B) Participants' categorization judgments in the conflict trial. A value closer to 1 indicates greater weighting of the object's skeleton; a value closer to 0 indicates greater weighting of the object's surface form. Although participants successfully matched objects by their skeletal structure or surface forms when each cue was presented in isolation, they were more likely to match objects by their skeleton, as opposed to their surface forms, when these cues conflicted with one another. (C) Histogram of participants' responses on the conflict trials. A value greater than zero indicates greater weighting of skeletal information. The majority of participants matched objects by their skeleton, demonstrating a consistent pattern of responses across participants.

Discussion

The ability to determine the similarity between shapes is crucial for object recognition. Shape skeletons may be a particularly useful representation in this context because they provide a compact descriptor of shape, as well as a common format from which to compute shape similarity. Nevertheless, few models of biological object recognition include skeletal descriptions in their implementation. Here we tested whether skeletal structures provide an important source of information for object recognition when compared with other models of vision. Our results showed that a model of skeletal similarity was most predictive of human object judgments when contrasted with models based on image-statistics or neural networks, as well another model of structure based on coarse spatial relations. Moreover, we found that skeletal structure was a privileged source of information when compared to other properties thought to be important for shape perception, such as object contours and component parts. Thus, our results suggest that not only does the visual system show sensitivity to the skeletal structure of objects (Hung et al., 2012; Kovács et al., 1998; Lowet et al., 2018), but also that perception and comparison of object skeletons may be crucial for successful object recognition.

The strength of skeletal models is that they provide a compact description of an object's global shape structure, as well as a metric by which to determine shape similarity. Indeed, shape skeletons may offer a concrete formalization of the oft poorly defined concept of global shape. Skeletal descriptions exhibit many properties ascribed to global shape percepts such as relative invariance to local contour variations (Ayzenberg, Chen, et al., 2019; Shaked & Bruckstein, 1998). Moreover, there exist many methods by which to compare skeletal structures, such as by their hierarchical organization (Shokoufandeh, Macrini, Dickinson, Siddiqi, & Zucker, 2005) or using distance metrics (as used here; Sebastian et al., 2004), thereby allowing for a quantitative description of shape similarity. Such a description may be particularly important when recognizing objects across previously unseen views or categorizing novel object exemplars.

A question that arises from the current findings is the extent to which the stimuli and tasks used here invoke the same mechanisms as rapid real-world object perception, also known as 'core' object recognition (Rajalingham et al., 2018). Indeed, one might ask whether the tightly controlled stimulus set, which was designed to vary in skeletal similarity, and the untimed tasks, where participants could directly compare the similarity of objects, invoke 'core' object recognition processes. It is well known that the visual system receives input from multiple systems (e.g., frontal and parietal regions) and incorporates recurrent processes to solve object recognition, particularly in cases of uncertainty (Bar et al., 2006; Tang et al., 2018; Van Dromme, Premereur, Verhoef, Vanduffel, & Janssen, 2016). Thus, it is possible that object recognition tasks in this study, and the implementation of skeletal models more generally, may have invoked higher-level processes. Although we acknowledge that object recognition is not a unitary process, with higher-level processes playing an important role, we suggest that our stimuli and tasks likely measured core object recognition. In particular, in Experiment 1, we found that the objects could be discriminated by models that are implemented during a feedforward sweep through the ventral visual stream, and these models were also predictive of human judgments. Moreover, we found that participants performed equally well when objects were presented for only 100 ms (see Supplemental Experiment 1). However, it is an open question whether shape skeletons are implemented using exclusively feedforward mechanisms or whether recurrent or generative processes are also needed (Ardila, Mihalas, Heydt, & Niebur, 2012; Elder, 2018; Trinh & Kimia, 2007). Nevertheless, our work highlights the importance of formalized models of shape for object recognition, particularly the unique, and possibly privileged, role that skeletal structures may play.

Supplementary Materials for "Skeletal descriptions of shape provide unique perceptual information for object recognition"

Supplemental Methods

Participants. 205 participants were tested in the current study (M_{age} = 19.75 years; *range* = 18.03 – 23.48 years). Of these participants, 16 were excluded for exhibiting chance, or below chance, performance (3 from Experiment 1; 7 from Experiment 2; 3 from the surface form discrimination task of Experiment 3). Because accuracy could not be evaluated in the NAP rating task of Experiment 3, we ensured that participants exhibited reliable performance; 6 participants were excluded from this experiment for failing to meet this criterion ($\alpha s < 0.7$). All participants provided informed consent and participated for course credit. Experimental procedures were approved by Emory University's Institutional Review Board (IRB). All experiments were performed in accordance with the relevant guidelines and regulations of the IRB.

Apparatus. All tasks were presented on a desktop computer with a 19-inch screen (1280 × 1024 px) and controlled using custom software written in Visual Basic (Microsoft). Participants sat at a distance of ~60 cm from the computer screen.

Experiment 1

Stimulus generation. Objects were procedurally generated using the Python API for Blender (Blender Foundation). Each skeleton was comprised of three segments created from Bezier curves of a random size and curvature scaled between .05 and .25 virtual Blender units (vu). The first axis segment was oriented forward towards the 'camera'. The second and third segments were oriented perpendicular to the first segment and attached to the first segment or second segment at a random point along their length. Surface forms were created by applying a circular bevel to the object's skeleton along with one of five taper properties that determined the shape of the surface form. Finally, the overall size of the object was normalized to .25 vu.

Skeletal similarity model. The coordinates of the skeleton for each object were extracted by sampling 999 points along the length of each axis segment (2997 points in total). Skeletal points were normalized by the length of each segment by subsampling points along the skeletal structure until these points were evenly spaced across the skeleton by 0.0005 vu (for ease of analysis, coordinates were rescaled by a factor of 1000). Skeletal similarity was calculated as the mean Euclidean distance between each point on one skeleton structure with the closest point on the second skeleton structure following maximal alignment. Maximal alignment was achieved by overlaying each structure by its center of mass and then iteratively rotating each object in the picture plane orientation by 15° until the smallest distance between structures was found.

Gabor-jet (GBJ) model. The GBJ model is a low-level model of image similarity inspired by the response profile of complex cells in early visual cortex (Margalit et al., 2016). It has been shown to scale with human psychophysical dissimilarity judgments of faces and simple objects (Yue, Biederman, Mangini, von der Malsburg, & Amir, 2012). To simulate the response profile of complex cell responses, the model overlays a 12 × 12 grid of Gabor filters (5 scales × 8 orientations) along the image. The image is convolved with each filter, and the magnitude and phase of the filtered image is stored as a feature vector. Dissimilarity between each image is computed as the mean Euclidean distance between feature vectors of each image. A single dissimilarity value was computed for each object pair by taking the mean Gabor activation distance for an object pair across the three orientations (30 objects × 3 orientations).

GIST. The GIST model is considered a mid-level model of image similarity that describes the content of an image through global image features (Oliva & Torralba, 2001). It has been shown to accurately describe the content of natural images, particularly as they relate to scene perception (Oliva & Torralba, 2006). The model overlays a grid of Gabor filters (4 scales × 8 orientations) on the image and then convolves the image with the filters, creating a feature activation map. This feature map is divided into 16 regions (based on the 4 × 4 grid) and then mean activation within

each region is computed and stored as a GIST feature vector. GIST dissimilarity between each image is computed as the mean Euclidean distance between feature vectors of each image. A single dissimilarity value was computed for each object pair by taking the mean Euclidean distance for an object pair across the three orientations (30 objects × 3 orientations).

HMAX. The HMAX (hierarchical MAX) model is a biologically inspired hierarchical neural network model that describes an image by max-pooling over a series of simple (S1, S2) and complex (C1, C2) units (Serre, Oliva, et al., 2007; Serre, Wolf, et al., 2007). It has been shown to match human performance on simple category judgment tasks (e.g., animals and non-animals) and exhibits some invariance to changes in position and scale. In the current study, we used the feature patches provided with the HMAX model. In the first layer (S1), each image is convolved with Gabor filters (8 scales × 4 orientations), the output of which is fed into a second layer (C1) that determines the local maximum over all positions and scales. The outputs of these layers are fed through a second set of simple and complex units (S2, C2). Dissimilarity was computed by extracting the C2 activations for each image and correlating (Pearson) it with the activations for every other image. A single dissimilarity value was computed for each object pair by taking the mean correlation for an object pair across the three orientations (30 objects × 3 orientations).

CNN. As a model of high-level vision, we used AlexNet, an eight layer CNN pre-trained to classify objects from the ImageNet database (Krizhevsky et al., 2012; Russakovsky et al., 2015). We adopted AlexNet rather than other CNNs in our analyses because its architecture is relatively simple by comparison and because it can identify objects with high accuracy. Importantly, AlexNet has been shown to be predictive of human object similarity judgments (Jozwik et al., 2017). CNN similarity for each object was calculated by extracting a feature vector from each convolutional, and fully connected, layer for each object image, and then computing the mean Euclidean distance between the feature vector for each image with every other image. A single difference value was computed for each object pair by taking the mean CNN difference for an object pair across the three

orientations (30 objects × 3 orientations). Dissimilarity values were calculated for every layer and correlated to participants' behavioral judgments. Participants' similarity judgments in Experiment 1 were most strongly correlated with fully-connected layer 6 (fc6) of AlexNet, $r = 0.34$ (see Figure 4A). Because fc6 was most predictive in this experiment, we tested only fc6 in subsequent experiments.

Discrimination task. On each trial, participants were shown images of two objects (side-by-side) presented simultaneously in one of three depth orientations (-30°, 0°, +30°). Objects were matched for surface form and either had the same or different skeleton. Participants were instructed to decide whether the two images showed the same or different object (independent of orientation). Each object was paired with every other object (including itself) during the experimental session. Participants were administered a total of 885 trials (435 different and 450 same trials). Each trial began with a fixation cross (500 ms), followed by a pair of objects, which remained onscreen until a response was made, followed by an inter-trial interval (500 ms). Each object was approximately 6° × 6° in size and subtended 9° from the center of the screen.

Experiment 2

Stimulus generation and model analyses. A *k*-means cluster analysis was conducted on participants' discrimination data from Experiment 1. This analysis revealed that objects were well described by four clusters. Based on these clusters, three perceptually matched objects were chosen whose skeletons could be altered without changing the coarse spatial relations (see Supplemental Figure 2). Importantly, these objects also had different coarse spatial relations (Set 1: two components below a third, pointing down, one component placed in front of the other; Set 2: one component on either side of a third, one pointing up and the other down; Set 3: one component on either side of a third, each pointing diagonally down). Six versions of each object (0%, 10%, 20%, 30%, 40%, and 50% skeleton difference) were generated by moving one segment along the length of the central segment in 10% increments (relative to the length of the central component). Objects

were rendered with only the thinnest surface form to prevent component parts from overlapping. Images of each object (18 total) were generated in three orientations (30°, 60°, 90°) intended to maximize the view of each object. Each object was analyzed and compared with every other object using the same models and procedure described in Experiment 1.

Discrimination task. On each trial, participants were shown images of two objects (side-by-side) presented simultaneously in one of the three depth orientations (30°, 60°, 90°). Each object was rendered with the same surface form (see Figure 5A) and either had the same or different coarse spatial relations. On each 'same' trial, participants were presented with both a reference object (0% skeletal difference) and another object that had the same coarse spatial relations but different skeleton in one of the increments described previously (objects in the same columns of Figure 5A). On each 'different' trial, participants were presented with two objects that had different coarse spatial relations (any possible skeleton; objects in the same rows of Figure 5A). Participants were instructed to decide whether the two images showed an object with the same or different "overall shape" (independent of orientation). Participants were given instructions and 8 sample trials (with feedback) using a separate set of objects to ensure that they understood that "overall shape" referred to objects with the same coarse spatial relations (4 same trials; 4 different trials). In the same trials, each skeletal difference was presented an equal number of times in each possible orientation. In the different trials, object pairs with different coarse spatial relations (any possible skeleton) were randomly selected and presented in randomly determined orientations. Participants were administered at total of 648 trials (324 same trials and 324 different trials). Each trial began with a fixation cross (500 ms), followed by a pair of objects, which remained onscreen until a response was made, followed by an inter-trial interval (500 ms). Each object was approximately 6° × 6° in size and subtended 9° from the center of the screen.

Experiment 3

Stimulus generation and model analyses. Four perceptually matched objects were chosen from the object clusters identified in Experiment 2. Images of each object (4 objects × 5 surface forms) were generated in three orientations (30°, 60°, 90°) intended to maximize the view of each object (see Supplemental Figure 2).

Surface form discrimination task. On each trial, participants were shown images of two objects (side-by-side) in one of the three depth orientations (30°, 60°, 90°). Objects had the same shape skeleton and either the same or different surface forms. Participants were instructed to decide whether the two images showed the same or different object (independent of orientation). Each surface form was paired with every other surface form an equal number of times for a total of 600 trials.

NAP task. To test whether surface forms were comprised of unique component parts, participants rated each surface form on the degree to which it exhibited a specific NAP. During a training phase, participants were taught a subset of NAPs (drawn from Amir et al. (2012)) and then shown a subset of objects that they were asked to rate on the degree to which they exhibited the specific NAP. The four NAPs were: (1) *taper,* defined as the degree to which the thickness of an object was reduced towards the end ((taper in the current study corresponds to 'expansion of cross-section' in Amir et al.Amir et al., 2012)); (2) *positive curvature*, defined as the degree to which an object part curved outwards; (3) *negative curvature,* defined as the degree to which an object part curved inwards; and (4) *convergence to vertex,* defined as the degree to which an object part ended in a point. We excluded the curved versus straight axis property of Amir et al. (Amir et al., 2012) because it was confounded with the object's skeleton. We also excluded the change in cross-section property (e.g., circular vs. rectangular shape) because all of the surface forms had a circular cross-section. Participants were tested on their understanding of the four NAPs with a task in which they were presented with pairs of single-part objects (simultaneously onscreen) where one exhibited an NAP and the other did not. Participants were instructed to select the object that

exhibited more/less of a particular NAP (e.g., "Which object exhibits more positive curvature?"; feedback was provided). During the rating phase, participants were shown each test stimulus with each surface form (4 objects × 5 surface forms) and asked to rate the degree to which each surface form exhibited a particular NAP (e.g., "To what extent do parts of this object exhibit taper?") on a 7-point Likert scale (1 = "not at all"; 7 = "a lot").

Match-to-sample task. On each trial, participants were shown one object (sample) placed centrally near the top of the screen above two objects near the bottom of the screen (target and distractor). Participants were instructed to choose which of the two bottom objects was most likely to be in the same category as the sample. Participants were presented with three possible trial types: skeleton and surface form trials, in which one object matched the sample in either skeleton or surface form, respectively (the other object matched on neither; see Figure 6A-B); and conflict trials in which one object matched in skeleton, but not surface form, and the other object matched in surface form, but not skeleton (see Figure 6C). Participants were administered a total of 480 trials (160 of each trial type). Each trial began with a fixation cross (500 ms), followed by the sample and choice objects, which remained onscreen until a response was made, followed by an inter-trial interval (500 ms). Each stimulus was approximately 6° × 6° in size, and choice objects subtended 9° from the center of the screen.

Supplemental Experiment 1

One potential concern with our stimuli is that they may have been difficult to discriminate because they represent a single class of unfamiliar objects with a high degree of visual similarity. Such objects may be less likely to elicit the same mechanisms that support 'core' object recognition, which is thought to occur within 100-200 ms via a feedforward sweep through the ventral stream (Rajalingham et al., 2018). Instead, discrimination of these objects may elicit additional high-level processes, such as mental rotation, not typically implemented when discriminating familiar objects (Gauthier et al., 2002). This possibility is difficult to rule out in the main experiments because

participants were given unlimited time to discriminate between objects. Thus, in a supplemental experiment, we tested participants in a speeded task where the target object was presented for a 100 ms. If performance on this speeded task were comparable to performance on the unspeeded task (see Experiment 1 in the main text), then it would suggest that both tasks measure core object recognition.

Participants (n = 14) were administered a sequential match-to-sample task where they were asked to decide which of two choice objects matched a previously presented sample object. Each trial began with a fixation cross (500 ms), followed by a display with the sample object (100 ms), and then a display with two choice objects which remained onscreen until a response was made. One choice object had the same skeleton as the sample, and the other choice object had a different skeleton. The choice objects always had the same surface form as the sample (randomly selected) but were presented from different orientations (-30°, 0°, 30°). Participants were instructed to ignore the orientations of the objects and to make their decision on the basis of visual similarity. Each object was pitted against every other object an equal number of times (435 trials). Each object was approximately 6° × 6° in size, and choice objects subtended 9° from the center of the screen.

Comparisons to chance (0.50) revealed that participants were able to match the sample object with the correct choice object, M = 0.82% (M_{RT} = 946 ms), $t(13) = 26.8$, $p < .001$, $d = 7.16$, with 14/14 participants displaying accuracy above 0.74. This result suggests that our objects differed sufficiently to allow object recognition to occur within 100 ms.

In a subsequent analysis, we tested whether participants' performance on this task differed from their performance in the unspeeded discrimination task used in Experiment 1. We found that participants performed comparably in the two tasks (M_{accuracy} = 0.82 vs. 0.80), with no statistical difference between groups (p = 0.46). These results are consistent with the speeded and unspeeded tasks recruiting similar perceptual processes, namely 'core' object recognition.

Supplemental Table 1. Linear regression results for each model used in Experiment 1.

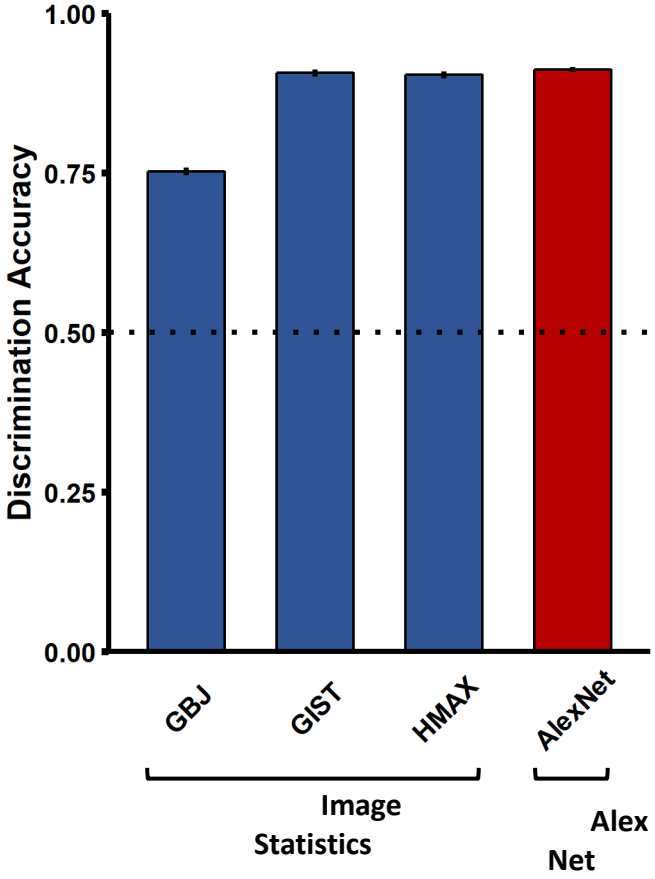
Model	Standardized Coefficient		
	β	*t*	*p*
(Constant)		1.82	
Skeleton	0.27	6.01	< .001
Gabor-Jet	0.31	3.17	0.002
GIST	-0.18	-2.03	0.043
HMAX	0.11	2.06	0.040
AlexNet-fc6	0.13	1.80	0.073

Supplemental Table 2. Coefficients displaying the percentage of unique and shared variance explained by each model and model combinations, as well as percentages of the total explainable variance (20.5%) explained by each model and model combinations.

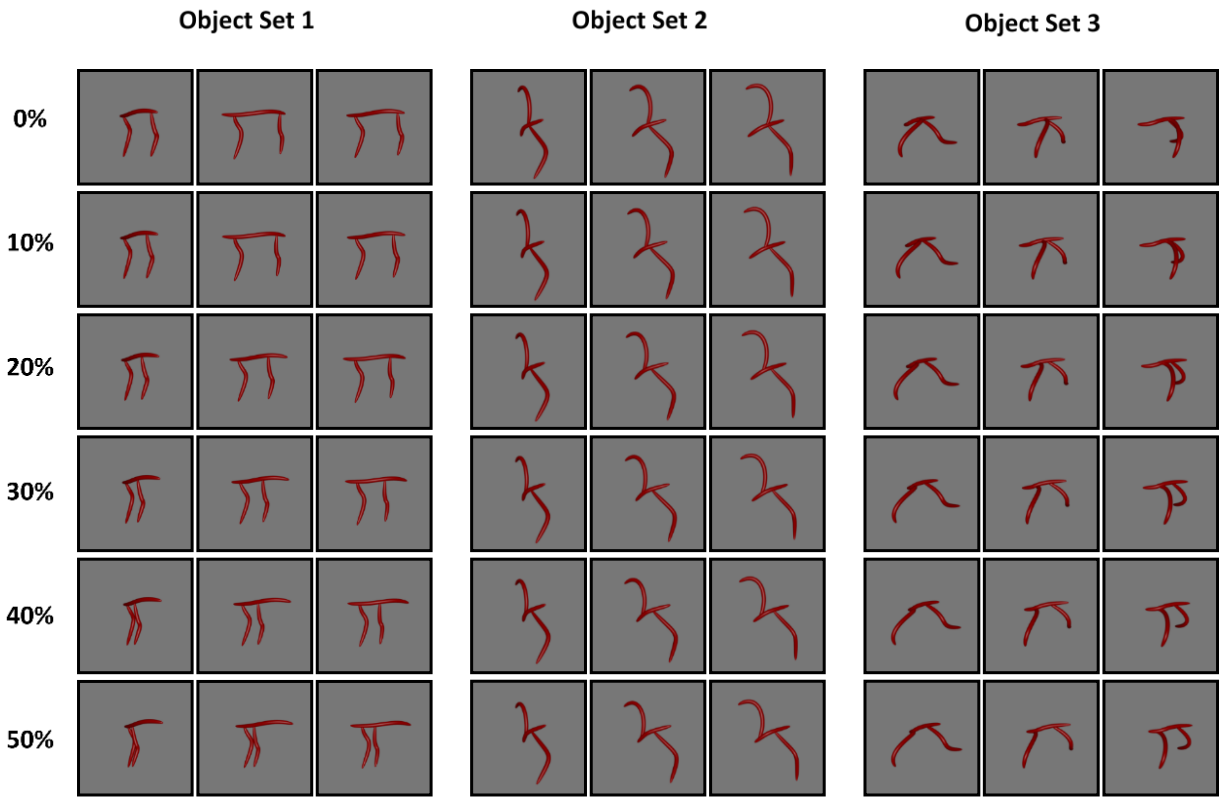
Model	Coefficient	Percentage of total
SKEL	6.631	33.305
GBJ	1.726	8.669
GIST	0.587	2.949
HMAX	0.732	3.678
fc6	0.499	2.508
SKEL + GBJ	-0.641	-3.218
SKEL + GIST	-0.241	-1.209
GBJ + GIST	-0.512	-2.569
SKEL + HMAX	-0.171	-0.859
GBJ + HMAX	0.344	1.730
GIST + HMAX	0.037	0.187
SKEL + fc6	1.005	5.046
GBJ + fc6	1.218	6.116
GIST + fc6	-0.105	-0.525
HMAX + fc6	0.445	2.235
SKEL + GBJ + GIST	0.224	1.125
SKEL + GBJ + HMAX	-0.095	-0.477
SKEL + GIST + HMAX	-0.009	-0.047
GBJ + GIST + HMAX	0.032	0.162
SKEL + GBJ + fc6	0.299	1.502
SKEL + GIST + fc6	-0.047	-0.237
GBJ + GIST + fc6	1.852	9.302
SKEL + HMAX + fc6	0.236	1.187
GBJ + HMAX + fc6	1.017	5.109
GIST + HMAX + fc6	-0.039	-0.196
SKEL + GBJ + GIST + HMAX	-0.006	-0.032
SKEL + GBJ + GIST + fc6	1.001	5.027
SKEL + GBJ + HMAX + fc6	0.037	0.186
SKEL + GIST + HMAX + fc6	0.009	0.045
GBJ + GIST + HMAX + fc6	3.084	15.493
SKEL + GBJ + GIST + HMAX + fc6	0.758	3.809

Supplemental Table 3. Random-effects regression results for each model used in Experiment 2. The standardized coefficients and *t*-values are drawn from a full regression model that included each model as a predictor. The χ^2 and *p*-values were calculated by iteratively testing the full regression model against ones without the predictor of interest.

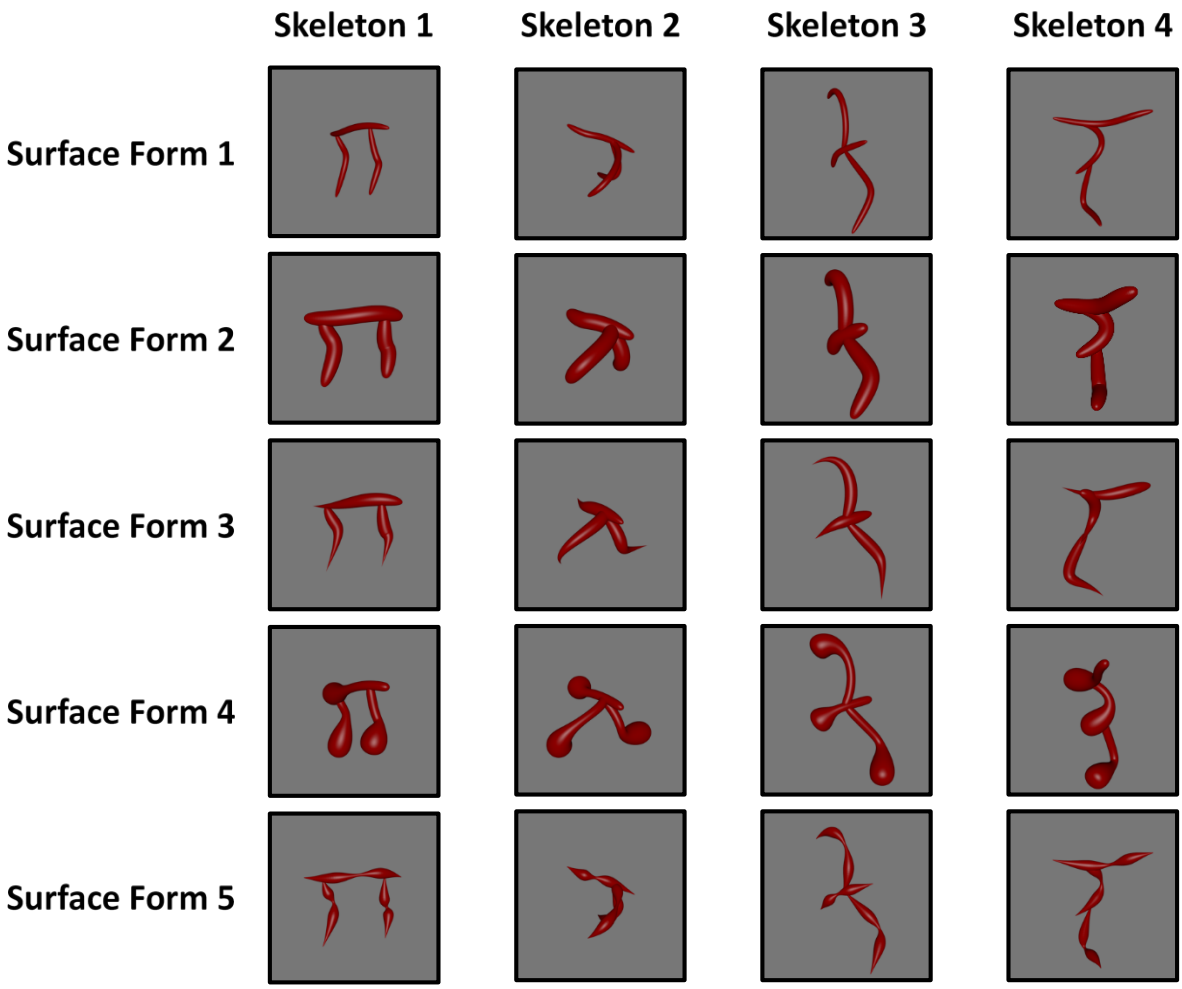
Model	Standardized Coefficient			
	β	*t*	χ^2	*p*
(Intercept)		22.51		
Skeleton	-1.24	-4.80	22.30	< 0.001
Gabor-Jet	0.32	1.44	1.81	0.18
GIST	0.63	4.46	19.55	< 0.001
HMAX	-0.07	-0.50	0.19	0.66
AlexNet-fc6	0.06	0.21	0.05	0.83



Supplemental Figure 1. Discrimination accuracies for all non-skeletal models. Each model was able to discriminate between objects significantly above chance (dotted line).



Supplemental Figure 2. All stimuli used in Experiment 2. Objects were comprised of three sets, each with distinct coarse spatial relations. Within each set, objects varied in skeletal similarity by increments of 0%, 10%, 20%, 30%, 40%, or 50% (each row). Each object could be presented in one of three orientations, each of which is depicted here (30°, 60°, 90°; each column within an object set).



Supplemental Figure 3. The stimulus set used in Experiment 3. Each column displays objects with the same skeleton, but different surface forms. Each row displays objects with the same surface form, but different skeletons. Each object could be presented in one of three orientations (30°, 60°, 90°); a subset are depicted here.

Chapter 3 - A dual role for shape skeletons in human vision: Perceptual organization and object recognition (Ayzenberg, Kamps, Dilks, & Lourenco, 2019)

A central goal of vision science is to understand how the human visual system represents the shapes of objects and how shape is ultimately used to recognize objects. Research from computer vision has suggested that shape representations can be created and then compared using computational models based on the medial axis, also known as the "shape skeleton." Although recent behavioral studies suggest that humans also represent shape skeletons (Ayzenberg & Lourenco, 2019b; Firestone & Scholl, 2014), it remains unknown whether they contribute to perceptual organization, object recognition, or both. Here we provide important neural evidence that shape skeletons may be involved in both functions.

One method to address whether shape skeletons are implicated in both perceptual organization and object recognition is to test whether regions of the brain involved in these processes also represent the shape skeleton, without an explicit task. If shape skeletons are used to create shape percepts, then they should be represented in area V3, a region consistently found to be involved in perceptual organization in humans (for review, see Sasaki, 2007). Indeed, V3 is the earliest stage of the visual hierarchy where symmetry structure has been decoded (Keefe et al., 2018; Sasaki, Vanduffel, Knutsen, Tyler, & Tootell, 2005; Van Meel, Baeck, Gillebert, Wagemans, & Op de Beeck, 2019). Moreover, it has been implicated in forming shape percepts from illusory contours (McMains & Kastner, 2010; Montaser-Kouhsari, Landy, Heeger, & Larsson, 2007) and motion (Caplovitz, Barroso, Hsieh, & Tse, 2008; Caplovitz & Peter, 2010). If shape skeletons are also used to recognize objects, then they should be represented in the lateral occipital cortex (LO), a region that is particularly sensitive to the spatial arrangement of object parts (Behrmann, Peterson, Moscovitch, & Suzuki, 2006; Margalit, Biederman, Tjan, & Shah, 2017), and known to be important for object recognition (Freud, Culham, Plaut, & Behrmann, 2017; Grill-Spector, Kourtzi, & Kanwisher, 2001; Grill-Spector, Kushnir, Edelman, Itzchak, & Malach, 1998),

Preliminary evidence for this hypothesis comes from neuroimaging studies with humans and monkeys. Using fMRI with humans, Lescroart and Biederman (2012) decoded the axis structure of objects in both V3 and LO. With monkeys, electrophysiological studies have demonstrated the existence of neurons in inferior temporal cortex (IT), the putative homolog of LO, which selectively code for object skeletons independently of their surface characteristics (Hung et al., 2012). Although this last study did not examine early visual regions, these two studies together are consistent with a role for shape skeletons in perceptual organization and object recognition. However, a major limitation of these studies is that they did not measure skeletal coding directly, nor did they compare skeletons to other models of vision. More specifically, these studies measured how neural populations changed in response to different skeletons. However, changes to object skeletons also induce changes along other visual dimensions. Because these studies did not measure *how much* an object's skeleton changes relative to other visual properties, it is impossible to know whether their results reflect skeletal coding or some other model of vision.

In the current study, we directly measured skeletal coding by varying object skeletons parametrically and then examining the unique contributions of skeletal information to neural responses. More specifically, we used representational similarity analysis (RSA) to test whether a model of skeletal similarity predicted the response patterns in *both* V3 and LO. Importantly, we also examined the specificity of the shape skeleton in these regions by controlling for other models of visual similarity that approximate early- (i.e., Gabor-jet; Margalit et al., 2016), mid- (i.e., GIST, and HMAX; Oliva & Torralba, 2006; Serre, Oliva, & Poggio, 2007), and high- (i.e., AlexNet-fc6; Krizhevsky et al., 2012) level visual processing. Finally, we ensured that the representation of shape skeletons could not be accounted for by lower-level shape properties (i.e., contours) by directly manipulating the object's contours while keeping the skeleton intact.

Materials and Methods

Participants

Twenty participants (M_{age} = 19.29 years, range = 20 – 36 years; 8 females) were recruited from the Emory University community. All participants gave written informed consent to participate and had normal or corrected-to-normal vision.

Stimuli

Twelve novel objects were selected from the stimulus set created by Ayzenberg and Lourenco (2019b, see Study 1 of current dissertation; see Figure 8A). The selected object set was composed of six distinct skeletons and two surface forms. The six skeletons were chosen by first conducting a *k*-means cluster analysis (k = 3) on skeletal similarity data for 30 unique objects (for details, see Ayzenberg & Lourenco, 2019). We selected six objects whose within- and between-cluster skeletal similarities were matched (2 per cluster). That is, the two objects from the same cluster were approximately as similar to one another as the two objects within the other clusters; objects in different clusters had comparable levels of dissimilarity to one another (see Figure 8B). This method of stimulus selection ensured that the stimulus set used in the present study contained objects with both similar and dissimilar skeletons. Importantly, to ensure that our stimulus selection criteria did not result in skeletons being the only salient visual feature, we also tested whether other, non-skeletal, models could discriminate between these objects. This analysis revealed that all models could accurately discriminate skeletons (80.2% - 95.3% accuracy; see Supplemental Materials for more details).

Each skeleton was rendered with one of two surface forms, which changed the contours and component parts of the object without altering the underlying skeleton. To provide the strongest test of a skeletal model, we chose the two surface forms (out of five) that a separate group of participants judged to be most dissimilar (Ayzenberg & Lourenco, 2019b). Importantly, the surface forms also had qualitatively different component parts, as measured by their non-accidental properties NAPs (Amir et al., 2012; Biederman, 1987), and differed in their image-level properties

(Margalit et al., 2016). Two additional objects were used as targets for an orthogonal target-detection task; these objects were not included in subsequent analyses.

We used a smaller stimulus set (compared to Ayzenberg & Lourenco, 2019b) to increase the number of presentations per objects and maximize the signal-to-noise ratio. Furthermore, using a smaller stimulus set allowed us to implement a continuous carry-over design with third-order counterbalancing, thereby minimizing carry-over effects across trials (see Methods; Aguirre, Mattar, & Magis-Weinberg, 2011). Stimulus presentation was controlled by a MacBook Pro running the Psychophysics Toolbox package (Brainard, 1997) in MATLAB (MathWorks). Images were projected onto a screen and viewed through a mirror mounted on the head coil.

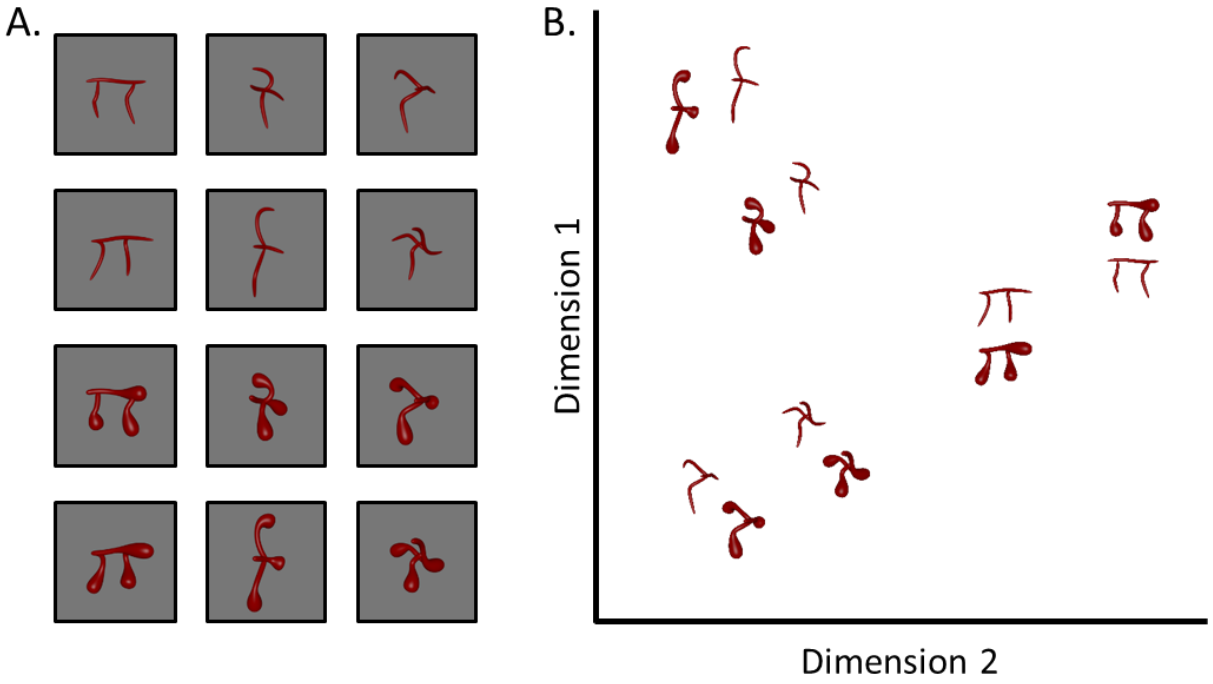


Figure 8. Stimuli used in the current study and a multi-dimensional scaling (MDS) plot illustrating the skeletal similarity between objects. (A) Six objects with unique skeletal structures were generated. Each object was rendered with two surface forms to change the objects' component parts without disrupting the skeleton. (B) To ensure that the stimulus set contained objects with both similar and dissimilar skeletons, objects were selected in pairs such that within- and between-pair skeletal similarity were approximately matched across objects.

Experimental design

First, we used a region of interest (ROI) approach, in which we independently localized the ROIs (localizer runs). Second, we used an independent set of data (experimental runs) to conduct representational similarity analyses in each ROI.

Localizer runs. We used a block design for the localizer runs. Participants viewed images of faces, bodies, objects, scenes, and scrambled objects, as previously described (Dilks, Julian, Kubilius, Spelke, & Kanwisher, 2011). Each participant completed three localizer runs, comprised of four blocks per stimulus category, each 400 s. Block order in each run was randomized. Each block contained 20 images randomly drawn from the same category. Each image was presented for 300 ms, followed by a 500 ms interstimulus interval (ISI), for a total of 16 s per block. We also included five 16 s fixation blocks: one at the beginning, three in the middle interleaved between each set of stimulus blocks, and one at the end of each run. To maintain attention, participants performed an orthogonal one-back task, responding to the repetition of an image on consecutive presentations.

Experimental runs. We used a continuous carry-over design for the experimental runs, wherein participants viewed images of each novel object. Each run was 360 s long. Using a de Bruijn sequence (Aguirre et al., 2011), we applied third-level counterbalancing on the image presentation order, which minimized any carry-over effects between stimuli. Importantly, this design supports smaller inter-stimulus intervals (ISIs) between stimuli (Aguirre et al., 2011; Drucker & Aguirre, 2009; Hatfield, McCloskey, & Park, 2016) and allowed for a greater number of presentations per image. Each image was presented for 600 ms, followed by a 200 ms ISI, and shown 225 times across the entire session. Each run began and ended with 6 s of fixation. To maintain attention, participants performed an orthogonal target-detection task. At the beginning of each experimental run, participants were shown one of two objects (not included in subsequent analyses) and were instructed to press a response button any time the target object appeared within the image stream.

MRI scan parameters

Scanning was done on a 3T Siemens Trio scanner at the Facility for Education and Research in Neuroscience (FERN) at Emory University. Functional images were acquired using a 32-channel head matrix coil and a gradient echo single-shot echoplanar imaging sequence. Thirty slices were acquired for both localizer and experimental runs. For all runs: repetition time = 2 s; echo time = 30 ms; flip angle = 90°; voxel size = 1.8 × 1.8 × 1.8 mm with a 0.2 mm interslice gap. Slices were oriented approximately parallel to the anterior and posterior cingulate, covering the occipital and temporal lobes. Whole-brain, high-resolution T1-weighted anatomical images (repetition time = 1900 ms; echo time = 2.27 ms; inversion time = 900 ms; voxel size = 1 × 1 ×1 mm) were also acquired for each participant for registration of the functional images. Analyses of the fMRI data were conducted using FSL software (Smith et al., 2004) and custom MATLAB code.

Data Analysis

Images were skull-stripped (Smith, 2002) and registered to participants' T1 weighted anatomical image (Jenkinson et al., 2002). Prior to statistical analyses, images were motion corrected, de-trended, and intensity normalized. Localizer, but not experimental, data were spatially smoothed (6 mm kernel). All data were fit with a general linear model consisting of covariates that were convolved with a double-gamma function to approximate the hemodynamic response function.

To investigate whether skeletal descriptions of shape play a role in the creation of shape percepts, we defined V3 bilaterally, using probabilistic parcels (Wang, Mruczek, Arcaro, & Kastner, 2014). As control regions, we also defined V1, V2, and V4 bilaterally from the same set of probabilistic parcels. Each parcel was registered from MNI standard space to participants' individual anatomical space.

To investigate whether skeletal descriptions of shape also play a role in object recognition, we functionally defined object-selective region LO, as well as pFs, bilaterally in each individual as

the voxels that responded more to images of intact objects than scrambled objects ($p < 10^{-4}$, uncorrected; Grill-Spector et al., 1998). Furthermore, to test the specificity of skeletal representations in object-selective regions, rather than higher-level visual regions more generally, we also defined the extrastriate body area (EBA; Downing, Jiang, Shuman, & Kanwisher, 2001) and fusiform body area (FBA; Peelen & Downing, 2005), as the voxels that responded more to images of bodies than objects ($p < 10^{-4}$, uncorrected). However, because EBA shows a high degree of overlap with LO, we subtracted any EBA voxels that overlapped with LO for each participant.

Analyses were conducted using the top 2000 voxels (1.8 × 1.8 × 1.8 mm) from each ROI (in each hemisphere) when available. For regions comprised of fewer than 2000 voxels, all voxels in the ROI were used (see Figure 9). To ensure that results were not related to the size of the ROI, we also conducted our primary analyses using 100, 500, and 1000 voxels. The same qualitative results were found for all ROI sizes. For each functionally defined ROI, we selected voxels that exhibited the greatest selectivity to the category of interest from the localizer runs (e.g., the 2000 most object-selective voxels in right LO). For the probabilistically-defined ROIs, we selected voxels with the greatest probability value (e.g., the 2000 voxels most likely to describe right V1). ROIs were analyzed by combining left and right hemispheric ROIs (4000 voxels total).

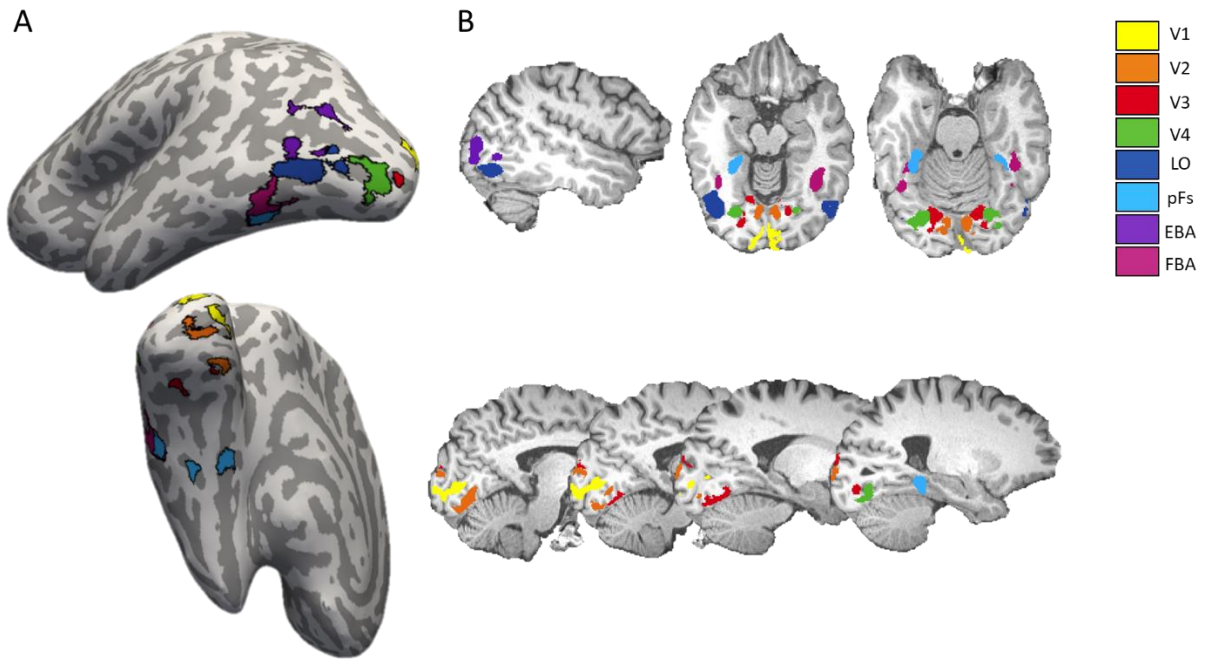


Figure 9. ROIs (2000 voxels) in a sample participant displayed on (A) the cortical surface and (B) in volumetric space. Each color corresponds to a different ROI. Early visual cortex ROIs (V1-V4) were defined using probabilistic maps. Higher-level visual regions (LO, pFs, EBA, FBA) were functionally defined in each participant using an independent localizer.

To investigate whether a model of skeletal similarity explained unique variance in each ROI, we used RSA (Kriegeskorte, Mur, & Bandettini, 2008). For each participant, parameter estimates for each stimulus (relative to fixation) were extracted for each voxel in an ROI. Responses to the stimuli in each voxel were then normalized by subtracting the mean response across all stimuli. A 12 × 12 symmetric neural representational dissimilarity matrix (RDM) was created for each ROI and participant by correlating (1-Pearson correlation) the voxel-wise responses for each stimulus with every other stimulus in a pairwise fashion. Neural RDMs were then Fisher transformed and averaged across participants separately for each ROI. Only the upper triangle of the resulting matrix (excluding the diagonal) was used in the following analyses. Although most dissimilarity measures produce similar results, we used Pearson correlation similarity because simulations have shown it to be more reliable than other similarity measures (Walther et al., 2016). Nevertheless, data were

also analyzed using cosine and squared Euclidean distance similarity metrics. The results were qualitatively the same across all measures.

Neural RDMs were compared to RDMs created from a model of skeletal similarity, as well as other models of visual similarity (GBJ, GIST, HMAX, and AlexNet-fc6). Skeletal similarity was calculated in 3D, object-centered, space as the mean Euclidean distance between each point on one skeleton and the closest point on the second skeleton following maximal alignment. Gabor-jet, GIST, and AlexNet (fc6-layer) similarity was calculated by extracting feature vectors from each model and computing the mean Euclidean distance between feature vectors for each feature vector. HMAX (C2-layer) similarity was calculated as the Pearson correlation between feature vectors. Because our primary analyses involve comparing the amount of unique variance explained by the skeletal model relative to the other models, we ensured that the skeletal model did not exhibit a high degree of multicollinearity with any other model, VIF = 2.61. Multicollinearity statistics for control models were also within an acceptable range (VIFs < 4.58; O'Brien, 2007).

Results

How are shape skeletons represented in the visual system?

We first tested whether skeletal similarity was predictive of the multivariate response pattern in each ROI by correlating the neural RDMs from each ROI with an RDM computed from a model of skeletal similarity. Significant correlations were found for V1-V4, and LO, rs = 0.35 – 0.67, R^2 = 12.5 – 50.1 (ps < .001; significance determined via permutation test with 10,000 permutations; see Figure 10). Skeletal similarity was not predictive of the response pattern in pFs, EBA, or FBA (ps > .23), revealing specificity in the predictive power of the skeletal model (see Table 1).

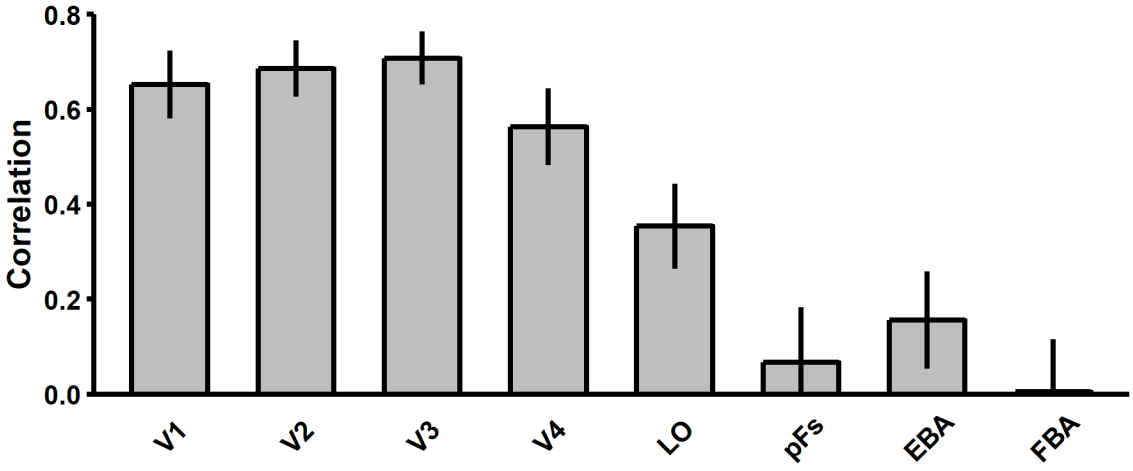


Figure 10. Bar plot displaying the correlations between the skeletal model and the multivariate response pattern in each ROI. A model of skeletal similarity was significantly correlated with response patterns in V1-V4 and LO. A skeletal model was not predictive of the response pattern in pFs, EBA, or FBA. Error bars represent bootstrapped SE.

Next, we tested whether skeletal similarity explained unique variance in each region, or whether these effects could be explained by another model of visual similarity (see Table 1 for correlations between the skeletal model and all other models). To test whether the skeletal model explained unique variance in each ROI, we conducted linear regression analyses with each neural RDM as the dependent variable, and the different models of visual similarity as predictors (Skeleton ∪ GBJ ∪ GIST ∪ HMAX ∪ AlexNet-fc6; see Figure 11A). These analyses revealed that the skeletal model explained unique variance in V3 (β = 0.46, p = .003) and LO (β = 0.49, p = .029), but not in the other regions (βs < 0.29, ps > .14).

We also conducted variance partitioning analyses (VPA) to determine how much unique variance was explained by the skeletal model in V3 and LO (Bonner & Epstein, 2018; Lescroart et al., 2015). These analyses allowed us to determine how much of the total explainable variance was unique to the different models and how much was shared by a combination of models (shared variance is illustrated in Supplemental Table 5). These analyses revealed that the skeletal model uniquely accounted for 9.0% of the total explainable variance in V3 and 25.5% of the explainable

variance in LO (see Figure 11B-C; details about other models are provided below). Thus, shape skeletons account for significant unique variance in V3 and LO even when compared with other models of visual similarity. Together, these results are consistent with the hypothesized dual role of shape skeletons in visual processing: namely, for perceptual organization and object recognition.

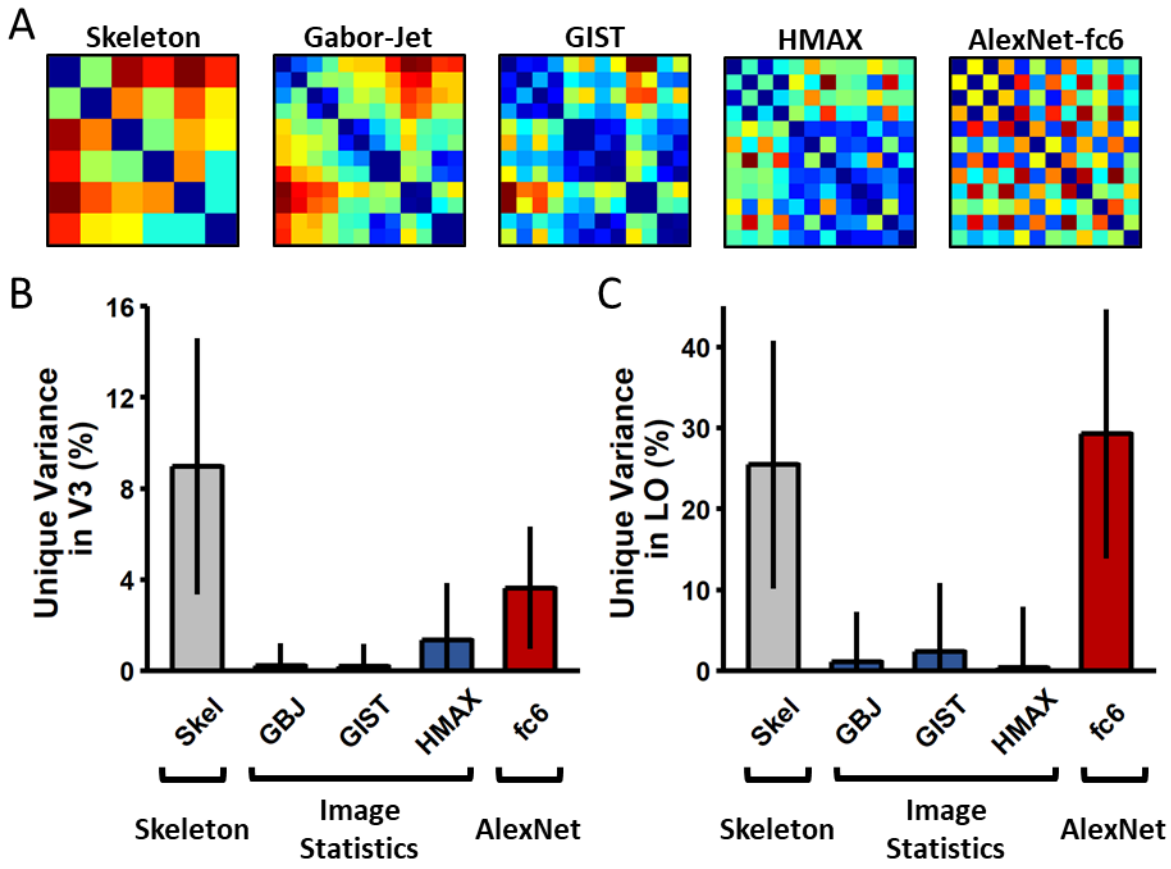


Figure 11. Variance partitioning results. (A) Dissimilarity matrices computed from models of skeletal similarity and other models of visual similarity. (B) Bar plot displaying the percentage of unique variance accounted for by each model in V3. (C) Bar plot displaying the percentage of unique variance accounted for by each model in LO. A model of skeletal similarity explained unique variance in both V3 and LO, but not in other cortical regions. Error bars represent bootstrapped SE.

Table 1. Results of the correlation, regression, and variance partitioning analyses for each ROI and each model. Correlation analyses were conducted by correlating RDMs created from the neural data from each ROI with RDMs created from each model. Regression analyses were conducted for each neural RDM by entering each model RDM as a predictor into a linear regression model. The R^2 values indicates the total explained variance by all of the models. Variance partitioning analyses were conducted by iteratively regressing each neural RDM on RDMs from each model and the combination of models, and then calculating the percentage of the total explained variance (R^2) uniquely explained by each model.

		Model Correlations			Model Regression			Variance Partitioning
ROI	Model	*r*	r^2	*p*	R^2	β	*p*	Percentage
V1		-	-	-	0.52	-	-	-
	Skeleton	0.65	0.43	< .001	-	0.26	.138	3.52
	Gabor-Jet	0.66	0.44	< .001	-	0.43	.070	5.27
	GIST	0.56	0.31	< .001	-	-0.09	.618	0.39
	HMAX	0.46	0.21	< .001	-	0.14	.297	1.72
	AlexNet-fc6	0.38	0.15	.002	-	0.13	.366	1.29
V2		-	-	-	0.63	-	-	-
	Skeleton	0.69	0.47	< .001	-	0.23	.140	2.19
	Gabor-Jet	0.71	0.51	< .001	-	0.55	.010	7.06
	GIST	0.60	0.36	< .001	-	-0.18	.275	1.19
	HMAX	0.53	0.28	< .001	-	0.13	.265	1.24
	AlexNet-fc6	0.48	0.23	< .001	-	0.24	.049	3.98
V3		-	-	-	0.63	-	-	-
	Skeleton	0.71	0.50	< .001	-	0.46	.003	8.97
	Gabor-Jet	0.67	0.44	< .001	-	0.10	.637	0.22
	GIST	0.62	0.39	< .001	-	0.07	.666	0.18
	HMAX	0.56	0.32	< .001	-	0.14	.239	1.36
	AlexNet-fc6	0.51	0.26	< .001	-	0.23	.056	3.63
V4		-	-	-	0.51	-	-	-
	Skeleton	0.56	0.32	< .001	-	0.18	.321	1.58
	Gabor-Jet	0.60	0.36	< .001	-	0.41	.085	4.83
	GIST	0.55	0.30	< .001	-	-0.11	.558	0.55
	HMAX	0.50	0.25	< .001	-	0.06	.678	0.27
	AlexNet-fc6	0.54	0.29	< .001	-	0.38	.009	11.53
LO		-	-	-	0.25	-	-	-
	Skeleton	0.35	0.13	.002	-	0.49	.029	25.46
	Gabor-Jet	0.25	0.06	.022	-	-0.13	.644	1.10
	GIST	0.23	0.05	.036	-	-0.16	.498	2.37
	HMAX	0.33	0.11	.004	-	-0.05	.776	0.42
	AlexNet-fc6	0.39	0.15	< .001	-	0.41	.020	29.27
pFs		-	-	-	0.04	-	-	-
	Skeleton	0.07	0.00	.294	-	0.29	.240	52.33
	Gabor-Jet	-0.02	0.00	.563	-	-0.08	.799	2.44
	GIST	-0.09	0.01	.761	-	-0.17	.505	16.71
	HMAX	-0.07	0.01	.727	-	-0.10	.590	10.90
	AlexNet-fc6	-0.10	0.01	.782	-	0.01	.959	0.10
EBA		-	-	-	0.07	-	-	-
	Skeleton	0.16	0.02	.107	-	0.26	.280	25.29
	Gabor-Jet	0.07	0.01	.287	-	0.04	.900	0.34
	GIST	0.00	0.00	.515	-	-0.30	.233	30.91

	HMAX	0.16	0.03	.096	-	0.05	.790	1.52
	AlexNet-fc6	0.12	0.02	.159	-	0.15	.428	13.57
FBA		-	-	-	0.06	-	-	-
	Skeleton	0.00	0.00	.484	-	-0.19	.442	16.68
	Gabor-Jet	0.04	0.00	.393	-	0.48	.142	61.84
	GIST	-0.06	0.00	.695	-	-0.42	.105	75.66
	HMAX	0.09	0.01	.242	-	0.12	.515	11.93
	AlexNet-fc6	0.05	0.00	.351	-	0.05	.785	2.09

Does skeletal coding in V3 and LO generalize across changes in surface form?

As described previously, a strength of skeletal models is that they can be used to describe an object's shape across variations in contours or component parts. Thus, if V3 and LO indeed incorporate a skeletal model, then, at the very least, these regions should represent objects by their skeletons across changes in surface form (see Figure 8). To test this prediction, new dissimilarity vectors were created from neural and model RDMs by extracting similarity values from only those object pairs whose surface forms differed and then correlating them to one another.

As predicted, skeletal similarity was a significant predictor of both V3 (r = 0.77, p < .001) and LO (r = 0.47, p < .001), even though object pairs were comprised of different surface forms. Notably, the finding that both V3 and LO represent shape skeletons across changes in surface form provides further evidence that skeletal coding in these regions cannot be accounted for by low-level shape properties such as contours and component parts.

But might another model of visual similarity account for these results? Here we conducted a similar regression analysis as above (neural RDMs ~ f[Skeleton ∪ GBJ ∪ GIST ∪ HMAX ∪ AlexNet-fc6]), but now included subject as the random effect because fewer object pairs were involved. This analysis revealed that the skeletal model explained the greatest amount of variance in both V3 (β = 0.28, p < .001) and LO (β = 0.21, p < .001; see Supplemental Table 5 for variance explained by the other models). Thus, not only are V3 and LO sensitive to object skeletons, the skeletal representations in these regions are invariant to changes in surface form.

Are V3 and LO predictive of participants' similarity judgments of objects?

Previous research has shown that shape skeletons are predictive of human participants' behavioral judgments of object similarity (Ayzenberg & Lourenco, 2019b; Destler et al., 2019; Lowet et al., 2018). Our neuroimaging results suggest that these judgments may be supported by areas V3 and LO. Here we tested this possibility by examining whether the response patterns of V3 and LO explain unique variance in humans' judgments of object similarity. Using discrimination data from Ayzenberg and Lourenco (2019), we created behavioral RDMs for the present objects. Using linear regression analyses, we first tested whether a model of skeletal similarity explained unique variance in behavioral judgments, after controlling for other models of visual similarity (GBJ, GIST, HMAX, AlexNet-fc6). We found that the skeletal model explained the greatest amount of variance in participants' judgments (VPA = 22%, $\beta = 0.28$, $p < .001$), replicating Ayzenberg and Lourenco (2019b). Next we tested whether the response profile of V3 and LO were also predictive of the behavioral RDM. These analyses revealed significant correlations for both V3 ($r = 0.81$, $p < .001$) and LO ($r = 0.46$, $p < .001$) and participants' judgments. In a final analysis, we tested whether V3 and LO were uniquely predictive of participants' behavioral judgments, or whether another region could explain this effect. We tested whether V3 and LO explained unique variance in participants' judgments by conducting separate regression analyses in which V3 and the other early visual regions (V1, V2, V4) were predictors, or LO and the other high-level visual regions (pFs, EBA, FBA) were predictors. The behavioral RDM was the dependent variable in both cases. These analyses revealed that V3 (VPA = 10%, $\beta = 0.83$, $p = .002$) and LO (VPA = 70%, $\beta = 0.70$, $p < .001$) explained unique variance in participants' similarity judgments, even when controlling for other early- and high-level visual regions, respectively.

What role do other models of visual similarity play in the visual processing of objects?

Although the skeletal model was predictive of the response profiles of V3 and LO, even across different surface forms, one might ask whether the other visual models still play a role in the neural processing of objects. For example, previous research has shown that other models of visual

similarity account for unique variance in participants' object similarity judgments (Ayzenberg & Lourenco, 2019). To explore whether other models play a role in the neural processing of objects, we tested whether these other models explained unique variance in the ROIs. Linear regression analyses revealed that the Gabor-jet model, which approximates V2-like complex cells, accounted for unique variance in the response profile of V2 (β = 0.55, p = .009), but not other regions. We also found that AlexNet-fc6, a model consisting of non-linear features, explained increasingly more variance in increasingly higher-level visual regions (V2: β = 0.24, p = .049; V4: β = 0.38, p = .009; LO: β = 0.41, p = .020). None of the models were predictive of the response profiles of V1, pFs, EBA, or FBA (ps > .070; see Table 1). Thus, the predictive power of these models of visual processing is largely consistent with the hypothesized regions they are meant to approximate.

Discussion

In the present study, we tested the hypothesis that shape skeletons are associated with two visual processes: namely, the creation of shape percepts and object recognition. Consistent with this hypothesis, we found that a model of skeletal similarity was predictive of the response pattern in V3, a region implicated in perceptual organization, and LO, a region involved in object recognition. Moreover, and crucially, skeletal representations in these regions could not be explained by low-, mid-, or high-level image properties, as described by other computational models of vision, nor by representations based on contours or component parts (i.e., surface forms) of the objects. These results provide novel neural evidence that the human visual system represents shape skeletons and may do so for both perceptual organization and object recognition.

The finding that V3 represents shape skeletons is consistent with human neuroimaging studies showing its involvement in perceptual organization. But how might shape skeletons arise in V3? One possibility is that shape skeletons reflect the response profile of grouping cells (G-cells), which play an important role within neural models of perceptual organization. More specifically, these models suggest that perceptual organization is accomplished by border ownership cells (B-

cells) in V2, which selectively respond to the contours of a figure (rather than the background), as well as G-cells in the subsequent visual region, which coordinate the firing of B-cells via top-down connections and help to specify the contours that belong to the same figure (von der Heydt, 2015; Zhou, Friedman, & von der Heydt, 2000). Interestingly, G-cells exhibit properties associated with shape skeletons. For example, G-cells specify the relations between contours, which may allow the visual system to determine an object's shape despite noisy or incomplete visual information (Craft, Schütze, Niebur, & von der Heydt, 2007; Martin & von der Heydt, 2015). Moreover, the response profile of G-cells within a shape corresponds to the points of the shape's skeleton (Craft et al., 2007), as would be expected if they implement a skeletal computation. Indeed, pruned shape skeletons, resembling those extracted from 2D shapes by human participants (Ayzenberg, Chen, et al., 2019), can be generated using a model of perceptual organization that incorporates the response profile of G-cells (Ardila, Mihalas, von der Heydt, et al., 2012).

Nevertheless, one might ask why we did not find evidence of skeletal representations in V2 or V4, given that these regions are also frequently implicated in perceptual organization (Cox et al., 2013; McMains & Kastner, 2010; Zhou et al., 2000), particularly in electrophysiology studies with monkeys (von der Heydt, 2015). First, if shape skeletons reflect the response profile of G-cells, then they would not arise in V2, which is primarily comprised of B-cells. Moreover, G-cells are thought to arise in the visual region directly following V2 (Craft et al., 2007; Martin & von der Heydt, 2015) which, in humans, is V3 but, in monkeys, is often delineated as V4 (DiCarlo et al., 2012; Gross, Rodman, Cochin, & Colombot, 1993; Serre, Oliva, et al., 2007). Studies have shown that V3 in humans is proportionally much larger than in monkeys, and there is debate regarding whether monkeys have a human-like V3 at all (Arcaro & Kastner, 2015; but, see Brewer, Press, Logothetis, & Wandell, 2002). Most relevant here is the fact that few studies on perceptual organization with monkeys have recorded from V3. Instead, these studies have focused primarily on V2 and V4 (Hegdé & Van Essen, 2006; Poort et al., 2012; Zhou et al., 2000). Our findings suggest that V3 may

be the locus of G-cells in humans and that the skeletal representations within V3 may be an emergent property of G-cell responses.

We also found evidence of shape skeletons in LO, which is consistent with a role for skeletons in object recognition. Much work has illustrated the importance of LO in using shape for object recognition (Chouinard, Whitwell, & Goodale, 2009; Grill-Spector, Kushnir, Hendler, & Malach, 2000). This region has been shown to be particularly sensitive to object-centered shape information and is tolerant to viewpoint changes and border perturbations (Grill-Spector et al., 2001; Grill-Spector et al., 1998). Our results suggest that LO may achieve such invariance by incorporating a skeletal description of shape, which provides a common format by which to compare shapes across variations in contours and component parts. Importantly, our results are consistent with electrophysiology work in monkeys in which the skeletal structure of 3D objects can be decoded from monkey IT across changes in both object orientation and surface form (Hung et al., 2012). Our findings are also consistent with patient studies in which damage to LO results in a specific impairment perceiving the spatial relations of component parts, but not the parts themselves, as would be predicted by a skeletal model (Behrmann et al., 2006; Konen, Behrmann, Nishimura, & Kastner, 2011). Building on these studies, the present work provides the first direct evidence of skeletal representations in human LO and, crucially, demonstrates that such representations cannot be accounted for by other models of visual processing.

Interestingly, we did not find evidence of skeletal representations in another object-selective region, namely pFs. This finding may reflect a division of labor between LO and pFs, following the posterior-to-anterior anatomical gradient of shape-to-category selectivity in the ventral stream (Bracci & Op de Beeck, 2016; Freud et al., 2017). More specifically, many studies have illustrated that shape selectivity peaks in posterior regions of the ventral stream and decreases in higher-level anterior regions (Brincat & Connor, 2004, 2006; Freud et al., 2017). By contrast, sensitivity to semantic category-level information, and other non-shape visual

information, progressively increases in anterior regions of the temporal lobe (Barense, Gaffan, & Graham, 2007; Behrmann, Lee, Geskin, Graham, & Barense, 2016). Given that skeletal models are exclusively descriptions of shape, such that they do not take semantic content into account, it follows that we did not find evidence of shape skeletons in pFs.

Importantly, we found that the predictive value of a skeletal model in V3 and LO held even when controlling for low- (i.e., Gabor-jet), mid- (i.e., GIST, and HMAX), and high-level (i.e., AlexNet-fc6) models of visual processing. Not only are these other models representative of different levels of visual processing, but they also approximate different theories of object recognition, such as those based on image-level similarity (i.e., Gabor-jet and HMAX; Tarr & Bülthoff, 1998) and feature descriptions (i.e., AlexNet-fc6; Ullman et al., 2016; Yamins et al., 2014). Moreover, by changing the object's surface forms, we changed the non-accidental properties of the object's component parts, thereby allowing for a test of component description theories (Biederman, 1987; Kayaert, Biederman, & Vogels, 2003). That skeletal models explained unique variance even when controlling for these other properties suggests that shape skeletons may play a privileged role in the visual processing of objects and highlights their importance in theories of object recognition.

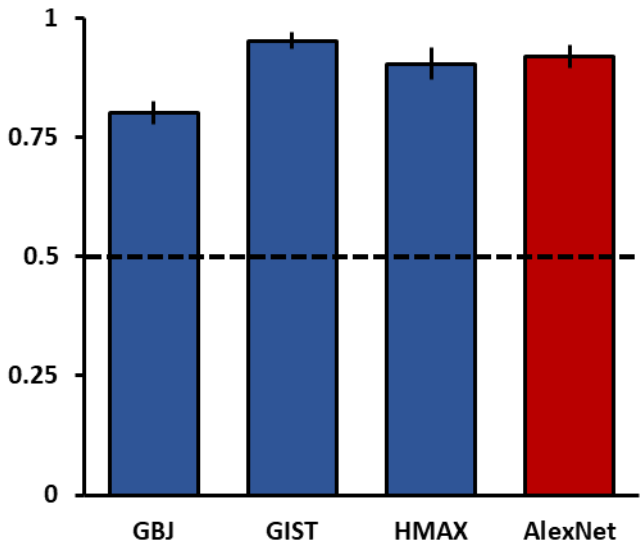
Yet our results also point to the contributions of two other models of vision to the neural processing of objects. We found that the Gabor-jet model was predictive of the response profile in V2 (earlier in the hierarchy than shape skeletons) and that AlexNet-fc6 was most predictive in LO (same region as skeletons). That these models are also associated with object processing is perhaps unsurprising given that other visual properties are important for solving a range of object recognition tasks. For instance, lower-level visual properties and feature descriptions may be particularly important for subordinate- (Biederman, Subramaniam, Bar, Kalocsai, & Fiser, 1999; Davitt, Cristino, Wong, & Leek, 2014) or superordinate-level (Long, Störmer, & Alvarez, 2017; Long, Yu, & Konkle, 2018) categorization. Collectively, our findings illustrate how the visual system may incorporate multiple models in parallel to create holistic object representations.

Although we have suggested that shape skeletons may be implicated in both perceptual organization and object recognition, the spatial and temporal resolution of fMRI places important qualifiers on these conclusions. First, although our results were consistent across different ROI sizes (see Supplemental Data), it is nevertheless possible that shape skeletons are represented in sub-populations of neurons within each region and that these regions have secondary functions. Indeed, V3 and LO have been shown to be sensitive to other types of visual cues, including motion (Dupont et al., 1997; Felleman & Van Essen, 1987) and depth (Parker, 2007; Welchman, 2016). Moreover, our own results showed that other models of vision (i.e., AlexNet-fc6) were represented in tandem with shape skeletons in these regions. Our data also cannot address whether skeletal representations in these regions arise via feedforward or feedback processes. Indeed, feedback processes are known to be important for both perceptual organization (Mannion, McDonald, & Clifford, 2010; Murray, Kersten, Olshausen, Schrater, & Woods, 2002; Wokke, Vandenbroucke, Scholte, & Lamme, 2013) and invariant object recognition (Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Tang et al., 2018), and may even lead to skeleton-like representations in early visual cortex (Lee, 1996; Li, 2000). A more complete understanding of V3 and LO, along with experiments designed to test the causal role of shape skeletons in human vision, will be needed to confirm the claims of the present research. Nevertheless, our findings are consistent with computer vision, behavioral, and neuroimaging work suggesting a role for shape skeletons in both perceptual organization and object recognition.

Supplemental Materials for "A dual role for shape skeletons in human vision: Perceptual organization and object recognition"

Can non-skeletal models discriminate the objects?

A potential concern with the current study is that, because we explicitly varied skeletal similarity, object skeletons were an especially salient cue. To address this concern, we tested whether non-skeletal models could discriminate the objects. If they could, it would suggest that the visual system need not necessarily rely on a skeletal model to discriminate these objects. A feature vector was extracted for every image (30 skeletons × 5 surface forms × 3 orientations) from each of these models (GBJ, GIST, HMAX, AlexNet-fc6). Then, for each model and object pair (same surface form), a linear support vector machine (SVM) classifier was trained to label objects using two object orientations; its ability to label the objects was tested using the third orientation. This procedure was repeated for every surface form and every combination of orientations between objects (0° × 0°; 0° × 30°; 0° × -30°; 30° × 30°; 30° × -30°; -30° × -30°). A final discrimination score was computed for each object pair by averaging the decoding accuracies across every surface form and combination of orientations. This analysis revealed that every model could discriminate between objects significantly above chance (0.50; $Ms > 0.80$), $ts > 11.78$, $ps < .001$, $ds > 3.30$ (see Supplemental Figure 4). Together, these findings demonstrate that the objects within our stimulus set were sufficiently different along other visual dimensions such that non-skeletal models could accurately discriminate them.



Supplemental Figure 4. Discrimination accuracies for all non-skeletal models. Each model was able to discriminate between objects significantly above chance (dotted line).

Supplemental Table 4. Percentages of unique and shared variance explained by each model and for each ROI.

	ROI							
Model	V1	V2	V3	V4	LO	PFS	EBA	FBA
Skel	3.5%	2.2%	9.0%	1.6%	25.5%	52.3%	25.3%	16.7%
GBJ	5.3%	7.1%	0.2%	4.8%	1.1%	2.4%	0.3%	61.8%
GIST	0.4%	1.2%	0.2%	0.5%	2.4%	16.7%	30.9%	75.7%
HMAX	1.7%	1.2%	1.4%	0.3%	0.4%	10.9%	1.5%	11.9%
fc6	1.3%	4.0%	3.6%	11.5%	29.3%	0.1%	13.6%	2.1%
Skel + GBJ	18.0%	17.5%	11.1%	12.2%	9.1%	17.8%	28.7%	-13.9%
Skel + GIST	0.3%	0.4%	-0.2%	0.3%	2.3%	8.2%	7.7%	-7.3%
GBJ + GIST	1.3%	-0.2%	0.9%	0.6%	7.3%	33.5%	19.3%	-55.0%
Skel + HMAX	2.3%	1.5%	3.6%	0.7%	0.6%	-9.7%	7.3%	-6.7%
GBJ + HMAX	-0.3%	-0.3%	-0.1%	-0.1%	-0.1%	-0.6%	-0.1%	-3.1%
GIST + HMAX	0.1%	0.2%	-0.1%	0.1%	-0.1%	-1.7%	1.0%	4.3%
Skel + fc6	-0.6%	-0.9%	-1.7%	-1.2%	-8.4%	0.9%	-5.6%	2.9%
GBJ + fc6	-0.3%	-0.7%	-0.1%	-1.0%	1.0%	0.1%	-0.2%	-1.3%
GIST + fc6	-0.3%	-1.0%	1.1%	-0.3%	-2.3%	1.3%	-10.0%	0.6%
HMAX + fc6	6.0%	9.5%	9.3%	13.6%	15.3%	6.3%	22.9%	23.1%
Skel + GBJ + GIST	28.5%	23.2%	24.5%	18.3%	-11.3%	-53.3%	-48.2%	7.2%
Skel + GBJ + HMAX	6.0%	5.4%	4.2%	2.8%	2.0%	-0.6%	9.1%	12.8%
Skel + GIST + HMAX	0.3%	0.4%	0.1%	0.2%	0.7%	0.8%	3.6%	0.8%
GBJ + GIST + HMAX	-0.5%	-0.4%	-0.2%	-0.1%	-0.2%	-3.9%	2.7%	0.4%
Skel + GBJ + fc6	-0.1%	-2.2%	-1.6%	-4.6%	-7.0%	4.2%	-5.7%	-3.5%
Skel + GIST + fc6	0.0%	-0.2%	-0.8%	-0.4%	-2.2%	7.4%	2.4%	-4.7%
GBJ + GIST + fc6	2.1%	3.6%	1.6%	6.0%	-6.5%	5.8%	-2.6%	-0.3%
Skel + HMAX + fc6	1.2%	1.5%	2.7%	1.8%	11.1%	-2.1%	8.4%	-4.0%
GBJ + HMAX + fc6	-1.2%	-1.8%	0.0%	-1.7%	2.2%	-1.0%	0.1%	-8.6%
GIST + HMAX + fc6	0.0%	-0.2%	2.6%	1.6%	1.7%	12.4%	-9.6%	-23.5%
Skel+ GBJ + GIST + HMAX	4.9%	3.8%	4.5%	2.4%	0.1%	4.9%	-10.2%	-9.3%
Skel+ GBJ + GIST + fc6	-0.7%	1.1%	1.8%	2.9%	14.3%	-10.3%	8.5%	5.1%
Skel+ GBJ + HMAX + fc6	-2.6%	-1.2%	-2.6%	0.6%	-2.9%	0.9%	-5.6%	2.7%
Skel+ GIST + HMAX + fc6	-0.8%	-0.7%	-0.9%	-0.6%	-2.2%	-4.7%	-6.4%	6.0%
GBJ+ GIST + HMAX + fc6	2.5%	3.1%	0.6%	2.6%	-2.3%	7.3%	-3.6%	11.8%
Skel+ GBJ + GIST + HMAX + fc6	21.8%	23.1%	25.3%	24.7%	19.2%	-6.5%	14.4%	-4.5%

Supplemental Table 5. Random-effects regression results for each model and for each ROI on objects with different surface forms.

		Model Correlations		Random Effects Regression	
ROI	Model	*r*	*p*	β	*p*
V1				-	-
	Skeleton	0.733	< .001	0.115	.132
	Gabor-Jet	0.639	< .001	-0.062	.534
	GIST	0.586	< .001	0.083	.268
	HMAX	0.526	< .001	-0.030	.548
	AlexNet-fc6	0.771	< .001	0.440	.001
V2		-	-	-	-
	Skeleton	0.762	< .001	0.062	.385
	Gabor-Jet	0.738	< .001	0.137	.142
	GIST	0.665	< .001	0.052	.456
	HMAX	0.482	.003	-0.044	.354
	AlexNet-fc6	0.752	< .001	0.550	.000
V3		-	-	-	-
	Skeleton	0.766	< .001	0.241	.001
	Gabor-Jet	0.688	< .001	-0.045	.615
	GIST	0.654	< .001	0.152	.026
	HMAX	0.513	.001	-0.031	.489
	AlexNet-fc6	0.744	< .001	0.512	.000
V4		-	-	-	-
	Skeleton	0.682	< .001	0.139	.068
	Gabor-Jet	0.620	< .001	0.001	.995
	GIST	0.563	< .001	0.077	.304
	HMAX	0.484	.003	0.006	.901
	AlexNet-fc6	0.676	< .001	0.330	.010
LO		-	-	-	-
	Skeleton	0.467	.004	0.280	.001
	Gabor-Jet	0.366	.027	-0.031	.767
	GIST	0.275	.104	-0.043	.593
	HMAX	0.196	.253	-0.050	.353
	AlexNet-fc6	0.321	.056	-0.007	.959
pFs		-	-	-	-
	Skeleton	0.037	.832	0.261	.001
	Gabor-Jet	-0.046	.793	-0.149	.162
	GIST	-0.076	.665	-0.036	.656
	HMAX	-0.072	.677	-0.045	.402
	AlexNet-fc6	-0.065	.705	-0.180	.188
EBA		-	-	-	-
	Skeleton	0.267	.117	0.145	.077
	Gabor-Jet	0.187	.271	0.029	.786
	GIST	0.100	.560	-0.084	.297
	HMAX	0.170	.318	0.047	.384
	AlexNet-fc6	0.164	.339	-0.116	.397
FBA		-	-	-	-

Skeleton	-0.057	.738	0.136	.095
Gabor-Jet	-0.134	.431	-0.126	.236
GIST	-0.142	.404	-0.011	.892
HMAX	-0.007	.964	0.026	.634
AlexNet-fc6	-0.076	.655	-0.111	.414

Chapter 4 - The shape skeleton supports single exemplar categorization in infants (Ayzenberg & Lourenco, 2019a)

Thus far, evidence from human adults demonstrates that participants spontaneously extract the skeletons of 2D shapes, even in the presence of incomplete contours (Ayzenberg, Chen, et al., 2019; Firestone & Scholl, 2014). Moreover, a 3D skeletal model predicted participants' object similarity and category judgments, even when controlling for other models of vision (Ayzenberg & Lourenco, 2019b). Finally, neuroimaging studies with human and nonhuman adult primates demonstrate that shape skeletons are represented in early visual areas (i.e., V3; Lescroart & Biederman, 2012), putatively involved in creating "shape percepts", and higher-level visual areas (e.g., inferiortemporal cortex; Hung et al., 2012), known for their involvement in object recognition.

Nevertheless, adults have had extensive experience categorizing objects across variations in appearance. Thus, it is unknown whether skeletal representations support rapid object learning early in development, when experience is more limited. To answer this question, we tested whether infants categorize objects with identical skeletons, but different component parts and image-level properties, as similar. To further minimize effects of experience, infants were tested with unfamiliar objects and required to categorize objects from a single exemplar.

Experiment 1

Using a habituation-dishabituation paradigm, we tested whether infants categorize objects by their shape skeletons across variations in surface form, which changed the component-part and image-level properties of the objects without altering the skeleton. If shape skeletons are used for categorization, then infants should look longer at an object with a novel skeleton compared to a familiar skeleton, even across different surface forms.

Methods

Participants. Thirty-four full-term infants (M = 9.53 months, range = 6.47 – 12.2 months; 18 female) participated in this experiment (6 additional infants were excluded: 5 for fussiness and

1 because of equipment failure). Sample size was determined by an a priori power-analysis with a hypothesized medium effect size (d = 0.50; 1 - β > .8).

Stimuli. Six videos of 3D novel objects were rendered from the stimulus set created by Ayzenberg and Lourenco (2019b; see Figure 12 and Supplemental Figure 5). The object set was comprised of three distinct skeletons chosen by first conducting a k-means cluster analysis (k = 3) on skeleton similarity data for 30 unique objects and then selecting three objects that adult participants judged to be equally dissimilar (see Supplemental Materials). Each object was also rendered with two different surface forms, which changed the component parts of the object without altering its skeleton (see Figure 12A). These two surface forms (out of five) were selected because a separate group of adult participants judged them to be most dissimilar (see Supplemental Materials).

We used data from Ayzenberg and Lourenco (2019b) to ensure that surface forms and skeletons were matched for perceptual discriminability and that surface forms were comprised of qualitatively different component parts (Biederman, 1987; see Supplemental Materials). Moreover, using the Gabor-jet model, a low-level model of image similarity based on the response profile of V1 (Margalit et al., 2016), we confirmed that objects with different surface forms (same skeletons) differed significantly in their image-level properties, and importantly, objects with different skeletons (same surface forms) were matched in their image-level properties (see Supplemental Materials).

Altogether, the present stimulus set provides an especially strong test of skeletal structure coding because objects with the same skeleton differed in both their component parts and image-level properties, such that if infants categorized objects on the basis of their skeletons, they would do so despite these other visual differences.

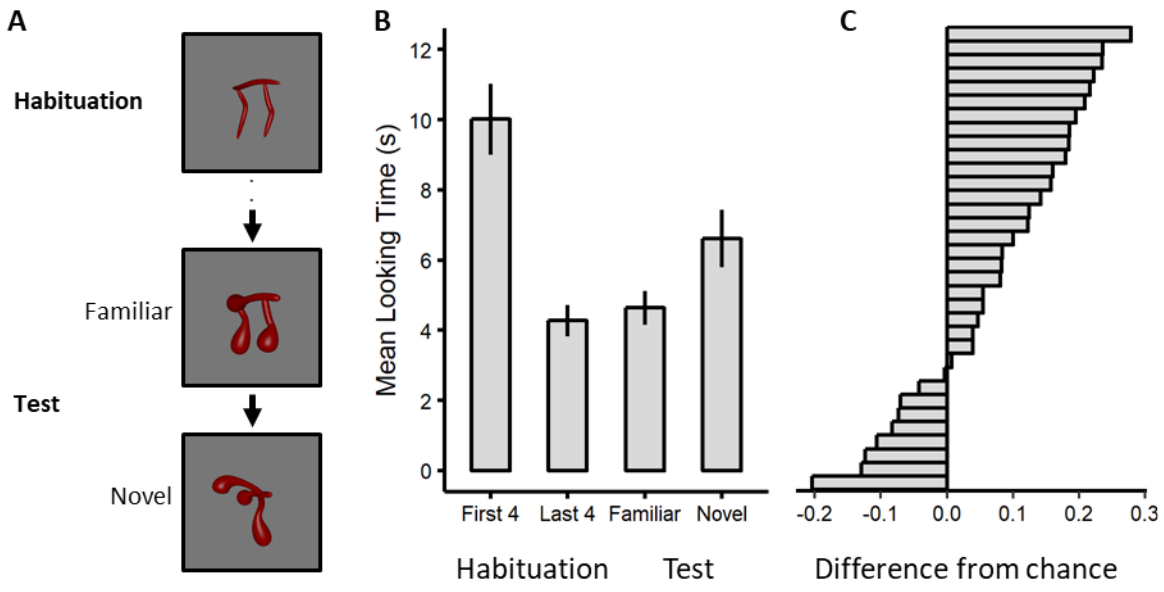


Figure 12. Experimental design and results for Experiment 1. (A) Infants were habituated to one object, and then tested with objects that had either a familiar or a novel skeleton. Both objects differed in surface form from the habituation object. (B) Mean looking times for each phase of the experiment. Results are shown for the first four and last four trials of the habituation phase, and for familiar and novel trials of the test phase. Error bars represent standard error. (C) Histogram of infants' responses on the test trials. A value greater than zero indicates greater looking time to the object with the novel skeleton.

Procedure. Infants were seated on a caregiver's lap approximately 60 cm from a 22-inch computer monitor (1920 × 1080 px). Caregivers were instructed to keep their eyes closed and to refrain from interacting with the infant during the study session. The experiment was controlled by a custom program written in Visual Basic (Microsoft) and gaze data were recorded with an Eye-Link 1000 plus eye tracker recording at 500 Hz (SR-Research). Prior to the start of the experiment, the eye tracker was calibrated to each infant using a 5-point calibration routine.

The experiment consisted of a habituation phase in which infants were presented with one object (single skeleton and surface form), and a test phase where categorization was tested using objects with familiar and novel skeletons, both of which differed in surface form from the habituated object (see Figure 12A). Each trial began with an attention-getting stimulus, which

remained onscreen until infants fixated it for 2 s. Infants were then shown a video of one object rotating across 60° (-30° to 30°; see Supplemental Materials). Videos remained onscreen for 60 s, or until infants looked away for 2 s.

Average looking times for the first four habituation trials were computed online for each infant. Infants met the habituation criterion when average looking time in the preceding four trials was below 50% of the average looking time in the first four trials of the experiment. Test trials were presented after infants habituated, or following 24 habituation trials, whichever came first. Each infant was habituated to an object with one of three possible skeletons (random assignment), with half of the infants habituated to each surface form.

All infants were presented with six test trials. The skeleton of the test objects was either identical to, or different from, the habituated object (alternating presentation; first test trial counterbalanced across infants). Each skeleton was used as the novel object in the test phase (random assignment).

Results

To ensure that results were not due to sample size decisions, we included non-parametric and Bayesian analyses in addition to null-hypothesis tests.

Parametric and non-parametric results are displayed in Figure 12. A significant decrease in looking times between the first four trials and last four trials of the habituation phase confirmed that infants habituated to the object exemplar, $t(33) = 8.39$, $p < .001$. An analysis of the test trials revealed that infants looked significantly longer to the test object with the novel skeleton ($M = 6.61$ s, $SD = 4.75$ s) than the familiar skeleton ($M = 4.64$ s, $SD = 2.84$ s), $t(33) = 3.04$, $p = .005$, $d = 0.52$, 95% CI [0.16, 0.88], with the majority of infants demonstrating this pattern of performance (73.5%, $p = .009$; binomial test). Likewise, a Bayes factor (BF) analysis (Jeffrey-Zellner-Siow prior; Jarosz & Wiley, 2014) suggested moderate support for the alternative hypothesis, $BF_{10} = 8.42$. Moreover, there was significant dishabituation to the object with the novel skeleton, $t(33) = 3.36$, $p = .002$, $d =$

0.58, BF_{10} = 17.47, but not the familiar skeleton, $t(33) = 1.00$, $p = .325$, $d = 0.17$, $BF_{10} = 0.29$. There were no effects of age (*ps* > .179) or gender (*ps* > .631) in the habituation or test phase. Altogether, these results present strong evidence that infants categorized the objects on the basis of their skeletons, despite only experiencing a single exemplar during habituation, and despite the familiar test object differing from the habituated object in both the component parts and image-level properties.

Experiment 2

Nevertheless, an alternative explanation to the findings in Experiment 1 is that infants categorized objects, not by their skeletons, but by the "coarse" spatial relations of object parts (Biederman, 1987; Hummel, 2000). Like shape skeletons, a coarse spatial-relations model describes the relations between parts. However, unlike shape skeletons, this description is qualitative, not quantitative. For example, in Figure 12, the familiar object could be characterized as two components below a third (like the habituated object), whereas the novel object could be characterized as one component on either side of a third (unlike the habituated object). Importantly, a coarse spatial-relations model predicts that only qualitative changes to the arrangement of object parts should influence categorization. To rule out this alternative model, we tested infants with objects whose coarse spatial relations were held constant and, thus, not diagnostic of category.

Methods

Participants. Forty-eight full-term infants (*M* = 9.12 months, range = 6.17 – 12.00 months; 20 female) participated in this experiment (4 additional infants were excluded: 3 for fussiness and 1 because of equipment failure). Because objects with the same coarse spatial relations also have more similar skeletons, and, thus, may be more difficult to discriminate (Ayzenberg & Lourenco, 2019), we hypothesized an attenuated effect of categorization in infants. Accordingly, to retain

adequate power, we chose to test 14 additional infants in this experiment (compared to Experiment 1), the exact number of which was determined according to a fully counterbalanced design.

Stimuli and Procedure. Four videos of 3D novel objects were rendered from the stimulus set created by Ayzenberg and Lourenco (2019; see Supplemental Figure 6). The object set was composed of two skeletons and two surface forms, the presentation of which (across habituation and test phases) was fully counterbalanced across infants. Importantly, objects consisted of the same coarse spatial arrangement of parts (i.e., a part on either side of a third; see Figure 13A), rendering these spatial relations uninformative. Thus, if infants' performance in Experiment 1 was based on coarse spatial relations, then they would be unable to categorize objects in the current experiment. However, if infants represent objects via a shape skeleton, then they should successfully categorize the objects. Finally, as in Experiment 1, we used the Gabor-jet model to ensure that objects with the same skeletons (different surface forms) differed significantly in their image properties, and importantly, that objects with different skeletons (same surface forms) did not differ in these properties (see Supplemental Materials). The procedure was identical to Experiment 1, except that objects were presented from orientations that maximized the visibility of their structure (30° to 90°; see Figure 13A and Supplemental Materials). The presentation order of test objects was counterbalanced across infants.

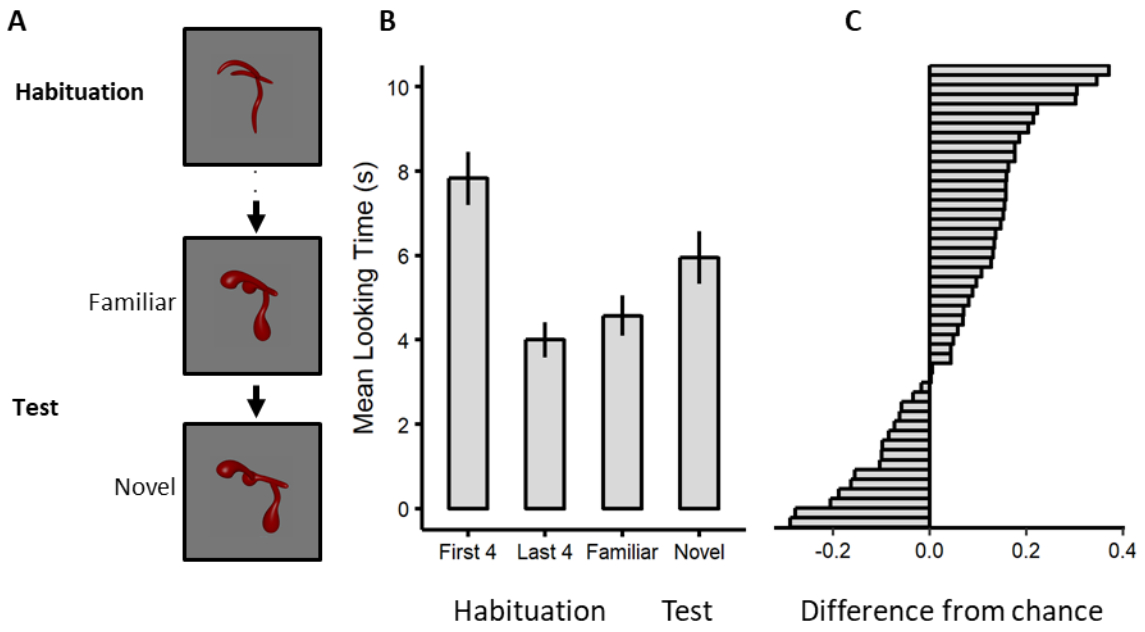


Figure 13. Experimental design and results for Experiment 2. (A) Infants were habituated to one object, and then tested with objects that had either a familiar or a novel skeleton. Both objects differed in surface form from the habituation object but had the same coarse spatial relations. (B) Mean looking times for each condition. For the habituation phase, results are shown for the first four and last four trials. For the test phase, results are shown for familiar and novel trials. Error bars represent standard error. (C) Histogram of infants' responses on the test trials. A value greater than zero indicates greater looking time to the object with the novel skeleton.

Results

Parametric and non-parametric results are displayed in Figure 13. A significant decrease in looking times between the first four trials and last four trials of the habituation phase confirmed that infants habituated to the object exemplar, $t(47) = 7.23$, $p < .001$. An analysis of the test trials revealed that infants looked significantly longer to the test object with the novel skeleton ($M = 5.96$ s, $SD = 4.32$ s) than the familiar skeleton ($M = 4.57$ s, $SD = 3.33$ s), $t(47) = 2.60$, $p = .012$, $d = 0.38$, 95% CI [0.08, 0.67], with the majority of infants demonstrating this pattern of performance (68.8%, $p = .013$; binomial test). Likewise, a Bayes factor (BF) analysis (Jeffrey-Zellner-Siow prior; Jarosz & Wiley, 2014) suggested moderate support for the alternative hypothesis, $BF_{10} = 3.18$. Moreover,

there was significant dishabituation to the object with the novel skeleton, $t(47) = 3.63$, $p < .001$, $d = 0.52$, $BF_{10} = 39.77$, but not the familiar skeleton, $t(47) = 1.42$, $p = .163$, $d = 0.16$, $BF_{10} = 0.40$. There were no effects of age (*ps* > .146) or gender (*ps* > .193) in the habituation or test phase. Finally, to ensure that effects were not unduly influenced by the larger sample size in this experiment, we computed bootstrapped CIs on a smaller sample. For each bootstrap procedure (10,000 iterations), we calculated Cohen's d on data that were resampled (without replacement) to match the sample size of Experiment 1 ($n = 34$). The bootstrapped effect size was greater than zero, 95% CI [0.20, 0.61], even with a smaller sample. These findings rule out infants' reliance on coarse spatial relations as a strategy for single exemplar categorization, providing further evidence for categorization on the basis of the object's shape skeleton.

Discussion

The ability to form robust object categories is often thought to rely on extensive experience with different exemplars, as well as conceptual object knowledge (Kibbe & Leslie, 2019; Slone, Smith, & Yu, 2019). Here we provide evidence for a perceptual mechanism that may support rapid object learning and may be innate to the visual system. Across two experiments, we found that infants were able to categorize unfamiliar objects by their skeletal structure following exposure to a single exemplar. Importantly, infants showed evidence of single exemplar categorization despite differences in the object's component parts, image-level properties, and even when the objects' coarse spatial-relations were identical. These results stand in stark contrast to state-of-the-art ANNs, which require thousands of training examples to learn an object category and have difficulty categorizing novel exemplars.

Supplemental Materials for "The shape skeleton supports single exemplar categorization in infants"

Skeleton selection

Experiment 1. Three novel objects with distinct skeletons were selected (from a set of 30) on the basis of their skeletal similarity. Skeletal similarity was calculated in 3D, object-centered, space as the mean Euclidean distance between each point on one skeleton and the closest point on the second skeleton following maximal alignment. A *k*-means cluster analysis ($k = 3$) was used to select three distinct objects, one from each cluster (see Supplemental Figure 5).

We ensured that the three selected objects were matched for discriminability by analyzing participants' discrimination judgments using data from Ayzenberg and Lourenco (2019b). Participants ($n = 42$) were shown images of two objects (side-by-side) that had either the same or different skeletons (same surface forms). Participants were instructed to decide whether the two images showed the same or different object. A repeated measures ANOVA, with skeleton pairs as the within-subject factor, revealed that the three skeletons did not significantly differ in their discriminability, $F(2, 64) = .11$, $p = .898$.

Experiment 2. We selected one object from Experiment 1 whose skeleton could be altered without changing the coarse spatial relations. We altered the object's skeleton by moving one segment 50% down the length of the central segment (see Supplemental Figure 6).

Surface form selection

Two surface forms were used in both experiments, a 'thin' (Surface Form 1) and 'bulbous' (Surface Form 2) form (see Supplemental Figures 5 and 6). Selection of these surface forms were based on adult participants' data from the study of Ayzenberg and Lourenco (2019). In a match-to-sample task, participants ($n = 39$) were shown one object (sample) placed centrally above two choice objects. One of the choice objects matched the sample's skeleton, but not surface form, and the other choice object matched the sample's surface form, but not skeleton. Participants were

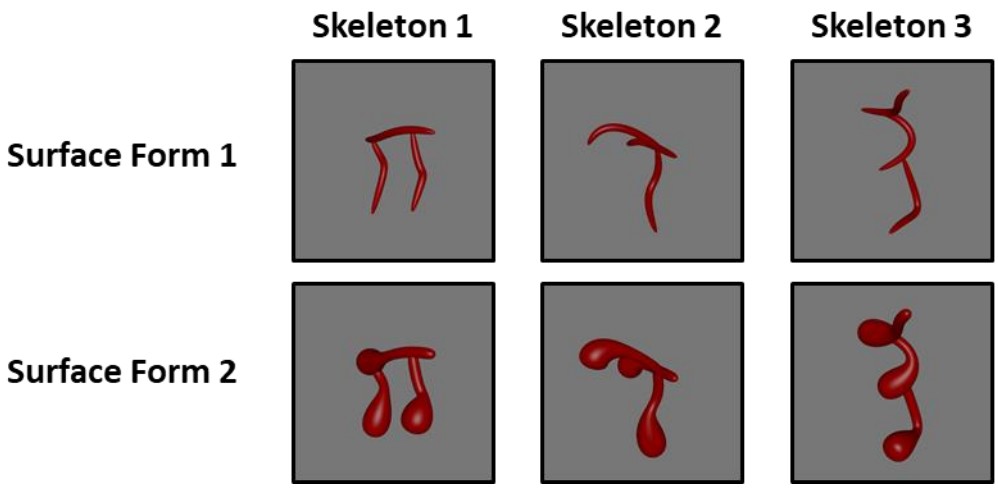
instructed to decide which of the two choice objects was most likely to be in the same category as the sample object. Participants performed worst at categorizing objects by their skeleton when Surface Form 1 was paired with Surface Form 2, $M = 0.58$, compared to the other surface forms (*Ms* = 0.61 - 0.78). Thus, by choosing the surface forms that presented adult participants with the greatest conflict, we provided infants with an especially strong test of skeletal structure coding.

To ensure that surface forms were matched in discriminability to the selected skeletons, participants ($n = 41$) conducted a surface form discrimination task, wherein they were shown images of two objects (side-by-side) that consisted of either the same or different surface forms (same skeleton). Participants were instructed to decide whether the two images showed the same or different object. Participants were found to discriminate between Surface Forms 1 and 2 significantly better than would be predicted by chance (0.50), $t(40) = 8.95$, $p < .001$, and importantly, discrimination accuracy between surface forms did not differ from discrimination accuracy between skeletons, $t(80) = 0.02$, $p = .981$.

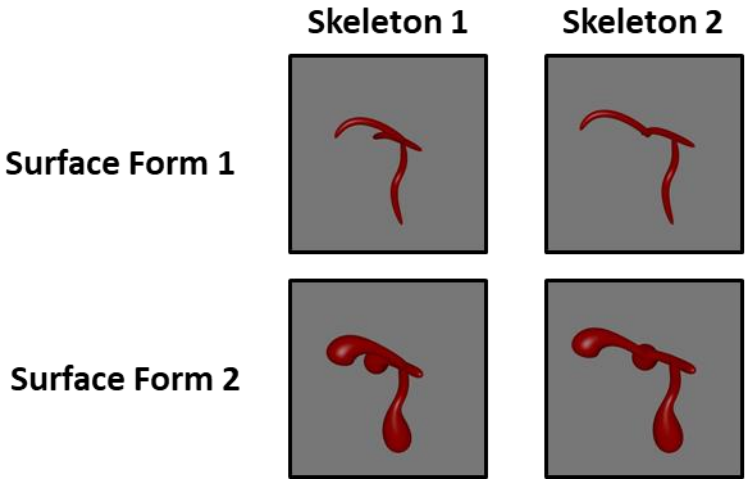
In a separate set of analyses, we tested whether surface forms were comprised of qualitatively different component parts by having participants rate each surface form on the degree to which it exhibited a specific non-accidental property (NAP). During a training phase, participants ($n = 34$) were taught four NAPs (drawn from Amir et al., 2012). They then rated the degree to which each surface form exhibited a particular NAP. The four NAPs were: (1) *taper*, defined as the degree to which the thickness of an object was reduced towards the end; (2) *positive curvature*, defined as the degree to which an object part curved outwards; (3) *negative curvature*, defined as the degree to which an object part curved inwards; and (4) *convergence to vertex*, defined as the degree to which an object part ended in a point. Prior to the statistical analyses, we ensured that all participants in this sample exhibited reliable performance ($\alpha s > 0.7$). A repeated measures ANOVA, with NAP as the within-subject factor and surface form as the between-subject factor, revealed a

significant main effect of surface form, $F(1, 66) = 64.00$, $p < .001$, suggesting that surface forms were comprised of different NAPs.

In a final analysis, we tested whether objects with different surface forms, but the same skeleton, had significantly different image-level properties. Each object video was converted into a sequence of images (30 frames/s; 300 frames total), which were analyzed with the Gabor-jet model (Margalit et al., 2016). This model overlays a 12 × 12 grid of Gabor filters (5 scales × 8 orientations) on each image. The image is convolved with each filter, and the magnitude and phase of the filtered image is stored as a feature vector. To test whether two object videos consisted of different image-level properties, we used paired *t*-tests to compare the feature vectors from each frame of one video to the corresponding frames of a second video. To provide an estimate of the image-level difference across the entire video, the resulting *p*-values from each *t*-test were then averaged across frames. This analysis revealed that objects with the same skeleton, but different surface forms, had significantly different image-level properties ($p = .002$), whereas objects with different skeletons, but same surface form, did not ($p = .09$).



Supplemental Figure 5. Stimulus set used in Experiment 1. Each column displays objects with the same skeleton, but different surface form. Each row displays objects with the same surface form, but different skeleton. Each object was presented as a video wherein the object rotated through 60° (-30° to 30°).



Supplemental Figure 6. Stimulus set used in Experiment 2. Each column displays objects with the same skeleton, but different surface form. Each row displays objects with the same surface form, but different skeleton. Objects with different skeletons consisted of the same "coarse" spatial relations. Each object was presented as a video wherein the object rotated through 60° (30° to 90°).

Chapter 5 - General Discussion

How do humans quickly form robust shape representations to accomplish object recognition? In the present dissertation, I hypothesized that a model of structure, known as the shape skeleton, supports the ability to both form global shape representations and recognize objects, with little visual experience. In Study 1, I found that shape skeletons were predictive of human object dissimilarity and category judgments, even when controlling for other models of vision. The findings from this study suggest that shape skeletons play a unique role in object recognition (Ayzenberg & Lourenco, 2019b). In Study 2, I found that a model of skeletal similarity was predictive of the multivariate patterns in V3 and LO, regions implicated in perceptual organization and object recognition (Ayzenberg, Kamps, et al., 2019). Finally, in Study 3, I found that infants, a population with little visual experience, could readily categorize never-before-seen objects using the shape skeleton, suggesting that it may support one-shot object categorization (Ayzenberg & Lourenco, 2019a). Together, these studies provide support for shape skeletons as a powerful mechanism by which humans form robust shape representations and accomplish object recognition.

Two roles for skeletons in human vision

Perceptual organization plays an important role at every level of visual analysis. At early stages of visual processing, it allows adjacent contours to be grouped into continuous line segments (Wagemans et al., 2012; Zhou et al., 2000); at the later stages, it groups local visual features into complete objects (Craft et al., 2007; Feldman et al., 2013; Martin & von der Heydt, 2015). This latter stage of processing is particularly complex because it requires the visual system to determine the relations between spatially distant elements (von der Heydt, 2015). Indeed, unlike other aspects of visual perception, neural network models do not approximate human perceptual organization processes (Linsley, Eberhardt, Sharma, Gupta, & Serre, 2017; Mehrani & Tsotsos, 2019), perhaps explaining why ANNs have difficulty recognizing objects when there are minor distortions to the

image background (Rosenfeld, Zemel, & Tsotsos, 2018; Szegedy et al., 2013; Wu, Wu, & Kreiman, 2018).

I have proposed that shape skeletons are ideally suited to support this latter stage of perceptual organization because they describe how local visual features (e.g., component parts) are arranged relative to one another to make a complete shape. Indeed, my masters work showed that skeletons were predictive of participants' responses in Kanisza shapes – stimuli where the shapes are inferred by integrating spatially distant elements (i.e., corners; Ayzenberg, Chen, et al., 2019). Moreover, Study 2 of my dissertation showed that a model of skeletal similarity explained the most variance in the multivariate pattern of area V3, a visual region consistently implicated in perceptual organization (Sasaki, 2007). Thus, shape skeletons not only offer a mechanism by which the human visual system accomplishes perceptual organization, but they also provide a quantitative method to improve these abilities in artificial neural networks.

As mentioned previously, shape skeletons are also particularly good models for object recognition because their structure remains stable across variations in viewpoint and exemplar. These properties make shape skeletons a strong candidate for explaining how organisms accomplish *invariant* object recognition – the ability to recognize never-before-seen viewpoints or exemplars of objects. Invariant object recognition has been a long-standing problem in the vision sciences because it requires the organism to form an object representation that is abstracted away from the image on the retina. Across the studies of this dissertation, I tested skeletons against multiple theories of invariant recognition: image-similarity models (e.g., Gabor-jet; Tarr & Bülthoff, 1998), component description theories (Biederman, 1987), and learned feature descriptors (i.e., AlexNet; Krizhevsky et al., 2012). In all three studies, I found that participants' behavioral and neural responses were best fit by a skeletal model. Shape skeletons explained more variance in participants' responses than either image models or ANNs, and they were a better predictor of human category judgments than models based on component parts and coarse spatial relations.

Shape skeletons can support invariance by ignoring visual features that may be considered noise. However, it remains unclear how the visual system determines which visual properties constitute noise. One possible determinant for what constitutes as noise follows from rules governing object-part segmentation (De Winter & Wagemans, 2006; Feldman et al., 2013). According to these rules, a visual feature may be less likely to be treated as noise if it is perceived as a new object part (Feldman & Singh, 2006). Thus, in current dissertation, participants readily generalized across surface form changes because they changed shape of individual component parts, but did not meet constitute as a new object part (De Winter & Wagemans, 2006; Dhandapani & Kimia, 2002; Hoffman & Richards, 1984; Singh, Seyranian, & Hoffman, 1999). By contrast, if we had added an additional part, or 'leg', to the objects, then participants would not have generalized.

A second determinant for what constitutes noise may depend on the level of perceptual detail required by the task. As described previously, the shape skeleton is organized hierarchically, such that there may be a series of parent axes that describe the shape's coarse global geometry, as well as smaller 'off-shoot' axes that describe individual component parts. Although these smaller off-shoot branches are typically pruned away during shape perception (Ayzenberg, Chen, et al., 2019), participants may extract a more detailed skeleton if the task required subordinate-level categorization and fine-grained shape discriminations (Biederman & Shiffrar, 1987; Tarr & Bülthoff, 1995). In the current dissertation we sought to characterize the nature of participants 'default' skeletal representations, however, future work should explore how skeletal representations change with different types of object manipulations and different task demands.

What is a shape?

Shape is thought to be a fundamental property of object representations (Wagemans et al., 2008). Yet, there is very little agreement regarding the format of human shape representations. Indeed, proposals have ranged from veridical representations that are closely tied to the physical input on the retina, such as models based on contours and image-level properties (Tarr & Bülthoff,

1998), to mentally constructed representations that abstract away from the input, such as those based on component parts and non-accidental properties (NAPs; Biederman, 1987). In the current dissertation, I have argued for a constructed representation of shape based on an extracted skeletal structure. Unlike veridical representations, shape skeletons (particularly pruned algorithms) are tolerant to variations in contours and image-level properties. Indeed, in all three studies of my dissertation, human behavioral and neural responses were better described by a skeletal model than models based on image or contour properties (e.g., Gabor-jet; surface forms). Moreover, even though participants were largely tested with 2D images (Study 1 and Study 2), their responses were best fit by a 3D viewpoint-invariant skeletal model, further suggesting that human representations of shape are abstracted away from the physical stimulus. Our results also suggest that skeletal models are a better descriptor of shape representations than other constructed models, such as those based on component parts and a coarse spatial structure. Indeed, both adults and infants identified objects by their skeletons across differences in surface form, which changed the non-accidental properties of objects, as well as when coarse spatial relations were held constant. Thus, shape skeletons are a biologically plausible mechanism by which humans represent shape information and may offer a quantitative formalization of the oft poorly defined concept of global shape.

In the current dissertation, I tested whether humans represent a 3D skeletal structure rather than a 2D skeleton that is arguably easier to compute from an image (Trinh & Kimia, 2011). This decision was motivated by behavioral (Lowet et al., 2018) and neuroimaging work (Hung et al., 2012; Lescroart & Biederman, 2012) suggesting sensitivity to 3D skeletons in the primate visual system, as well as accumulating evidence that object perception (at least for novel objects) is best described by a 3D object-centered shape representation (Erdogan & Jacobs, 2017; Yamane et al., 2008). However, it remains unknown how a 3D skeletal structure arises from 2D images on the retina. One possibility is that skeletal computations in the visual system invoke generative shape

processes (Elder, Oleskiw, Yakubovich, & Peyré, 2013; Trinh & Kimia, 2007). These processes may be able to recover an object's 3D skeletal structure from retinal images by incorporating a small number of image-computable 2D skeletons (e.g., one from each eye; Qiu, Hatori, & Sakai, 2015). Alternatively, it is possible that an object's 3D structure may be recovered from a single image by first creating a representation based on depth properties and surface orientation, a so-called 2.5D sketch (Marr & Nishihara, 1978). Indeed, recent neural network models have been able to successfully reconstruct an object's 3D shape from single images by incorporating 2.5D sketches (Wu et al., 2017). Moreover, preliminary data from our lab suggest that input from depth perception regions of the parietal cortex are particularly important for recovering an object's 3D shape when its presented from rotated viewpoints (Ayzenberg, Kubert, Dilks, & Lourenco, in prep). Together, these studies suggest that monocular and binocular depth information may be particularly important for 3D skeleton generation.

Developmental origins of shape skeletons

One question raised by these findings is: how do skeletal representations arise in development? Although we found evidence of skeletal representations in infants as young as 6 months of age, it remains unclear whether infants are born with skeletal representations. One possibility is that children learn to rely on shape skeletons because they are a stable property of many object categories. That is, although image-level properties and component parts vary across views and exemplars, the shape skeleton remains constant and may come to serve as a privileged source of information. However, another possibility is that shape skeletons are an emergent property of early developing visual regions. More specifically, it is well known that early visual regions (i.e., V1-V3) are relatively mature at birth (Johnson, 2011; Kellman & Arterberry, 2006; Rakig, 1977). In Study 2, we found evidence of skeletal coding in V3, which may suggest that young infants are able to rely on mature V3 representations to categorize objects using the shape skeleton.

Thus, the early maturity of V3, and its ability to represent the shape skeleton, may help bootstrap object learning across development.

Yet, the results from Study 3 stand in stark contrast to other developmental work that suggests that global shape perception is experience dependent. For instance, many studies have suggested that even 8- and 9-year-old children have difficulty perceiving global shape and, instead, may focus on local object features (Davidoff & Roberson, 2002; Scherf, Behrmann, Kimchi, & Luna, 2009; Wakui et al., 2013). Moreover, the development of the 'shape bias' has been shown to depend on prior experience, such that children will begin to favor shape cues (over color and texture) only once they have sufficient linguistic experience with other categories (Landau, Smith, & Jones, 1998; Smith, 2003). How can these disparate findings be reconciled? One possibility is that these results reflect a performance-competence distinction. More specifically, many of the aforementioned studies rely on trials wherein global shape is placed in conflict with another cue (e.g., texture). To succeed on these trials, children must exhibit sufficient inhibitory control to avoid responding to the distracting cue – a feat which is difficult for children of this age (Logan, Schachar, & Tannock, 1997). Moreover, it's unclear whether studies linking object categorization to linguistic experience are measuring children's perceptual abilities to form categories, or their ability to link words to an object referent. Indeed, other work from our lab has found that 3-year-old children can readily recognize objects by their global shapes when inhibitory control and language demands are reduced (Ayzenberg et al., under review). Together, our findings suggest that the ability to categorize objects by a skeletal representation of global shape is present from early in development and may underlie later perceptual abilities.

A role for other visual properties

Although our findings suggest that skeletal descriptors play an important role in perceptual organization and object recognition, we would not argue that skeletal descriptions alone are sufficient. Humans do not perceive the visual environment simply as a collection of skeletons, but

rather, as complete objects where local contours, textures, and colors are integrated with a skeletal structure. Indeed, our results in Studies 1 and 2 showed that other models of vision were also predictive of participants' behavioral and neural responses to varying degrees. That other models were also predictive may be unsurprising, given that both shape- and non-shape-related properties are known to play important roles in object recognition. For instance, local contour information may be particularly useful for making subordinate-level category distinctions where pruned object skeletons of objects are roughly the same (Davitt et al., 2014; Hummel & Stankiewicz, 1996). Similarly, texture statistics and feature descriptions have been shown to be important indicators of both basic (Elder & Velisavljević, 2009; Ullman et al., 2016) and superordinate-level (Long et al., 2017; Long et al., 2018) object distinctions. Nevertheless, our work highlights the importance of formalized models of shape for object recognition, particularly the unique, and possibly privileged, role that skeletal structures may play.

Conclusion

How is it that, from early in development, humans are able to form robust and seemingly abstract representations of objects to support recognition? Using a combination of behavioral, neural, computational, and developmental methods we found that a model of structure known as the medial axis, or shape skeleton, can support the creation of robust shape representations, recognition, and rapid object learning. Together, these findings offer insights into the object processing mechanisms of the human mind and brain, and shed light on the developmental mechanisms that support mature perception and recognition.

References

Aguirre, G. K., Mattar, M. G., & Magis-Weinberg, L. (2011). de Bruijn cycles for neural decoding. *Neuroimage, 56*(3), 1293-1300. doi:https://doi.org/10.1016/j.neuroimage.2011.02.005

Amir, O., Biederman, I., & Hayworth, K. J. (2012). Sensitivity to nonaccidental properties across various shape dimensions. *Vision Research, 62*, 35-43. doi:10.1016/j.visres.2012.03.020

Arcaro, M. J., & Kastner, S. (2015). Topographic organization of areas V3 and V4 and its relation to supra-areal organization of the primate visual system. *Visual Neuroscience, 32*, E014. doi:10.1017/S0952523815000115

Ardila, D., Mihalas, S., Heydt, R. v. d., & Niebur, E. (2012, 21-23 March 2012). *Medial axis generation in a model of perceptual organization.* Paper presented at the 2012 46th Annual Conference on Information Sciences and Systems (CISS).

Ardila, D., Mihalas, S., von der Heydt, R., & Niebur, E. (2012). Medial axis generation in a model of perceptual organization. *Conference on Information Sciences and Systems (CISS)*, 1-4.

Ayzenberg, V., Chen, Y., Yousif, S. R., & Lourenco, S. F. (2019). Skeletal representations of shape in human vision: Evidence for a pruned medial axis model. *Journal of Vision, 19*(6), 1-21. doi:10.1167/19.6.6

Ayzenberg, V., Kamps, F. S., Dilks, D. D., & Lourenco, S. F. (2019). A dual role for shape skeletons in human vision: perceptual organization and object recognition. *bioRxiv*, 799650.

Ayzenberg, V., Kubert, J., Dilks, D. D., & Lourenco, S. F. (in prep). The dorsal stream faciliatates viewpoint-invariant object recognition.

Ayzenberg, V., & Lourenco, S. F. (2019a). The shape skeleton supports single exemplar categorization in infants. *PsyArxiv*.

Ayzenberg, V., & Lourenco, S. F. (2019b). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific reports, 9*(1), 1-13. doi:10.1038/s41598-019-45268-y

Ayzenberg, V., Sener, S. B., & Lourenco, S. F. (under review). Core object recognition in children: Mechanistic insights from neural networks.

Baker, N., & Kellman, P. J. (2018). Abstract shape representation in human visual perception. *Journal of Experimental Psychology: General, 147*(9), 1295-1308. doi:10.1037/xge0000409

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. (2018). Deep Convolutional Networks do not Make Classifications Based on Global Object Shape. *Journal of Vision, 18*(10), 904-904. doi:10.1167/18.10.904

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., . . . Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences of the United States of America, 103*(2), 449-454. doi:10.1073/pnas.0507062103

Barenholtz, E., & Tarr, M. J. (2006). Reconsidering the role of structure in vision. *Psychology of learning and motivation, 47*, 157-180.

Barenholtz, E., & Tarr, M. J. (2008). Visual judgment of similarity across shape transformations: Evidence for a compositional model of articulated objects. *Acta Psychologica, 128*(2), 331-338. doi:10.1016/j.actpsy.2008.03.007

Barense, M. D., Gaffan, D., & Graham, K. S. (2007). The human medial temporal lobe processes online representations of complex objects. *Neuropsychologia, 45*(13), 2963-2974.

Behrmann, M., Lee, A. C. H., Geskin, J. Z., Graham, K. S., & Barense, M. D. (2016). Temporal lobe contribution to perceptual function: A tale of three patient groups. *Neuropsychologia, 90*, 33-45. doi:https://doi.org/10.1016/j.neuropsychologia.2016.05.002

Behrmann, M., Peterson, M. A., Moscovitch, M., & Suzuki, S. (2006). Independent representation of parts and the relations between them: evidence from integrative agnosia. *Journal of Experimental Psychology: Human Perception and Performance, 32*(5), 1169-1184.

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

Biederman, I. (2000). Recognizing depth-rotated objects: A review of recent research and theory. *Spatial Vision, 13*(2), 241-253.

Biederman, I., & Bar, M. (1999). One-shot viewpoint invariance in matching novel objects. *Vision Research, 39*(17), 2885-2899.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19*(6), 1162-1182.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology, 20*(1), 38-64. doi:10.1016/0010-0285(88)90024-2

Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: a case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 640-645.

Biederman, I., Subramaniam, S., Bar, M., Kalocsai, P., & Fiser, J. (1999). Subordinate-level object classification reexamined. *Psychological research, 62*(2), 131-153. doi:10.1007/s004260050047

Blum, H. (1967). A transformation for extracting descriptors of shape. In W. Wathen-Dunn (Ed.), *Models for the Perception of Speech and Visual Form* (pp. 362-380). Cambridge, MA: MIT Press.

Blum, H. (1973). Biological shape and visual science (Part I). *Journal of Theoretical Biology, 38*(2), 205-287.

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLOS Computational Biology, 14*(4), e1006111. doi:10.1371/journal.pcbi.1006111

Bracci, S., & Op de Beeck, H. (2016). Dissociations and associations between shape and category representations in the two visual pathways. *Journal of Neuroscience, 36*(2), 432-444.

Brewer, A. A., Press, W. A., Logothetis, N. K., & Wandell, B. A. (2002). Visual areas in macaque cortex measured using functional magnetic resonance imaging. *Journal of Neuroscience, 22*(23), 10416-10426.

Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience, 7*, 880. doi:10.1038/nn1278

Brincat, S. L., & Connor, C. E. (2006). Dynamic Shape Synthesis in Posterior Inferotemporal Cortex. *Neuron, 49*(1), 17-24. doi:https://doi.org/10.1016/j.neuron.2005.11.026

Caplovitz, G. P., Barroso, D. J., Hsieh, P. J., & Tse, P. U. (2008). fMRI reveals that non-local processing in ventral retinotopic cortex underlies perceptual grouping by temporal synchrony. *Human Brain Mapping, 29*(6), 651-661.

Caplovitz, G. P., & Peter, U. T. (2010). Extrastriate cortical activity reflects segmentation of motion into independent sources. *Neuropsychologia, 48*(9), 2699-2708.

Chouinard, P. A., Whitwell, R. L., & Goodale, M. A. (2009). The lateral-occipital and the inferior-frontal cortex play different roles during the naming of visually presented objects. *Human Brain Mapping, 30*(12), 3851-3864. doi:doi:10.1002/hbm.20812

Cox, M. A., Schmid, M. C., Peters, A. J., Saunders, R. C., Leopold, D. A., & Maier, A. (2013). Receptive field focus of visual area V4 neurons determines responses to illusory surfaces. *Proceedings of the National Academy of Sciences, 110*(42), 17095-17100. doi:10.1073/pnas.1310806110

Craft, E., Schütze, H., Niebur, E., & von der Heydt, R. (2007). A Neural Model of Figure–Ground Organization. *Journal of Neurophysiology, 97*(6), 4310-4326. doi:10.1152/jn.00203.2007

Crouzet, S., & Serre, T. (2011). What are the Visual Features Underlying Rapid Object Recognition? *Frontiers in Psychology, 2*(326). doi:10.3389/fpsyg.2011.00326

Davidoff, J., & Roberson, D. (2002). Development of Animal Recognition: A Difference between Parts and Wholes. *Journal of Experimental Child Psychology, 81*(3), 217-234. doi:https://doi.org/10.1006/jecp.2002.2659

Davitt, L. I., Cristino, F., Wong, A. C. N., & Leek, E. C. (2014). Shape information mediating basic- and subordinate-level object recognition revealed by analyses of eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 40*(2), 451-456. doi:10.1037/a0034983

De Winter, J., & Wagemans, J. (2006). Segmentation of object outlines into parts: A large-scale integrative study. *Cognition, 99*(3), 275-325. doi:http://dx.doi.org/10.1016/j.cognition.2005.03.004

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

Dhandapani, R., & Kimia, B. B. (2002, 22-25 Sept. 2002). *Role of scale in partitioning shape.* Paper presented at the Proceedings. International Conference on Image Processing.

DiCarlo, James J., Zoccolan, D., & Rust, Nicole C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron, 73*(3), 415-434. doi:https://doi.org/10.1016/j.neuron.2012.01.010

Dilks, D. D., Julian, J. B., Kubilius, J., Spelke, E. S., & Kanwisher, N. (2011). Mirror-image sensitivity and invariance in object and scene processing pathways. *The Journal of Neuroscience, 31*(31), 11305-11312.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. *Science, 293*(5539), 2470-2473. doi:10.1126/science.1063414

Drucker, D. M., & Aguirre, G. K. (2009). Different Spatial Scales of Shape Similarity Representation in Lateral and Ventral LOC. *Cerebral Cortex, 19*(10), 2269-2280. doi:10.1093/cercor/bhn244

Dupont, P., De Bruyn, B., Vandenberghe, R., Rosier, A.-M., Michiels, J., Marchal, G., . . . Orban, G. (1997). The kinetic occipital region in human visual cortex. *Cerebral Cortex, 7*(3), 283-292.

Elder, J. H. (1999). Are Edges Incomplete? *International Journal of Computer Vision, 34*(2), 97-122. doi:10.1023/a:1008183703117

Elder, J. H. (2018). Shape from Contour: Computation and Representation. *Annual Review of Vision Science, 4*(1), 423-450. doi:10.1146/annurev-vision-091517-034110

Elder, J. H., Oleskiw, T. D., Yakubovich, A., & Peyré, G. (2013). On growth and formlets: Sparse multi-scale coding of planar shape. *Image and Vision Computing, 31*(1), 1-13. doi:10.1016/j.imavis.2012.11.002

Elder, J. H., & Velisavljević, L. (2009). Cue dynamics underlying rapid detection of animals in natural scenes. *Journal of Vision, 9*(7), 1-20. doi:10.1167/9.7.7

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

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

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

Felleman, D. J., & Van Essen, D. C. (1987). Receptive field properties of neurons in area V3 of macaque monkey extrastriate cortex. *Journal of Neurophysiology, 57*(4), 889-920.

Firestone, C., & Scholl, B. J. (2014). "Please tap the shape, anywhere you like" shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological Science, 25*(2), 377-386.

Freud, E., Culham, J. C., Plaut, D. C., & Behrmann, M. (2017). The large-scale organization of shape processing in the ventral and dorsal pathways. *eLife, 6*, e27576.

Gauthier, I., Hayward, W. G., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (2002). BOLD activity during mental rotation and viewpoint-dependent object recognition. *Neuron, 34*(1), 161-171.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research, 41*(10), 1409-1422.

Grill-Spector, K., Kushnir, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). Cue-invariant activation in object-related areas of the human occipital lobe. *Neuron, 21*(1), 191-202.

Grill-Spector, K., Kushnir, T., Hendler, T., & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience, 3*(8), 837-843.

Gross, C. G., Rodman, H. R., Cochin, P. M., & Colombot, M. W. (1993). *Inferior temporal cortex as a pattern recognition device.* Paper presented at the Computational Learning & Cognition: Proceedings of the Third NEC Research Symposium.

Harrison, S. J., & Feldman, J. (2009). The influence of shape and skeletal axis structure on texture perception. *Journal of Vision, 9*(6), 1-21. doi:10.1167/9.6.13

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100-108. doi:10.2307/2346830

Hatfield, M., McCloskey, M., & Park, S. (2016). Neural representation of object orientation: A dissociation between MVPA and Repetition Suppression. *Neuroimage, 139*, 136-148.

Hegdé, J., & Van Essen, D. C. (2006). A Comparative Study of Shape Representation in Macaque Visual Areas V2 and V4. *Cerebral Cortex, 17*(5), 1100-1116. doi:10.1093/cercor/bhl020

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition, 18*(1), 65-96. doi:https://doi.org/10.1016/0010-0277(84)90022-2

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 157-185). Hillsdale, NJ: Erlbaum.

Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition, 8*(3-5), 489-517. doi:10.1080/13506280143000214

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*(3), 480.

Hummel, J. E., & Stankiewicz, B. J. (1996). Categorical relations in shape perception. *Spatial Vision, 10*(3), 201-236.

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

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving, 7*(1), 2-9.

Johnson, S. P. (2011). Development of visual perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(5), 515-528. doi:10.1002/wcs.128

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology, 8*(1726), 1-18. doi:10.3389/fpsyg.2017.01726

Kanizsa, G. (1976). Subjective contours. *Scientific American, 234*(4), 48-52.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience, 22*(6), 974-983. doi:10.1038/s41593-019-0392-5

Kayaert, G., Biederman, I., & Vogels, R. (2003). Shape Tuning in Macaque Inferior Temporal Cortex. *The Journal of Neuroscience, 23*(7), 3016-3027. doi:10.1523/jneurosci.23-07-03016.2003

Keefe, B. D., Gouws, A. D., Sheldon, A. A., Vernon, R. J., Lawrence, S. J., McKeefry, D. J., . . . Morland, A. B. (2018). Emergence of symmetry selectivity in the visual areas of the human brain: fMRI

responses to symmetry presented in both frontoparallel and slanted planes. *Human Brain Mapping, 39*(10), 3813-3826.

Kellman, P. J., & Arterberry, M. E. (2006). Infant visual perception. *Handbook of child psychology*.

Kibbe, M. M., & Leslie, A. M. (2019). Conceptually Rich, Perceptually Sparse: Object Representations in 6-Month-Old Infants' Working Memory. *Psychological Science*, 1-14.

Kimia, B. B. (2003). On the role of medial geometry in human vision. *Journal of Physiology-Paris, 97*(2), 155-190.

Konen, Christina S., Behrmann, M., Nishimura, M., & Kastner, S. (2011). The functional neuroanatomy of object agnosia: A case study. *Neuron, 71*(1), 49-60.

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

Kovacs, I., & Julesz, B. (1994). Perceptual sensitivity maps within globally defined visual shapes. *Nature, 370*(6491), 644-646.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*(4). doi:10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097-1105.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLOS Computational Biology, 12*(4), e1004896. doi:10.1371/journal.pcbi.1004896

Landau, B., Smith, L., & Jones, S. (1998). Object perception and object naming in early development. *Trends in Cognitive Sciences, 2*(1), 19-24.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*(3), 299-321.

Lee, T. S. (1996). *Neurophysiological evidence for image segmentation and medial axis computation in primate V1.* Paper presented at the Computation and Neural Systems: Proceedings of the Fourth Annual Computational Neuroscience Conference.

Lescroart, M. D., & Biederman, I. (2012). Cortical representation of medial axis structure. *Cerebral Cortex, 23*(3), 629-637.

Lescroart, M. D., Stansbury, D. E., & Gallant, J. L. (2015). Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. *Frontiers in Computational Neuroscience, 9*(135), 1-20. doi:10.3389/fncom.2015.00135

Leyton, M. (1989). Inferring Causal History froms Shape. *Cognitive Science, 13*(3), 357-387. doi:10.1207/s15516709cog1303_2

Li, Z. (2000). Can V1 mechanisms account for figure–ground and medial axis effects. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems* (Vol. 12, pp. 136-142): MIT Press.

Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., & Serre, T. (2017). *What are the visual features underlying human versus machine vision?* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Liu, T.-L., & Geiger, D. (1999). Approximate tree matching and shape similarity. *Proceedings of the Seventh IEEE International Conference on Computer Vision, 1*, 456-462.

Logan, G. D., Schachar, R. J., & Tannock, R. (1997). Impulsivity and inhibitory control. *Psychological Science, 8*(1), 60-64.

Long, B., Störmer, V. S., & Alvarez, G. A. (2017). Mid-level perceptual features contain early cues to animacy. *Journal of Vision, 17*(6), 1-20. doi:10.1167/17.6.20

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences, 115*(38), E9015-E9024. doi:10.1073/pnas.1719616115

Lowet, A. S., Firestone, C., & Scholl, B. J. (2018). Seeing structure: Shape skeletons modulate perceived similarity. *Attention, Perception, & Psychophysics, 80*(5), 1278-1289. doi:10.3758/s13414-017-1457-8

Mannion, D. J., McDonald, J. S., & Clifford, C. W. (2010). The influence of global form on local orientation anisotropies in human visual cortex. *Neuroimage, 52*(2), 600-605.

Margalit, E., Biederman, I., Herald, S. B., Yue, X., & von der Malsburg, C. (2016). An applet for the Gabor similarity scaling of the differences between complex stimuli. *Attention, Perception, & Psychophysics, 78*(8), 2298-2306. doi:10.3758/s13414-016-1191-7

Margalit, E., Biederman, I., Tjan, B. S., & Shah, M. P. (2017). What is actually affected by the scrambling of objects when localizing the lateral occipital complex? *Journal of Cognitive Neuroscience, 29*(9), 1595-1604.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences, 200*(1140), 269-294.

Martin, A. B., & von der Heydt, R. (2015). Spike synchrony reveals emergence of proto-objects in visual cortex. *Journal of Neuroscience, 35*(17), 6860-6870.

Mash, C., Arterberry, M. E., & Bornstein, M. H. (2007). Mechanisms of visual object tecognition in infancy: Five-month-olds generalize beyond the interpolation of familiar views. *Infancy, 12*(1), 31-43. doi:10.1111/j.1532-7078.2007.tb00232.x

McMains, S. A., & Kastner, S. (2010). Defining the units of competition: influences of perceptual organization on competitive interactions in human visual cortex. *Journal of Cognitive Neuroscience, 22*(11), 2417-2426.

Mehrani, P., & Tsotsos, J. K. (2019). Early recurrence enables figure border ownership. *arXiv preprint arXiv:1901.03201*.

Montaser-Kouhsari, L., Landy, M. S., Heeger, D. J., & Larsson, J. (2007). Orientation-selective adaptation to illusory contours in human visual cortex. *Journal of Neuroscience, 27*(9), 2186-2195.

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences, 99*(23), 15164-15169.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & quantity, 41*(5), 673-690.

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision, 42*(3), 145-175. doi:10.1023/A:1011139631724

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in brain research, 155*, 23-36.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607-609.

Op de Beeck, H. P., Torfs, K., & Wagemans, J. (2008). Perceived Shape Similarity among Unfamiliar Objects and the Organization of the Human Object Vision Pathway. *The Journal of Neuroscience, 28*(40), 10111-10123. doi:10.1523/jneurosci.2511-08.2008

Parker, A. J. (2007). Binocular depth perception and the cerebral cortex. *Nature Reviews Neuroscience, 8*(5), 379-391. doi:10.1038/nrn2131

Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience, 5*(12), 1332-1338.

Peelen, M. V., & Downing, P. E. (2005). Selectivity for the Human Body in the Fusiform Gyrus. *Journal of Neurophysiology, 93*(1), 603-608. doi:10.1152/jn.00513.2004

Poort, J., Raudies, F., Wannig, A., Lamme, V. A., Neumann, H., & Roelfsema, P. R. (2012). The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. *Neuron, 75*(1), 143-156.

Psotka, J. (1978). Perceptual processes that may create stick figures and balance. *Journal of Experimental Psychology: Human Perception and Performance, 4*(1), 101-111.

Qiu, W., Hatori, Y., & Sakai, K. (2015). Neural construction of 3D medial axis from the binocular fusion of 2D MAs. *Neurocomputing, 149, Part B*, 546-558. doi:10.1016/j.neucom.2014.08.019

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *The Journal of Neuroscience, 38*(33), 7255-7269. doi:10.1523/jneurosci.0388-18.2018

Rakig, P. (1977). Prenatal development of the visual system in rhesus monkey. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 278*(961), 245-260.

Rezanejad, M., Downs, G., Wilder, J., Walther, D. B., Jepson, A., Dickinson, S., & Siddiqi, K. (2019). *Scene Categorization from Contours: Medial Axis Based Salience Measures.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Rosenfeld, A., Zemel, R., & Tsotsos, J. K. (2018). The elephant in the room. *arXiv preprint arXiv:1808.03305*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211-252.

Sanocki, T. (1993). Time course of object identification: Evidence for a global-to-local contingency. *Journal of Experimental Psychology: Human Perception and Performance, 19*(4), 878-898. doi:10.1037/0096-1523.19.4.878

Sasaki, Y. (2007). Processing local signals into global patterns. *Current Opinion in Neurobiology, 17*(2), 132-139.

Sasaki, Y., Vanduffel, W., Knutsen, T., Tyler, C., & Tootell, R. (2005). Symmetry activates extrastriate visual cortex in human and nonhuman primates. *Proceedings of the National Academy of Sciences of the United States of America, 102*(8), 3159-3163. doi:10.1073/pnas.0500319102

Scherf, K. S., Behrmann, M., Kimchi, R., & Luna, B. (2009). Emergence of Global Shape Processing Continues Through Adolescence. *Child Development, 80*(1), 162-177. doi:10.1111/j.1467-8624.2008.01252.x

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., . . . DiCarlo, J. J. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*. doi:10.1101/407007

Sebastian, T. B., Klein, P. N., & Kimia, B. B. (2004). Recognition of shapes by editing their shock graphs. *IEEE Transactions on pattern Analysis and Machine Intelligence, 26*(5), 550-571.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, 104*(15), 6424-6429. doi:10.1073/pnas.0700622104

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis & Machine Intelligence*(3), 411-426.

Shaked, D., & Bruckstein, A. M. (1998). Pruning medial axes. *Computer Vision and Image Understanding, 69*(2), 156-169. doi:10.1006/cviu.1997.0598

Shokoufandeh, A., Macrini, D., Dickinson, S., Siddiqi, K., & Zucker, S. W. (2005). Indexing hierarchical structures using graph spectra. *IEEE Transactions on pattern Analysis and Machine Intelligence, 27*(7), 1125-1140. doi:10.1109/TPAMI.2005.142

Singh, M., Seyranian, G. D., & Hoffman, D. D. (1999). Parsing silhouettes: The short-cut rule. *Perception & Psychophysics, 61*(4), 636-660. doi:10.3758/bf03205536

Slone, L. K., Smith, L. B., & Yu, C. (2019). Self-generated variability in object images predicts vocabulary growth. *Developmental Science, 0*(0), e12816. doi:10.1111/desc.12816

Smith, L. B. (2003). Learning to recognize objects. *Psychological Science, 14*(3), 244-250.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., . . . Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage, 23*, S208-S219. doi:https://doi.org/10.1016/j.neuroimage.2004.07.051

Spröte, P., & Fleming, R. W. (2016). Bent out of shape: The visual inference of non-rigid shape transformations applied to objects. *Vision Research, 126*, 330-346. doi:10.1016/j.visres.2015.08.009

Spröte, P., Schmidt, F., & Fleming, R. W. (2016). Visual perception of shape altered by inferred causal history. *Scientific reports, 6*, 1-11. doi:10.1038/srep36245

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv*.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., . . . Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences, 115*(35), 8835-8840. doi:10.1073/pnas.1719397115

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., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. *Cognition, 67*(1), 1-20. doi:http://dx.doi.org/10.1016/S0010-0277(98)00026-2

Trinh, N. H., & Kimia, B. B. (2007). A Symmetry-Based Generative Model for Shape. *11th International Conference on Computer Vision*, 1-8. doi:10.1109/ICCV.2007.4409022

Trinh, N. H., & Kimia, B. B. (2011). Skeleton search: Category-specific object recognition and segmentation using a skeletal shape model. *International Journal of Computer Vision, 94*(2), 215-240.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences, 113*(10), 2744-2749.

Van Dromme, I. C., Premereur, E., Verhoef, B.-E., Vanduffel, W., & Janssen, P. (2016). Posterior Parietal Cortex Drives Inferotemporal Activations During Three-Dimensional Object Vision. *PLOS Biology, 14*(4), e1002445. doi:10.1371/journal.pbio.1002445

Van Meel, C., Baeck, A., Gillebert, C. R., Wagemans, J., & Op de Beeck, H. P. (2019). The representation of symmetry in multi-voxel response patterns and functional connectivity throughout the ventral visual stream. *Neuroimage, 191*, 216-224.

Vogels, R., Biederman, I., Bar, M., & Lorincz, A. (2001). Inferior temporal neurons show greater sensitivity to nonaccidental than to metric shape differences. *Journal of Cognitive Neuroscience, 13*(4), 444-453.

von der Heydt, R. (2015). Figure–ground organization and the emergence of proto-objects in the visual cortex. *Frontiers in Psychology, 6*(1695). doi:10.3389/fpsyg.2015.01695

Wagemans, J., De Winter, J., de Beeck, H. O., Ploeger, A., Beckers, T., & Vanroose, P. (2008). Identification of everyday objects on the basis of silhouette and outline versions. *Perception, 37*(2), 207-244.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin, 138*(6), 1172-1217.

Wakui, E., Jüttner, M., Petters, D., Kaur, S., Hummel, J. E., & Davidoff, J. (2013). Earlier development of analytical than holistic object recognition in adolescence. *PLoS ONE, 8*(4), e61041. doi:10.1371/journal.pone.0061041

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage, 137*, 188-200.

Wang, L., Mruczek, R. E. B., Arcaro, M. J., & Kastner, S. (2014). Probabilistic Maps of Visual Topography in Human Cortex. *Cerebral Cortex, 25*(10), 3911-3931. doi:10.1093/cercor/bhu277

Welchman, A. E. (2016). The human brain in depth: how we see in 3D. *Annual Review of Vision Science, 2*, 345-376.

Wieser, E., Seidl, M., & Zeppelzauer, M. (2017). A study on skeletonization of complex petroglyph shapes. *Multimedia Tools and Applications, 76*(6), 8285-8303. doi:10.1007/s11042-016-3395-1

Wilder, J., Feldman, J., & Singh, M. (2011). Superordinate shape classification using natural shape statistics. *Cognition, 119*(3), 325-340. doi:10.1016/j.cognition.2011.01.009

Wilder, J., Rezanejad, M., Dickinson, S., Siddiqi, K., Jepson, A., & Walther, D. B. (2019). Local contour symmetry facilitates scene categorization. *Cognition, 182*, 307-317. doi:https://doi.org/10.1016/j.cognition.2018.09.014

Wokke, M. E., Vandenbroucke, A. R., Scholte, H. S., & Lamme, V. A. (2013). Confuse your illusion: feedback to early visual cortex contributes to perceptual completion. *Psychological Science, 24*(1), 63-71.

Wood, J. N. (2013). Newborn chickens generate invariant object representations at the onset of visual object experience. *Proceedings of the National Academy of Sciences, 110*(34), 14000-14005.

Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, B., & Tenenbaum, J. (2017). Marrnet: 3d shape reconstruction via 2.5 d sketches. *Advances in Neural Information Processing Systems*, 540-550.

Wu, K., Wu, E., & Kreiman, G. (2018). *Learning scene gist with convolutional neural networks to improve object recognition.* Paper presented at the 2018 52nd Annual Conference on Information Sciences and Systems (CISS).

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., & Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience, 11*, 1352-1360. doi:10.1038/nn.2202

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

Yue, X., Biederman, I., Mangini, M. C., Malsburg, C. v. d., & Amir, O. (2012). Predicting the psychophysical similarity of faces and non-face complex shapes by image-based measures. *Vision Research, 55*, 41-46. doi:https://doi.org/10.1016/j.visres.2011.12.012

Yue, X., Biederman, I., Mangini, M. C., von der Malsburg, C., & Amir, O. (2012). Predicting the psychophysical similarity of faces and non-face complex shapes by image-based measures. *Vision Research, 55*, 41-46.

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications, 10*(1), 3770. doi:10.1038/s41467-019-11786-6

Zhou, H., Friedman, H. S., & von der Heydt, R. (2000). Coding of Border Ownership in Monkey Visual Cortex. *The Journal of Neuroscience, 20*(17), 6594-6611. doi:10.1523/jneurosci.20-17-06594.2000

Zoccolan, D., Oertelt, N., DiCarlo, J. J., & Cox, D. D. (2009). A rodent model for the study of invariant visual object recognition. *Proceedings of the National Academy of Sciences, 106*(21), 8748-8753.