Attention, Perception, & Psychophysics (2018) 80:1278–1289
https://doi.org/10.3758/s13414-017-1457-8

Seeing structure: Shape skeletons modulate perceived similarity

Adam S. Lowet[1] · Chaz Firestone[1,2] · Brian J. Scholl[1]

Published online: 15 March 2018

Abstract
An intrinsic part of seeing objects is seeing how similar or different they are relative to one another. This experience requires that objects be mentally represented in a common format over which such comparisons can be carried out. What is that representational format? Objects could be compared in terms of their superficial features (e.g., degree of pixel-by-pixel overlap), but a more intriguing possibility is that they are compared on the basis of a deeper structure. One especially promising candidate that has enjoyed success in the computer vision literature is the *shape skeleton*—a geometric transformation that represents objects according to their inferred underlying organization. Despite several hints that shape skeletons are computed in human vision, it remains unclear how much they actually *matter* for subsequent performance. Here, we explore the possibility that shape skeletons help mediate the ability to extract visual similarity. Observers completed a same/different task in which two shapes could vary either in their skeletal structure (without changing superficial features such as size, orientation, and internal angular separation) or in large surface-level ways (without changing overall skeletal organization). Discrimination was better for skeletally dissimilar shapes: observers had difficulty appreciating even surprisingly large differences when those differences did not reorganize the underlying skeletons. This pattern also generalized beyond line drawings to 3-D volumes whose skeletons were less readily inferable from the shapes' visible contours. These results show how shape skeletons may influence the perception of similarity—and more generally, how they have important consequences for downstream visual processing.

Keywords Shape skeletons · Medial axis · Visual similarity

Objects in the world routinely strike us as being *similar* or *dissimilar* to one another, or to themselves at different times. Indeed, comparisons of this sort are often crucial in everyday life, as when we judge that a novel object belongs to an existing category, or when we determine whether a given object is one that we have seen before. This capacity is especially critical for objects that can take on multiple visually distinct configurations, such as an animal that may assume different postures or a man-made artifact with movable parts (e.g., a collapsible umbrella or a folding chair).

The ability to compare objects in this way requires (almost by definition) that the mind represent objects in a common language or *format* that could enable these comparisons. What is this format, such that we can compare individual objects across time and space?

Superficial features versus deeper structure

In order to determine how similar or different two objects are, the mind could compare them using a variety of different approaches, which have classically fallen into two categories. One approach prioritizes superficial features of the objects, such as their visible contours. For example, the visual system could align the two objects as best as possible and then calculate the degree of pixel-by-pixel overlap between them, or assess the degree to which they share similar features—such as the extents of their spatial envelopes, the lengths of their perimeters, or the angles between edges. Such approaches have been posited to explain aspects of object recognition in human vision (e.g., Corballis, 1988; Tarr & Pinker, 1989; Ullman, 1989) and have also been implemented in limited ways in computer vision systems (e.g., Bolles & Cain, 1982; Cortese & Dyre, 1996; Ferrari, Jurie, & Schmid, 2010; Zhang & Lu, 2002).

✉ Brian J. Scholl
brian.scholl@yale.edu

[1] Department of Psychology, Yale University, Box 208205, New Haven, CT 06520-8205, USA

[2] Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, USA

However, a prominent difficulty with such approaches is that they are unable to categorize objects as being the *same* when they fail to share the relevant superficial features. Consider, for example, a human hand: Such an "object" can assume a variety of different poses—from a clenched fist to a precision grip to an extended wave—but feature-based approaches will be frustrated by such transformations and may too readily conclude that they are entirely different objects (see Fig. 1a).

An alternative approach, then, is to represent shapes at a deeper level of organization—one that remains constant over these sorts of transformations to an object's global shape. Such approaches infer an object's underlying *structure* and take advantage of invariant relationships between parts of that structure when computing similarity (e.g., in the geons of Biederman, 1987; Hummel & Biederman, 1992; the part structure based on curvature minima of Hoffman & Richards, 1984; and the generalized cylinders of Marr & Nishihara, 1978), in a way that may mirror the organization of the object itself. For example, just as the literal parts of a hand–its bones and joints–remain connected to one another in the same way regardless of the hand's pose, an object's inferred interior structure would remain similarly invariant over such transformations. Thus, if two objects share the same underlying structure, they can be represented as such.

Shape skeletons

An especially intriguing candidate for this underlying structure is the *shape skeleton*, a geometric transformation that defines such a structure in terms of an object's local symmetry axes. The shape skeleton is typically formalized as the set of all points equidistant from two or more points on a shape's boundary—a construct known as the "medial axis" (Blum, 1973; for related definitions, see Aichholzer, Aurenhammer, Alberts, & Gärtner, 1995; Feldman & Singh, 2006; Serra, 1986). For some simple shapes (such as the triangle in Fig. 2a), this definition merely picks out the global symmetry axes themselves, but for others (such as the rectangle in Fig. 2b) it picks out more complex collections of points.

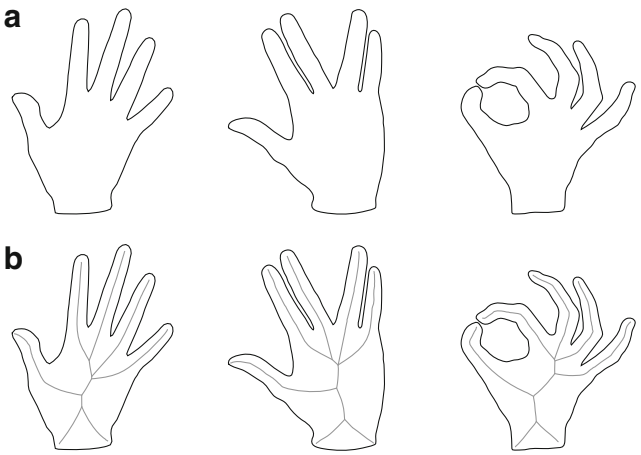

Fig. 1 **a** Everyday objects, such as the human hand, can assume many poses, each with a very different global shape. **b** Despite such differences in global shape, the internal structure of the hand remains the same across transformations, and the organization of the shape skeleton reflects this invariance (Feldman & Singh, 2006; skeletons computed using Jacob Feldman and Manish Singh's ShapeToolbox 1.0)

Further analysis can then group this collection of points into a hierarchical structure, with some points being localized onto more peripheral "branches" that are seen as stemming from a more central "trunk". This yields a representation that emphasizes the underlying connectivity and topology of a shape (as explored by Feldman & Singh, 2006), in a way that overcomes the challenge of comparing superficially different forms and allows for objects to be recognized as similar even when contextual factors, such as the object's orientation or the angle of viewing, are dramatically different. Moreover, when applied to images of natural or everyday objects (as in Fig. 1b), the shape skeleton tends to capture an object's essential part structure—and, in the case of a living thing, may even respect biomechanical constraints, such as the possible articulations of its limbs. For these sorts of reasons, representing and comparing objects based on their shape skeletons—in a way that can directly fuel similarity judgments—has enjoyed considerable success as an object recognition strategy in computer vision systems (e.g., Bai & Latecki, 2008; Liu & Geiger, 1999; Sebastian & Kimia, 2005; Siddiqi, Shokoufandeh, Dickinson, & Zucker, 1999; Torsello & Hancock, 2004; Trinh & Kimia, 2011; Zhu & Yuille, 1996; for a review, see Siddiqi & Pizer, 2008).

The success of skeletal shape representations in computer vision has raised the possibility that human vision, too, has converged on this solution for representing and recognizing objects (Kimia, 2003). And indeed, there are several studies suggesting that such skeletal representations exist in visual processing (e.g., Harrison & Feldman, 2009; Kovacs, Feher, & Julesz, 1998; Kovacs & Julesz, 1994). For example, if many people are simply asked to tap a shape (presented on a tablet computer) once with their finger, wherever they wish, the collective patterns of taps conform to the medial axis, as depicted in Figs. 2c–d (Firestone & Scholl, 2014; see also Psotka, 1978). And, beyond documenting their existence, other work has suggested that such skeletal representations might also influence certain types of higher-level subjective judgments—such as how aesthetically pleasing a shape (or even a real-world structure such as a rock garden) is (Palmer & Guidi, 2011; van Tonder, Lyons, & Ejima, 2002), or what that shape should be called (Wilder, Feldman, & Singh, 2011).

The current project seeks to build on this previous research, with a novel empirical focus on shape skeletons and perceived similarity. Whereas previous work has explored possible *theoretical* connections between shape

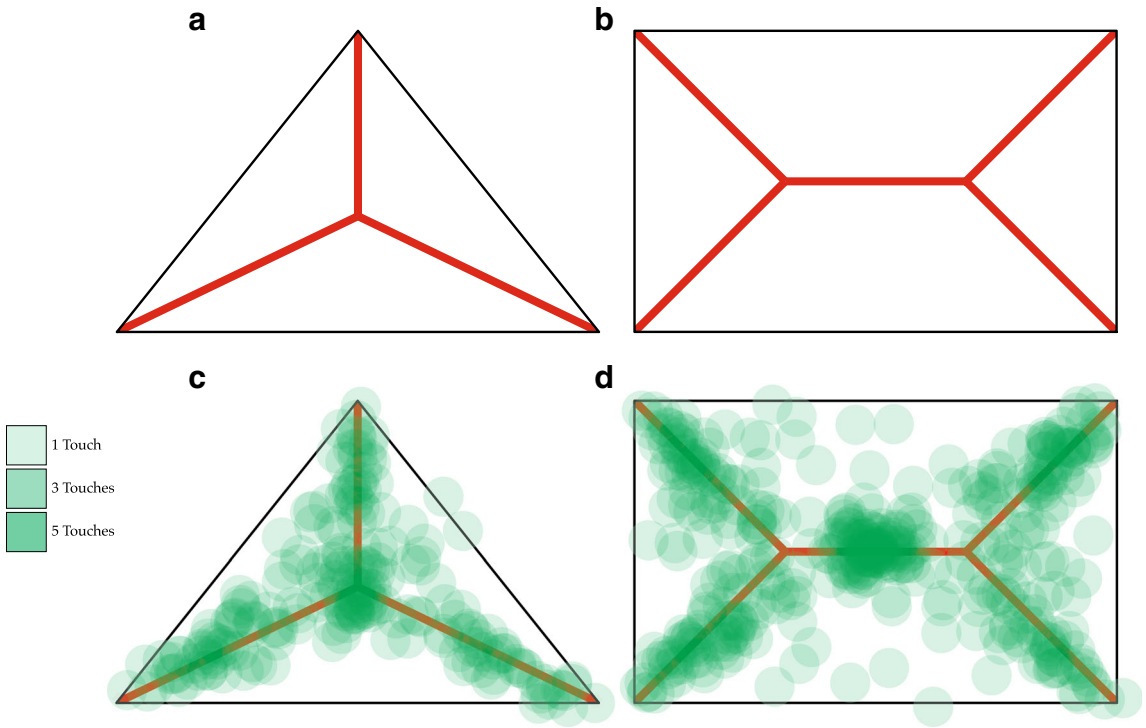

Fig. 2 The medial axis, depicted for (**a**) a triangle and (**b**) a rectangle. **c** and **d** When subjects were asked to tap these shapes once, anywhere they chose, the aggregated touches clustered around the medial axes, revealing the psychological reality of shape skeletons (Firestone & Scholl, 2014, Experiments 1 and 2). (Color figure online)

skeletons and similarity judgments, we aim to collect *empirical* data on this possibility. Whereas quite a lot of previous work has explored the possibility that similarity judgments could be explained by appeal to more *general* types of structural shape representations (e.g., geons, generalized cylinders, part structure based on curvature minima), we aim to forge a novel empirical connection between perceived similarity and shape skeletons, per se. And whereas past work has explored the possible psychological reality of shape skeletons in several ways, we aim to do so for the first time in the context of visual similarity.

The current project: Shape skeletons and visual similarity

Might shape skeletons actually influence our objective ability to correctly identify whether two shapes are the same or different? In the present study, to find out, observers were repeatedly shown pairs of shapes—both simultaneously visible, side-by-side (see Fig. 3). The two shapes in each pair, when they differed, could do so either in their underlying skeletal structure (without changing superficial features such as size, orientation, and internal angular separation) or in larger

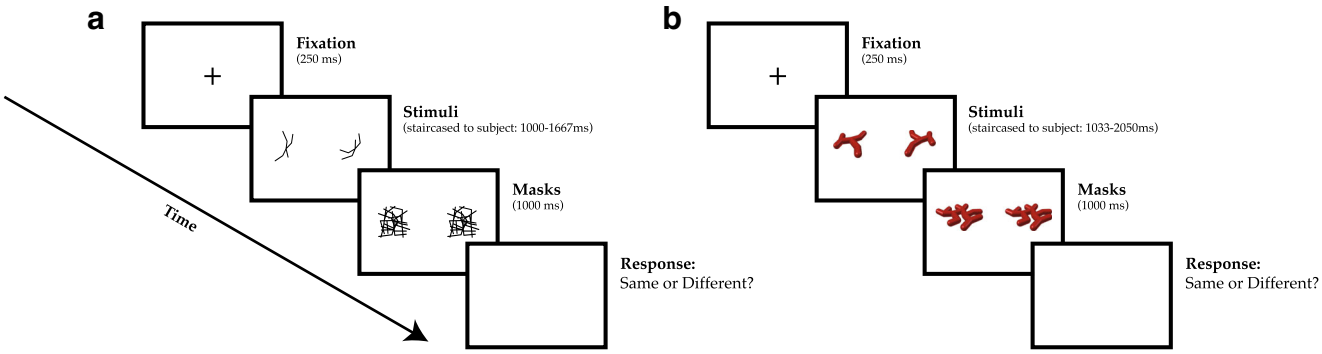

Fig. 3 Schematic illustrations of the same/different tasks in (**a**) Experiment 1 and (**b**) Experiment 2. In both experiments, each trial began with a 250-ms presentation of a black fixation cross centered on a white background. Two shapes from the same "family" then appeared side by side on the screen for the staircased duration, followed by masks, which appeared for 1 s. Subjects indicated with a keypress whether they thought the two shapes were the same or different. Following a 500-ms pause, the next trial began. (Color figure online)

surface-level ways (without changing the overall skeletal organization). Observers simply had to determine on each trial whether the two shapes were identical or not (disregarding orientation)—and on trials where the shapes differed, their performance could be correlated with the actual degrees of skeletal similarity versus superficial similarity. If performance is predicted better by skeletal similarity than by superficial similarity, then this would be evidence that shape skeletons don't merely exist in the mind but actually modulate objective visual performance.

Experiment 1: Skeletal or featural similarity?

We first contrasted skeletal versus featural similarity using stick-figure shapes (see Fig. 4) that made skeletal structure (a) especially apparent, (b) unambiguous (in that every computational approach to shape skeletons would output the same skeleton for these shapes), and (c) dissociable from the shapes' lower-level features. These stick-figures thus intuitively resemble the skeletons themselves, but this in no way entailed that similarity perception would be best explained by the underlying skeletal structure. As we explore in detail, these figures also differ in a host of lower-level ways (such as the different areas of their convex hulls) that are not confounded with skeletal structure and that could in principle influence similarity perception to an even greater extent. For this reason, we take care to demonstrate the explanatory power of shape skeletons over and above a variety of lower-level features that could account for similarity perception without making any reference to skeletal structure.

Method

Subjects Ten naïve observers (with normal or corrected-to-normal acuity) from the Yale community completed individual 30-minute sessions in exchange for a small monetary payment. This sample size was determined a priori based on previous experiments of this sort (e.g., Barenholtz, Cohen, Feldman, & Singh, 2003), and was the same for both experiments reported here.

Apparatus The experiment was conducted with custom software written in Python with the PsychoPy libraries (Peirce, 2007). The observers sat approximately 60 cm (without restraint) from a 36.4 × 27.3 cm CRT display with a 60 Hz refresh rate.

Stimuli The stimulus set consisted of 400 families of six shapes each. The shapes in each family were derived from a Parent shape composed of four branches emanating from a central node (see Fig. 4). Three of the branches—referred to below as "arms"—had two segments connected at a joint. The remaining branch—referred to below as the "stub"—had only one segment. Each segment of the Parent shape was 0.24 cm wide and capped by a semicircle with a diameter of 0.24 cm such that the segment terminated smoothly. (The distance measurements that follow do not include the contribution of

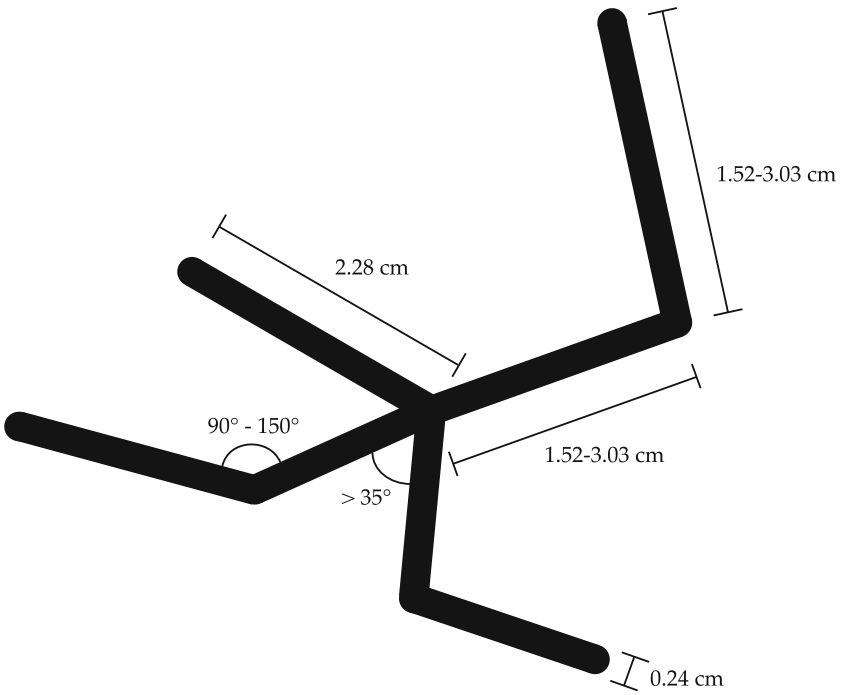

Fig. 4 Schematic illustration of stimuli used in Experiment 1. Values displayed for one branch apply also to every corresponding branch

this cap and are always computed from the center of any given node or joint.) The length of each arm segment was a randomly chosen value between 1.52 and 3.03 cm. The inner arm segments were separated by different randomly chosen angles (of at least 35°) for each Parent shape. Each outer arm segment was oriented at a randomly chosen angle (between 90° and 150°) relative to its inner arm segment. The stub was always 2.28 cm long and was inserted between the two arms with the largest gap between them, at a random orientation no less than 35° from either arm.

To construct the other five members of each shape family from its Parent shape, we modified the Parent shape in five distinct ways, as depicted in Fig. 5. The first four of these changes were merely featural, and did not alter the shape's skeletal structure: (1) In *Single Arm* changes, the outer segment of a single randomly chosen arm pivoted by a randomly chosen angle between 45° and 90°; (2) in *Stub* changes, the stub pivoted by a randomly chosen angle of at least 45°; (3) in *Arms* changes, the outer segment of each arm pivoted by a different randomly chosen angle between 45° and 90°; and (4) in *Arms + Stub* changes, both the Stub changes and Arms changes were combined. The final type of change manipulated the shape's skeletal structure while minimizing perturbations to the shape's other features: (5) In *Skeletal* changes, the base of the stub translated from the central node to a randomly chosen joint without changing the stub's orientation. The crucial aspect of these modifications is that although Skeletal changes are the only ones that reorganize the shape skeleton, these changes were actually quite small in terms of number of pixels that moved: indeed, many fewer pixels change in Skeletal modifications compared to Arms and Arms + Stub modifications—and a

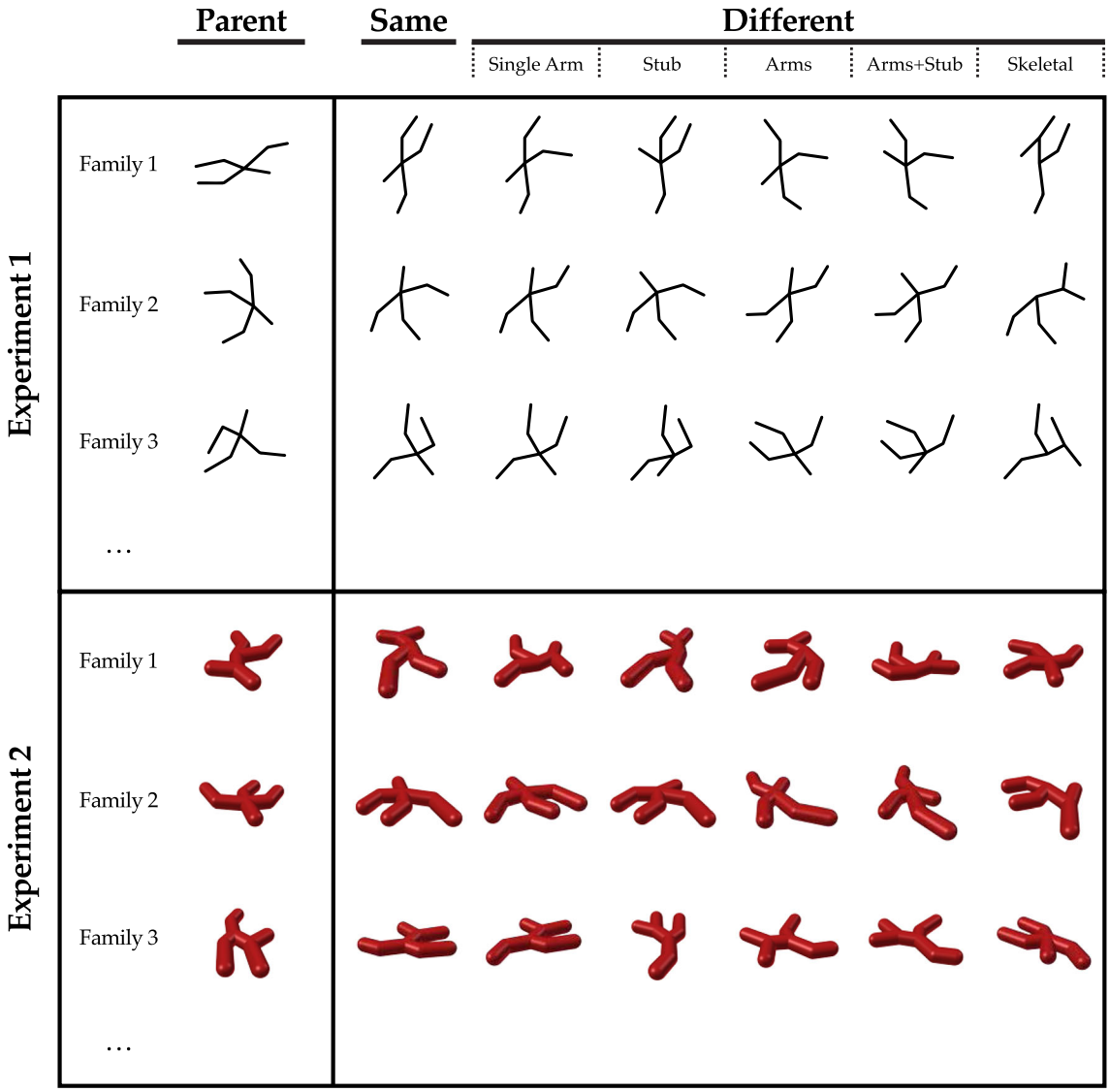

Fig. 5 Sample stimuli used in Experiments 1 and 2. Each row represents a different shape family (see text for details)

roughly equal number of pixels move compared to Single Arm and Stub modifications.

Every modified shape in every family had the same aggregate arm length of 13.65 cm, each endpoint was no less than 0.73 cm away from its nearest neighbor, and no two branches intersected (except, of course, at the central node).

Procedure As outlined in Fig. 3, each trial began with a 250-ms presentation of a black fixation cross (5.0 × 5.0 cm) centered on a white background. Two shapes then appeared side-by-side on the screen (for a duration that was staircased for each subject as detailed below), with each shape's central node centered in its half of the display. One of the shapes was always a Parent, presented in a random orientation and on a randomly chosen side of the display. Half of the trials were *Same* trials, on which the other shape was also that Parent shape, presented in a new random orientation that differed by at least at 90° from the first shape. The other half of the trials were *Different* trials, in which the second shape was drawn from one of the five types of shape modifications and appeared in a new random orientation that differed by at least at 90° from the first shape.

On all trials, the shapes were replaced by masks (consisting of overlapping line segments designed to mimic the low-level features of the shapes and occupy roughly the same area; see Fig. 3a), which remained present for 1 s. On each trial, subjects indicated by a keypress (that could be made as soon as the masks appeared) whether they thought the two shapes were identical or not (disregarding orientation). After a 500-ms pause, the next trial began.

Subjects completed 400 trials (each based on a different Parent shape—with a different random assignment of Parent shapes to each trial computed for each subject), of which 200 were Same trials and 200 were Different trials. The Different trials consisted of 40 trials of each of the five types of shape modifications (Single Arm, Stub, Arms, Arms + Stub, Skeletal). The trials were divided into five 80-trial blocks, with a self-paced rest period between each block. Each block contained 40 Same trials and 40 Different trials, with eight trials each of the five different types of shape modifications—all presented in a different random order for each subject.

Subjects first completed three practice trials followed by a staircasing procedure that manipulated stimulus presentation duration to bring accuracy to 70%. (An initial presentation duration of 1,500 ms was reduced whenever subjects answered two consecutive trials correctly and was increased whenever they answered a single trial incorrectly. The decrements and increments themselves decreased over the course of staircasing from 83 ms to 33 ms. The staircasing procedure continued until the subject reached a minimum of 15 trials and three reversals.) After the session, each subject completed a funneled debriefing procedure during which they were asked about their experiences and about any particular strategies that they had employed.

Results and discussion

As can be seen in Fig. 6a, changes to the shape's underlying skeleton were easier to detect than any other kind of change. These impressions were verified by the following analyses. A repeated-measures ANOVA revealed a main effect of shape modification type, $F(4, 36) = 17.545$, $p < .001$, $\eta = .661$, and planned comparisons confirmed that Skeletal differences (86.3%) were detected better than any of the surface-feature changes that did not change skeletal structure: Single Arm, 51.8%, $t(9) = 6.23$, $p < .001$; Stub, 64.0%, $t(9) = 4.37$, $p = .002$; Arms, 67.8%, $t(9) = 4.53$, $p = .001$; and Arms + Stub, 75.5%, $t(9) = 4.12$, $p = .003$. This effect was also reliable nonparametrically: Every single observer performed better on Skeletal trials compared to Single Arm and Stub trials (two-tailed binomial test, $p = .002$), and 9/10 ($p = .021$) and 8/10 ($p = .109$) subjects performed better on Skeletal trials than on Arms and Arms + Stub trials, respectively.

The performance boost for Skeletal trials was not due to strategic differences such as giving a rushed response in the other Different trials (i.e., a speed–accuracy trade-off); in fact, subjects also responded fastest on Skeletal trials. Excluding response times (RTs) greater than two standard deviations above the mean (1.1% of all trials), RTs on correct trials were significantly faster for Skeletal shapes (421 ms) compared to Single Arm shapes (509 ms), $t(9) = 3.56$, $p = .006$; Stub shapes (506ms), $t(9) = 3.01$, $p = .015$; and Arms shapes (530 ms), $t(9) = 3.88$, $p = .004$; and were numerically faster (and certainly not slower) compared to Arms + Stub shapes (457ms), $t(9) = 1.40$, $p = .195$. These same trends were again observed nonparametrically, with 9 out of 10 subjects responding fastest on Skeletal trials compared to all other Different trials (ps = .021). Thus, beyond being detected most accurately, Skeletal changes may also be quicker and easier to detect.

We conducted three independent analyses to show that the performance boost for Skeletal changes could be attributed to skeletal structure, per se, over and above lower-level visual properties. First, because of how the shapes were constructed, the simplest possible image-based analysis (i.e., degree of pixel-by-pixel overlap) will never find that Skeletal shapes differed the most from the Parent shapes. In particular, One and Stub changes were roughly equivalent in pixelwise magnitude to Skeletal changes—and Arms and Arms + Stubs changes were even more extreme than Skeletal changes. This is because only one segment changes position during Skeletal changes, but three and four segments (and thus, three and four times the number of pixels) change their locations during Arms and Arms + Stub changes, respectively. Thus, on the basis of this intuitive and frequently used heuristic (e.g., Ullman, 1989), Skeletal shapes changed less than many other types of shapes in terms of such lower-level visual properties.

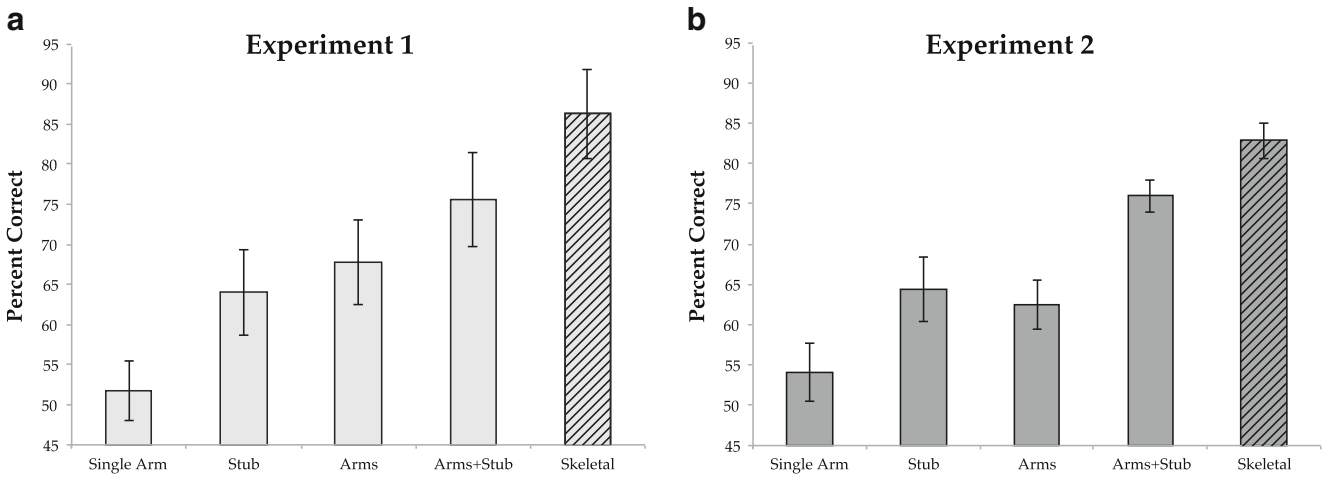

Fig. 6 Performance on the same/different tasks in (**a**) Experiment 1 and (**b**) Experiment 2. In both experiments, performance on Skeletal changes differed from performance on every other kind of change ($ps < .016$). Error bars depict ±1 *SEM*

Second, we considered an extensive list of specific lower-level properties that may have changed across shape modifications, to rule out the possibility that such properties might independently explain these results. We considered (a) the smallest angle between any two intersecting branches, (b) the largest angle between any two intersecting branches, (c) the area bounded by the shape's convex hull, (d) the shortest distance between any two branches' terminal points, (e) the average distance between all branches' terminal points, and (f) the standard deviation of the distances between all branches' terminal points. (These attributes exhaust all of the low-level features we considered to be possibly relevant in advance, along with every low-level feature suggested to us in debriefing by subjects who were asked to reflect on their strategies for discriminating the shapes.) For each of these attributes, we calculated the absolute value of the difference between a given Different shape and its Parent shape. These values are presented in Table 1. For five of the six attributes, the stimuli were not confounded to begin with: Skeletal shapes had numerically *smaller* changes on average than one or more other shape types. In fact, the Arms + Stub shape was, on average, *more* different from the Parent shape on each of these five unconfounded dimensions than was the Skeletal shape—and yet was *harder* to distinguish from the Parent shape than was the Skeletal shape.

The only attribute that was in fact confounded was the unsigned difference in the largest angle between any two intersecting branches—which was largest for Skeletal changes, on average. However, this was not the case for every shape family. So, to rule out this confound, we simply rank-ordered the shape families by this difference, and then progressively eliminated families until the confound disappeared. For the remaining 203 shape families (that were not confounded in this way—i.e., the largest ordered subset for which the difference in largest angle between any two intersecting branches was *smaller* on average for Skeletal changes than for Arms + Stub changes), Skeletal changes were still detected better than were all other types of changes ($ps < .015$). Thus, this potential confound cannot explain our results.

Finally, in addition to these shape features, we also computed an independent measure of shape similarity using the Malsburg Gabor-jet model (Lades et al., 1993; Margalit, Biederman, Herald, Yue, & von der Malsburg, 2016), which has been shown to robustly track human discrimination performance for metric differences between shapes (Yue, Biederman, Mangini, von der Malsburg, & Amir, 2012). Inspired by the Gabor-like filtering of simple cells in V1 (Jones & Palmer, 1987), this model overlays sets (or "jets") of 40 Gabor filters (5 scales × 8 orientations) on each pixel of a 128 × 128-pixel image and calculates the convolution of the input image with each filter, storing both the magnitude and the phase of the filtered image. This yields two separate feature vectors, one for magnitude and one for phase, each with 655,360 (128 × 128 × 40) values. The difference between image pairs was computed in two ways (for magnitude and phase, individually): Euclidean distance (which has been shown to correlate most highly with human discrimination performance; Yue et al., 2012) and cosine distance (which is invariant to the scaling of a vector). The results are shown in Table 2. As expected from the pixelwise differences alone, distances in the 655,360-dimensional space relative to the Parent shape are comparable for Skeletal, One, and Stub shapes, and are far greater for Arms and Arms + Stub shapes. Thus, even according to this fully pixelwise analysis (that makes no explicit reference to any specific shape properties), Skeletal shapes are objectively more similar than Arms and Arms + Stub shapes to their Parent shapes.

These results collectively suggest that visual dissimilarity is most accurately perceived for changes that influence a shape's underlying skeletal structure—even when those

Table 1 Lower-level features that could have been confounded with changes in skeletal structure

Modification Type	Comparison Feature					
	(a) Smallest angle	(b) Largest angle	(c) Convex hull area	(d) Shortest distance	(e) Average distance	(f) *SD* of distance
Experiment 1						
Single Arm	0°	1.615°	2.700 cm^2	0.050 cm	0.087 cm	0.090 cm
Stub	13.565°	2.33°	1.754 cm^2	0.127 cm	0.056 cm	0.065 cm
Arms	0°	4.823°	3.723 cm^2	0.094 cm	0.158 cm	0.141 cm
Arms + Stub	13.565°	6.795°	3.650 cm^2	0.162 cm	0.163 cm	0.151 cm
Skeletal	11.078°	16.58°	2.209 cm^2	0.096 cm	0.142 cm	0.084 cm
Experiment 2						
Single Arm	0°	1.05°	10852u^2	0.117u	4.751u	4.588u
Stub	18.65°	5.125°	10356u^2	1.401u	5.833u	6.473u
Arms	0.006°	3.166°	14618u^2	0.201u	7.607u	6.919u
Arms + Stub	18.67°	7.666°	16303u^2	1.530u	10.06u	9.220u
Skeletal	17.39°	16.04°	12109u^2	0.920u	10.38u	5.089u

Note. Each entry is the absolute value of the difference between a Parent shape and the relevant Different shape for that feature, averaged across the entire stimulus set. The features that we computed exhausted all of those we identified a priori as being possibly relevant, combined with all those mentioned by the subjects during debriefing: (**a**) the smallest angle between any two intersecting branches, (**b**) the largest angle between any two intersecting branches, (**c**) the area bounded by the shape's convex hull, (**d**) the shortest distance between any two branches' terminal points, (**e**) the average distance between all branches' terminal points, and (**f**) the standard deviation of the distances between all branches' terminal points. (u = Blender units. See text for details)

changes are objectively *smaller* in terms of their lower-level properties, as measured both by geometric features of the images (e.g., distances and angles between various segments) and by features prioritized during lower-level stages in visual processing (i.e., the outputs of the Gabor-jet model).

Table 2 Average psychophysical distances according to the Gabor-jet model (Lades et al., 1993; Margalit et al., 2016)

Modification Type	Gabor-jet Distance			
	Euclidean		Cosine	
	Magnitude	Phase	Magnitude	Phase
Experiment 1				
Single Arm	7137au	1219rad	0.1285	0.0866
Stub	6492au	1070rad	0.1052	0.0669
Arms	12420au	1873rad	0.3846	0.2033
Arms + Stub	13935au	2004rad	0.4843	0.2326
Skeletal	6287au	1142rad	0.0987	0.0760
Experiment 2				
Single Arm	3034au	2074rad	0.3788	0.2494
Stub	3095au	2074rad	0.3992	0.2494
Arms	3022au	2076rad	0.3777	0.2498
Arms + Stub	3067au	2073rad	0.3897	0.2491
Skeletal	3022au	2075rad	0.3788	0.2495

Note. Cosine distance is measured as one minus the cosine of the included angle between vectors, such that greater values indicate greater differences. As expected, the shapes that are more different on a pixel-by-pixel basis (i.e. Arms and Arms + Stub) are also more different according to this model. (au = arbitrary units, rad = radians. See text for details)

Experiment 2: Skeletal similarity in 3-D objects

The promise of shape skeletons is that they may serve as an effective format for real-world object representation, and accordingly a great deal of work in computer vision has been devoted to different ways of actually deriving skeletons from real-world 3-D images (e.g., Borgefors, Nyström, & Di Baja, 1999; Sundar et al., 2003). This challenge did not even exist for the stimuli in Experiment 1, however: By design, the shapes were line drawings, where each point on the shape was a point on the skeleton, and vice versa. To see if the performance boost for skeletal changes observed in Experiment 1 depended on these constraints, we replicated the experiment with volumetric 3-D objects of the sort depicted in Fig. 3b and in Fig. 5. These objects still closely approximated their underlying skeletons, but (a) they depicted 3-D structure from shading rather than being 2-D line drawings, and (b) it was no longer the case that every point on the shape was a point on the skeleton, and vice versa. This makes such stimuli somewhat unique in the psychological literature on skeletal shape representations, which have so often used (only) 2-D images (e.g., Denisova, Feldman, Su, & Singh, 2016; Firestone & Scholl, 2014; Kovacs & Julesz, 1994; Wilder et al., 2011; cf. Hung, Carlson, & Connor, 2012; Lescroart & Biederman, 2012).

In addition, this experiment eliminated another possible confound from Experiment 1. Because Skeletal shapes moved a branch from the central node to a more peripheral joint, such shapes had only three segments intersecting at the central node, whereas every other type of shape continued to have four such segments. Thus, an especially savvy subject could have succeeded at the task without actually engaging in comparison per se, simply by responding "different" any time a shape appeared that had three central segments. Even though this strategy exploits the shape's skeletal organization, we wanted to ensure that the task required active comparison. For this reason, the stimuli in this experiment included two types of shape families: one that was structurally identical to the shape families from Experiment 1 (with four branches intersecting at the central node) and a new type in which the Parent shape started with three central segments and was then transformed by the Skeletal manipulation into a shape with four central segments. Thus, the design of this experiment was identical except for the fact that the number of central segments in an object was never a reliable cue to the correct response—and so subjects had no choice but to actively compare the shapes.

Method

This experiment was identical to Experiment 1, except as noted. Ten new naïve subjects participated (with this sample size chosen ahead of time to match that of Experiment 1).

The results of Experiment 1 were robust even in just the first four blocks, so we truncated the session to this length. The stimulus set therefore consisted of 320 shape families, generated using the 3-D rendering program Blender (see Fig. 5). The stimuli were rendered with realistic shading from a single point-light source. Because the 3-D stimuli were presented as a 2-D projection (and so subject to foreshortening), we give their dimensions here in arbitrary units and note that for branches that were orthogonal to the camera and centered at the central node (and so not subject to foreshortening), 100 units corresponded to 1.62 cm on the testing monitor. Each segment of the Parent shape was 100 units wide and capped by a hemisphere with a diameter of 100 units so that the segment terminated smoothly. The length of each arm segment was a randomly chosen value between 100 and 200 units, while the stub was always 150 units. Every modified shape in every family had the same aggregate arm length of 900 units, and no two branches intersected, except at the node(s). The mask stimuli were similarly redesigned to more closely approximate the new 3-D stimuli being used (as depicted in Fig. 3b). Just as in Experiment 1, it is worth emphasizing that even though Skeletal changes reorganized the underlying shape skeletons, fewer pixels were actually altered during such changes compared to the other kinds of shape modifications.

The stimuli were rendered in advance using a camera angle that was randomly chosen for each shape family (with at least a 90° difference between the Parent shape of a given family and all other shapes in that family) but then fixed across subjects (as in Fig. 5). The Blender camera itself was positioned 600 units above and 780 units in front of the central node of the shape, and was aimed directly at the central node. The point light source sat immediately above the central node at a height of 1,000 units.

Half of the trials of every given type (and in every 80-trial block) involved a Parent shape with four central segments, and a Skeletal manipulation that resulted in three central segments (as in Experiment 1). The other half involved a Parent shape with three central segments, and a Skeletal manipulation that resulted in four central segments—with all manipulations designed in a manner corresponding to those in Experiment 1.

Results and discussion

Just as in Experiment 1, changes to the shape's underlying skeleton were easier to detect than any other kind of change. A repeated-measures ANOVA revealed a main effect of shape modification type, $F(4, 36) = 18.107$, $p < .001$, $\eta = .668$, and planned comparisons confirmed that Skeletal differences (82.8%) were detected better than any of the surface-feature changes that did not change skeletal structure: Single Arm, 54.1%, $t(9) = 6.74$, $p < .001$; Stub, 64.4%, $t(9) = 3.51$, $p = .007$; Arms, 62.5%, $t(9) = 5.17$, $p < .001$; and Arms + Stub, 75.9%, $t(9) = 2.96$, $p = .016$. Nonparametric data showed similar trends, as every single subject performed better on Skeletal trials compared to Single Arm and Arms trials (two-tailed binomial test, $p = .002$), and 9/10 ($p = .021$) and 8/10 ($p = .109$) subjects performed better on Skeletal trials than on Stub and Arms + Stub trials, respectively.

Also in agreement with Experiment 1, these results cannot be explained by a speed–accuracy trade-off. After excluding RTs that were greater than two standard deviations above the mean, subjects were numerically faster (and thus certainly not slower) to respond correctly on Skeletal trials (823 ms) than on all other types of trials (Single Arm: 922 ms; Stub: 887 ms; Arms: 906 ms; Arms + Stub: 869 ms).

We also tested and ruled out the same set of possible confounds as in Experiment 1—keeping in mind that the stimuli were again constructed such that Arms and Arms + Stub changes differed more from their Parent shapes in terms of pixelwise overlap than did all other sorts of changes, including Skeletal changes. As detailed in Table 1, most of the six features we tested were again not confounded in the first place. However, two features did differ most for Skeletal changes: (b) the largest angle between any two intersecting branches,

and (e) the average distance between all branches' terminal points. An analysis identical to that described in Experiment 1 ruled out both of these confounds: Considering the largest ordered subset of families in which these confounds simply were not present on average (202 of 320 families for largest angle; 313 of 320 families for average distance), Skeletal changes were still easier to notice than all other changes ($ps < .023$). Therefore, the performance boost observed with Skeletal changes is influenced by shape skeletons per se, rather than any of the lower-level features that may be correlated with changes to a skeleton.

Finally, we performed a similar analysis with the Gabor-jet model (as detailed in Experiment 1), and found that Skeletal changes were no greater on this pixelwise metric than were the other shape changes (see Table 2). (And even this is surely a conservative estimate of pixelwise shape differences because the 3-D images could not be brought into maximal alignment within the image plane—so they varied not only metrically but also in terms of their viewpoints. If they had been aligned, they would have produced a pattern much more similar to that of Experiment 1.)

General discussion

This project was motivated by a simple but profound question about visual experience: How do we perceive that two objects are similar or different? And this invites another foundational question from the perspective of cognitive science: What is the underlying representational format that makes such comparisons between objects possible? Whereas decades of research have proposed shape skeletons as a useful answer to this question in the context of computer vision, the current results provide the first direct evidence that *human* perception of similarity is likewise influenced by shape skeletons. Thus, beyond *existing* in human vision in the first place (e.g., Firestone & Scholl, 2014; Kovacs & Julesz, 1994) and perhaps guiding subjective judgments (e.g., van Tonder et al., 2002; Wilder et al., 2011), shape skeletons actually matter for objective visual *performance*.

Across two experiments, we demonstrated a robust advantage in the ability to discriminate shapes (both 2-D line drawings and 3-D volumes) as different when they had different skeletal structures—even when the structurally similar shapes differed to a greater degree in many types of lower-level attributes. Moreover, this performance boost occurred while the objects were simultaneously visible, implying that shape skeletons influence *perceived* similarity per se, rather than only influencing how shapes are remembered after the fact.

Future work could explore the power and generalizability of this result in at least two ways. First, given that the stimuli in the present studies were all (2-D or 3-D) stick figures, it will be important to determine the degree to which skeletal structure also influences objective similarity perception in stimuli whose shape skeletons are less similar to their visible contours. The fact that medial axis representations have been found to underlie the perception of many other types of shapes—for example, polygons (Firestone & Scholl, 2014), ellipses (Kovacs & Julesz, 1994), cardioids (Kovacs et al., 1998), and even silhouettes of plants and animals (Wilder et al., 2011)—provides some reason for suspecting that objective similarity judgments might similarly be influenced by skeletal structure in such cases, but this needs to be empirically tested. Second, the present work demonstrates an influence of shape skeletons on the perception of similarity, but of course we do not suggest that this is the *only* such factor (or even the principal one). As such, it might prove interesting in future work to directly compare the influence of skeletal structure with many other sorts of factors—for example closure (e.g., Elder & Zucker, 1993; Kovacs & Julesz, 1993), connectedness (e.g., Palmer & Rock, 1994; Saiki & Hummel, 1998), and topology (e.g., Chen, 1982, 2005). (And of course, unconfounding these factors would also require a wider array of shape types.) Comparisons to these others factors might help reveal not only *whether* skeletal structure influences the perception of similarity (as the current study demonstrates) but also how central it is within a broader hierarchy of visual features.

Parts, skeletons, and similarity

One reason why shape skeletons have captivated some human vision researchers is that they seem so counterintuitive. (In fact, this counterintuitiveness can be confirmed and measured directly; see Firestone & Scholl, 2014, Experiment 8.) As such, it may seem surprising that the familiar experience of visual similarity should be influenced by such an abstract geometric construct. But shape skeletons in fact respect many subjective aspects of perceived similarity—most notably, the sense in which two objects that share an underlying structure can and do look similar even when their superficial shapes are very different (as in Fig. 1). Frequently, objects in the world—both biological and man-made—*do* have real underlying structures that permit certain kinds of changes (such as articulations of limbs) but forbid others (such as translocation and/or reattachment of parts), and so perhaps it should not be such a surprise after all that such structure plays a role in human vision.

Indeed, this same insight has motivated other investigations into how the visual system represents *parts* of objects, even without explicitly invoking shape skeletons. Changes to the number of parts a shape has are readily detected (e.g., Barenholtz et al., 2003; Bertamini & Farrant, 2005), and shapes whose parts are articulated in ways that obey these part boundaries are explicitly judged to be more similar looking

(Barenholtz & Tarr, 2008). However, shape skeletons have been proposed as a better way to recover part structure, both because they can be used to represent a shape's structure hierarchically (Feldman & Singh, 2006) and because the transformations that are possible for an object also tend to be those that preserve patterns of skeletal connectivity. Indeed, representations of skeletal structure have recently been invoked to explain our sensitivity to certain part changes, such as articulations (Denisova et al., 2016); however, this study tested only changes to a particular part (such as bending, extending, or sliding a given branch), and not to a shape's overall skeletal organization. By contrast, the shapes in the present studies changed in their overall connectivity while *not* changing the brute physical appearance of any of the individual parts—and while carefully controlling for confounds (such as the minimum and average distances between parts' endpoints) that may otherwise be present in such changes (cf. Keane, Hayward, & Burke, 2003). (In fact, Denisova et al., 2016, found the *lowest* sensitivity for "sliding"' a part along the branch to which it is connected. This result amplifies the strength of the present findings, which suggest that part translations go from being the *least* detectable kind of change to the *most* detectable kind of change the moment such a translation alters the shape's skeletal organization.)

Overall, the present studies are the first to implicate skeletal organization per se in perceived similarity, beyond lower-level surface features and beyond part-structure. Shape skeletons thus not only influence subjective impressions of our environment but also alter our objective ability to compare and recognize objects in the world.

Author note For helpful conversation and/or comments on previous drafts, we thank Vladislav Ayzenberg, Jeremy Wolfe, and the members of the Yale Perception & Cognition Laboratory. This project was funded by ONR MURI #N00014-16-1-2007 awarded to B.J.S.

References

Aichholzer, O., Aurenhammer, F., Alberts, D., & Gärtner, B. (1995). A novel type of skeleton for polygons. *Journal of Universal Computer Science, 1,* 752–761.

Bai, X., & Latecki, L. J. (2008). Path similarity skeleton graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*, 1282–1292.

Barenholtz, E., Cohen, E. H., Feldman, J., & Singh, M. (2003). Detection of change in shape: An advantage for concavities. *Cognition, 89*, 1–9.

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

Bertamini, M., & Farrant, T. (2005). Detection of change in shape and its relation to part structure. *Acta Psychologica, 120,* 35–54.

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

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

Bolles, R. C., & Cain, R. A. (1982). Recognizing and locating partially visible objects: The local-feature-focus method. *The International Journal of Robotics Research, 1,* 57–82.

Borgefors, G., Nyström, I., & Di Baja, G. S. (1999). Computing skeletons in three dimensions. *Pattern Recognition*, *32*, 1225–1236.

Chen, L. (1982). Topological structure in visual perception. *Science*, *218*, 699–700.

Chen, L. (2005). The topological approach to perceptual organization. *Visual Cognition*, *12*, 553–637.

Corballis, M. C. (1988). Recognition of disoriented shapes. *Psychological Review*, *95*, 115–123.

Cortese, J. M., & Dyre, B. P. (1996). Perceptual similarity of shapes generated from Fourier descriptors. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 133–143.

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

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

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

Ferrari, V., Jurie, F., & Schmid, C. (2010). From images to shape models for object detection. *International Journal of Computer Vision*, *87*, 284–303.

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,* 377–386.

Harrison, S. J., & Feldman, J. (2009). The influence of shape and skeletal axis structure on texture perception. *Journal of Vision*, *9*, 1–21.

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

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*, 480–517.

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

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*, 1233–1258.

Keane, S., Hayward, W. G., & Burke, D. (2003). Detection of three types of changes to novel objects. *Visual Cognition*, *10*, 101–127.

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

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

Kovacs, I., & Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation. *Proceedings of the National Academy of Sciences of the United States of America*, *90*, 7495–7497.

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

Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R. P., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42, 300–311.

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

Liu, T. L., & Geiger, D. (1999). Approximate tree matching and shape similarity. In *Proceedings of the Seventh IEEE International Conference on Computer Vision* (Vol. 1, pp. 456–462). Los Alamitos, CA: IEEE.

Margalit, E., Biederman, I., Herald, S. B., Yue, X., & von der Malsburg, C. (2016). An applet for the Gabor scaling of the differences

between complex stimuli. *Attention, Perception, & Psychophysics*, *78*, 2298–2306.

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

Palmer, S. E., & Guidi, S. (2011). Mapping the perceptual structure of rectangles through goodness-of-fit ratings. *Perception*, *40*, 1428–1446.

Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review*, *1*, 29–55.

Peirce, J. W. (2007). PsychoPy: Psychophysics software in Python. *Journal of Neuroscience Methods, 162,* 8–13.

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

Saiki, J., & Hummel, J. E. (1998). Connectedness and the integration of parts with relations in shape perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 227–251.

Sebastian, T. B., & Kimia, B. B. (2005). Curves vs. skeletons in object recognition. *Signal Processing*, *85*, 247–263.

Serra, J. (1986). Introduction to mathematical morphology. *Computer Vision, Graphics, and Image Processing, 35,* 283–305.

Siddiqi, K., & Pizer, S. (2008). *Medial representations: Mathematics, algorithms, and applications*. New York, NY: Springer.

Siddiqi, K., Shokoufandeh, A., Dickinson, S., & Zucker, S. (1999). Shock graphs and shape matching. *International Journal of Computer Vision*, *30*, 13–32.

Sundar, H., Silver, D., Gagvani, N., & Dickinson, S. (2003). Skeleton based shape matching and retrieval. In *Shape Modeling International, 2003* (pp. 130–139). Seoul, South Korea: IEEE.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, *21*, 233–282.

Torsello, A., & Hancock, E. R. (2004). A skeletal measure of 2D shape similarity. *Computer Vision and Image Understanding*, *95*, 1–29.

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*, 215–240.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition, 32,* 193–254.

van Tonder, G. J., Lyons, M. J., & Ejima, Y. (2002). Visual structure of a Japanese Zen garden. *Nature*, *419*, 359–360.

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

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.

Zhang, D., & Lu, G. (2002). Shape-based image retrieval using generic Fourier descriptor. *Signal Processing: Image Communication*, *17*, 825–848.

Zhu, S. C., & Yuille, A. L. (1996). FORMS: A flexible object recognition and modelling system. *International Journal of Computer Vision*, *20*, 187–212.