Orthogonal multi-view three-dimensional object representations in memory revealed by serial reproduction

Thomas A. Langlois (thomas.langlois@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94720-1650 USA

Nori Jacoby (nori.viola@gmail.com)
Max Planck Institute for Empirical Aesthetics
Grüneburgweg 14, 60322 Frankfurt am Main, Germany

Jordan Suchow (jws@stevens.edu)
Stevens Institute of Technology
Hoboken, NJ 07030, USA

Thomas L. Griffiths (tomg@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08540, USA

Abstract

The internal representations of three dimensional objects within visual memory are only partially understood. Previous research suggests that 3D object perception is viewpoint dependent, and that the visual system stores viewpoint perspectives in a biased manner. The aim of this project was to obtain detailed estimates of the distributions of 3D object views in shared human memory. We devised a novel experimental paradigm based on transmission chains to investigate memory biases for the 3D orientation of objects. We found that memory tends to be biased towards orthogonal diagrammatic perspectives aligned with the ends of the standard basis for a set of common 3D objects, and that these biases are strongest for side views as well as top or bottom views for a small set of bilaterally symmetric objects. Finally, we found that views sampled from the modes were easier to categorize in a recognition task.

Keywords: Memory; 3D object perception; Serial reproduction; Iterated learning; Vision.

Introduction

Humans do not possess photographic memories of the things they see. Instead, visual memory is known to be biased towards systematic and simplified representations. The perception of 3D objects is known to be viewpoint dependent, but detailed estimates of the distributions of 3D object views in shared human memory remain unknown. For a given object, towards what views does visual memory tend to be biased? Are the number of views the same across different objects? How many views are there? Evidence from prior work points to systematic viewpoint-specific biases in 3D object perception such as so-called "canonical" views of common everyday objects (Palmer & Rosch, 1981). Canonical views are associated with improvements in categorization accuracy and recognition (as measured using response-time latencies). While the human visual system is largely robust to perspective transformations, this work provided early evidence for viewpoint dependence in human object perception, a finding that was corroborated in subsequent work (Bülthoff et al., 1995). However, none of this work fully characterized the object-specific distributions of views that bias visual memory, and provided mostly indirect evidence for them. We therefore attempt to provide a detailed picture of the structure of memory biases for the orientation of 3D objects.

We aimed to uncover the distributions of 3D object views in shared human memory. Doing so is of particular interest to disambiguate theoretical explanations for viewpoint dependence in 3D object perception, and to determine if biases in remembered views of objects correspond to canonical views. Two theoretical explanations have been suggested in order to explain canonical views: the "frequency hypothesis" and the "maximal information hypothesis" (Mezuman & Weiss, 2012). The "frequency hypothesis" states that privileged views correspond to the views that are most commonly taken when viewing or interacting with everyday objects, while the "maximal information hypothesis" states that these views change the least under small local perspective transformations. The "frequency hypothesis" is most consistent with the notion of a statistical "prior" in Bayesian accounts of perception and memory. However, it remains an open question as to whether memory representations for 3D objects resemble canonical views, and if these representations are shaped by statistical priors.

To answer this question, we used transmission chains adapted to a 3D orientation memory experiment. Under experimentally verifiable conditions, transmission chains are known to approximate samples from shared priors (Xu & Griffiths, 2010), and can be used to characterize shared biases in reconstructive memory. In this paper, we start by outlining past computational approaches and empirical findings regarding 3D object representations, as well as theoretical properties of transmission chains. Next, we present our novel findings revealing hitherto unknown distributions of 3D memory biases for a range of everyday objects. We find that these distributions are characterized by systematic patterns of biases towards diagrammatic orthogonal views that appear to be aligned with the faces of the objects (strong side views,

front and back views, top and bottom views). These views do not appear to match known canonical views, which are typically semi-profile views, although they are consistent with past findings that revealed similar biases in visual inspection of novel objects in adults (Perrett et al., 1992), as well as infants (Pereira et al., 2010). We also find that these views were associated with improved categorization accuracy relative to views sampled from areas far from the modes in these distributions.

Background

Transmission chains and experimental methods Transmission chains are analogous to the so-called "telephone game." In the most famous and early example, Bartlett had a series of people reproduce a drawing of an owl hieroglyph, and as the the reproductions of the image progressed through the chain, what began as an imperfect but recognizable facsimile of the hieroglyph morphed into an image of a cat (Bartlett, 1932), revealing that the participants shared a common bias to distort the unusual image into an image for which they had a strong collective prior.

Transmission chains have since been adopted to study phenomena in many fields, including evolutionary biology, cognitive science, anthropology, vision science, and music cognition (Kirby et al., 2008; Jacoby & McDermott, 2017; Lew & Vul, 2015). A recent analysis of reconstruction from memory examined how information should change as it is transmitted through a chain of rational agents (Xu & Griffiths, 2010). Under the rational analysis, reconstruction from memory is defined as the problem of inferring the most accurate state of the world despite a noisy or imperfect sensory input (such as an imperfect memory trace of a scene or an object in the world). Using the framework of Bayesian statistics, this problem can be captured as follows: Previous experience is characterized by a prior distribution over possible world states (a hypothesis space of all conceivable world states, such as all possible 3D orientations of an object). The posterior is computed by integrating that prior with the likelihood, which in this case simply describes the probability of observing a world state (such as an object in a particular orientation), given a hypothesis about the true state of the world. In this work, (Xu & Griffiths, 2010) found that a transmission chain populated by rational Bayesian agents defines a Markov chain with the following transition probabilities:

$$p(x_{n+1} \mid x_n) = \int p(x_{n+1} \mid \mu) p(\mu \mid x_n) d\mu,$$

where x is a noisy stimulus (such as noisy recollection of the orientation of a previously viewed object) and μ is the true state of the world that generated that stimulus. This Markov chain captures the probability of a new stimulus x_{n+1} being created as a reconstruction of a previously seen stimulus x_n in each iteration in the transmission chain, and has a stationary distribution which defines the probability of observing a stimulus x when μ is sampled from the prior:

$$p(x) = \int p(x \mid \mu) p(\mu) d\mu.$$

This process approximates a Gibbs sampler for the joint distribution on x and μ defined by multiplying $p(x \mid \mu)$ and $p(\mu)$. In other words, assuming that participants share common inductive biases, the transmission chain will converge to a sample from their shared prior.

Computational theories of 3D representations To date, a significant body of work has explored the nature of human representations of 3D objects and a great deal of experimental work has been done to elucidate the characteristics of human perceptual representations of 3D objects and scenes. (Palmer & Rosch, 1981) provided early evidence for the existence of privileged "canonical" views that facilitate 3D object recognition, in keeping with principles of categorization (Rosch, 1999) that introduced the notion of "prototype exemplars." Later work introduced the recognition-by-components (RBC) theory of image understanding (Biederman, 1987). This work proposed that representations of objects in memory are accessed when components ("geons") derived from perceptual mechanisms (Lowe, 2012; Rock, 1983) are combined, and that these components form a perceptual basis for a "componential representation of real world objects in memory." A third computational theory argues that objects are represented as lists of viewpoint-invariant properties (A piano has keys, pedals, legs) (Bülthoff et al., 1995; B. Tversky & Hemenway, 1984; A. Tversky, 1977), or by points in abstract multidimensional feature spaces (Carr et al., 2001; **?**; Su et al., 2015).

Theories based on list-based feature descriptors or viewpoint-invariant parts have been difficult to reconcile with experimental data showing systematic view-specific variations in human response-time latencies and recognition accuracy (Bülthoff et al., 1995; Tarr et al., 1998). These results have tended to favor theories that postulate viewpoint-specific and largely 2D representations (Vetter et al., 1995; Bülthoff et al., 1995) as forming the basis for human object representations. However, to our knowledge, little work has been done to devise an experimental method for revealing the distributions of viewpoint-specific biases in memory representations.

Canonical perspectives were discovered for objects that were bilaterally symmetric due to experimental constraints, and although (Palmer & Rosch, 1981) confirmed the presence of privileged views for each, it is possible that other canonical views, such as the mirror images of bilaterally symmetric objects exist. In fact, work using online images returned by search engines estimated the modes of the distribution of 3D perspectives for a variety of objects, and found that canonical views for bilaterally symmetric objects are typically bi-modal (Mezuman & Weiss, 2012). In this paper we adapted transmission chains to a memory paradigm in which we probed collective biases in reconstructive memory for the 3D orientation of a handful of everyday objects in order to uncover any and all biases in 3D reconstructive memory.

Methods

Participants

All participants were recruited online using Amazon Mechanical Turk and gave informed consent. according to a protocol was approved by The Committee for the Protection of Human Subjects (CPHS) at the University of California, Berkeley. Each experiment required approximately 100 participants.

Stimuli

The stimuli used in these experiments were 3D objects that could be viewed from any angle by rotating a camera oriented towards the origin of the object, and at a fixed distance (travelling on the surface of a sphere around the object) and with the camera tilted (in a direction tangent to the sphere). We started with a detailed mesh model of a typical teapot, and shoe. In addition, we used grayscale versions of the teapot and shoe, as well as a grayscale 3D model of a car, alarm clock, armchair, coffee maker, camera, and grand piano, see Figure 1A. We selected objects matching the objects in (Palmer & Rosch, 1981) as closely as possible.

Procedure

For each object, we ran a serial reproduction experiment with 250 chains and 20 iterations (see Figure 1B). Participants viewed timed displays of the 3D object. The chains were initialized as camera views over the surface of a unit sphere with the object in the center. The camera frame orientation was always oriented towards the center of the object, but was tilted at random angles orthogonal to the sphere (the "up" vector, see Figure 1D). The position of the camera and the view were sampled uniformly from the Haar measure on $SO(3)$ (Perez-Sala et al., 2013). Following the timed display, and 1000 ms retention phase when the screen went blank, a probe screen containing the object at a new random orientation was shown.

Participants were instructed to orient the object (which is equivalent to rotating the camera view) so that it matched the original orientation of the object that was shown during the initial timed display. Participants were not given time constraints during the probe, and could change their responses as many times as they needed. The object on the screen could be rotated by means of the mouse, as well as a set of buttons (see Figure 1C). Participants were given 10 practice trials during which the initial display was shown for 4000 ms in order to familiarize them with the nature of the task, and the user interface. Only after they completed the practice trials was the presentation time reduced to 1000 ms. In addition, they were given trial-by-trial feedback based on their performance (either a green message saying "Well done! Your response was sufficiently accurate", or a red message stating: "Your response was insufficiently accurate"), see Figure 1C.

Results

By the final iteration of the transmission chain process, a clear pattern emerges: 3D views are biased towards a small set of orthogonal "diagramatic" views that are aligned with the top,

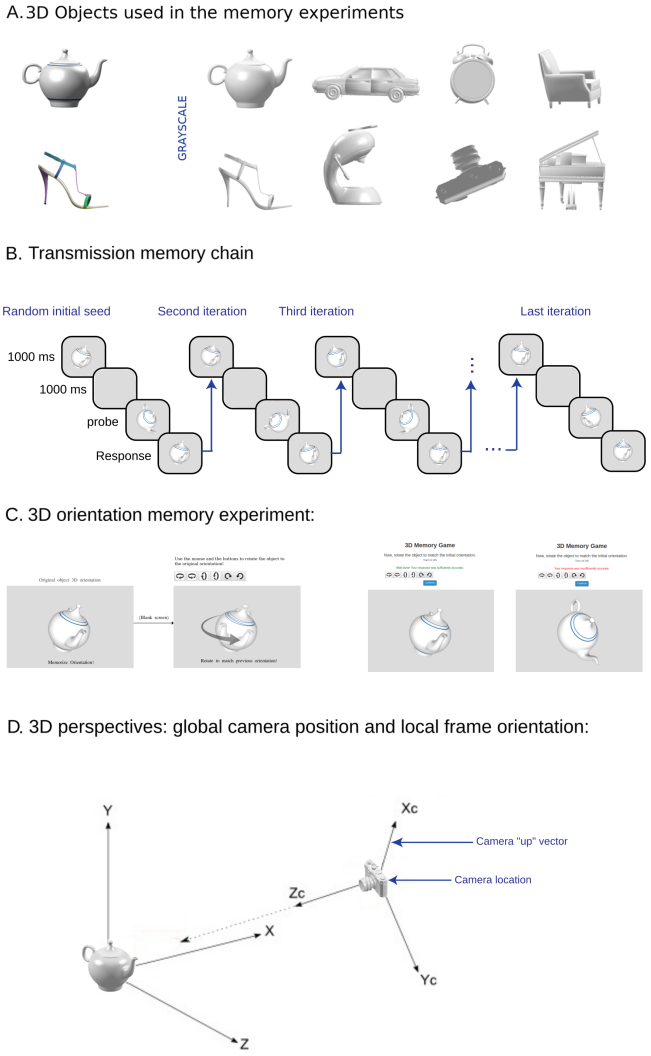

Figure 1: 3D objects, experiment structure, task and geometry. A. Textured and grayscale 3D objects used in the transmission chain experiments. B. Transmission chain structure: A 3D view of an object (teapot) is initialized somewhere at random over a unit sphere. This view is presented as a stimulus to a subject who then reconstructs this view from memory. The subject's response is then presented as the stimulus to a second subject, who must reproduce this second view, and so on. C. The experiment instructions and trial structure. Participants could rotate the object with the mouse and a set of buttons displayed over the image. They were instructed to reproduce the view they saw as accurately as possible, and were given feedback on their performance. D. Geometry of 3D object views adopted in the experiment. Views (cameras) were always positioned on the surface of a sphere centered at the object, and was always pointed towards the center of the sphere (towards the object). The local frame of the camera could vary according to the "up" vector, which controls the tilt of the camera

bottom, and side views of the objects. In some cases, the views in the modes correspond to the front and back (for the clock in particular), see Figure 2 for the results obtained with the textured teapot and shoe. While all the starting views of the chains are camera views sampled uniformally over the sphere surrounding the objects, the distributions quickly change and become clustered around four distinct modes as the chains progress. Figure 2A shows the initial distribution, the distribution at the 5th, 10th, 15th and 20th iteration of the transmission chains for the teapot and shoe.

Figure 2B shows the distributions of all points across all iterations for the shoe and teapot. In addition, the four modes with respect to the camera directions are plotted in four colors for both distributions. Next to each of these distributions, we show the corresponding histograms of the angles of the "up" vectors at the modes, where the direction most aligned with the data is centered to 90 degrees for the first two modes of the teapot and shoe. Surprisingly, while the "up" vectors in modes I, II (side views) of the Teapot are centered mainly in one direction, those in modes III and IV (the top and bottom views) show a bimodal distribution (top and bottom views are remembered with the handle and spout oriented vertically, while the side views show them to be oriented horizontally, orthogonal to the vertical orientations in the top and bottom views). We don't find this pattern in the case of the shoe, where the distributions of "up" vector angles were unimodal for all modes (I, II, III, and IV). This suggests that memory representations contain interaction patterns where some objects are memorized with a specific location *and* orientation, while memory for views of other objects are not necessarily associated with particular angular orientations. The columns on the far right of Figure 2B show spherical kernel density estimates (KDEs) of the final iteration data oriented according to the top four modes. Thumbnail insets to the right of the KDE modes show the corresponding object views. For both objects, the top two views are side views, while the remaining two modes correspond to the top and bottom views.

In order to verify if our chains showed convergence, we measured the mean copying error of the camera views for the textured teapot and shoe objects (See Figure 2C). The copying error was computed separately for each iteration by averaging the difference between the remembered camera view responses and stimulus views. We found that the copying error tends to reduce over the course of the experiment. Indeed, whereas the copying error for the first iterations was significantly smaller compared with the last iteration ($t(364)=6.6$, $p<0.001$ and $t(386)=5.6$, $p<0.001$ for the teapot and shoe, respectively), the difference between the copying error in the last iteration was not significantly different from the preceding four iterations ($p > 0.1$). For all cases this holds true even with Bonferroni corrections for multiple comparisons). This suggests that convergence occurs by the last five iterations of the chains.

In order to control for effects of colors and texture on 3D memory biases, as well as to control these factors for the

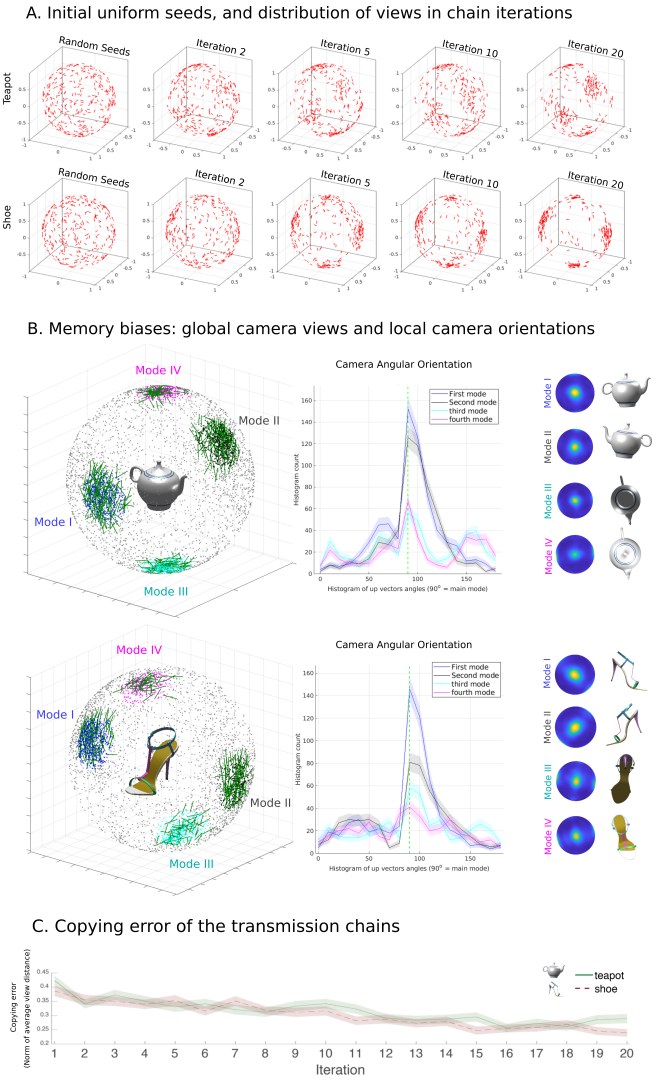

Figure 2: Transmission chain results for a teapot and shoe. A. Scatter plots showing camera views and "up" vectors for four chain iterations, and the initial uniform random seed locations. First row shows results for the teapot (initial seed, 5th, 10th, 15th, and 20th iteration distributions), and second row shows results for the shoe. B. Modes in the 20th and final distributions of views for the teapot and shoe. Four modes are clearly discernible: the side views of the objects, and the top and bottom views. Spherical subplots show a superposition of camera views across all iterations, highlighted are the four modes obtained by the 20th and final iteration of the chains. These correspond to the side views as well as the top and bottom views. The central subplots show histograms of the "up" vector angles, which show the frequency of local camera orientations at each of the modes. They reveal that perspectives in the first two modes (side views in both cases) are biased towards views where the camera is oriented towards a 90 degree angle, which yields views of the objects that are upright. These views are visualized in the far right columns, for each object, along with views of the modes in spherical Kernel Density Estimates (KDEs) of the 20th iteration data. C. Copying error across the chain iterations.

recognition experiments that follow, we ran a set of grayscale objects. Figure 3A the data for the novel objects. The results for the teapot and shoe were largely consistent with the results obtained with the textured versions of these objects, although only three clear modes were observed for the grayscale shoe (side views, and top view). Similarly, the modes for the car reveal four orthogonal views: left and right sides, as well as top, and bottom views. For the remaining objects, either three or two modes were present. Only two primary modes revealing frontal views, and back views were observed for the clock. Finally, the results for the remaining objects reveal primarily three orthogonal views. In sum, we find that 3D object memory representations are not equivalent to canonical views, and are characterized by multi-modal biases that may reflect the symmetry of the objects, a finding that corroborates findings revealing the presence of bi-modal views in distributions of online images (Mezuman & Weiss, 2012), although those were not views aligned with the faces of the objects, nor were they necessarily orthogonal. The memory representations we uncovered replicate past findings showing systematic biases towards the same views in a variety of visual inspection tasks in both infants and adults, suggesting that memory biases may be influenced by encoding precision and angular discrimination.

Figure 3 shows the results of a categorization experiment in which we compared the categorization accuracy for the set of eight grayscale objects when they were presented from views sampled in the modes of our memory KDEs, or from views far from the modes (sampling 4 nearest neighbors around the points that were farthest from the modes on the sphere, in the initial seed distributions of the chains). Figure 3B shows example views, and the experimental task: subjects were presented with a view for 100 ms, and then asked to categorize the object. The eight object labels were shown, as well as two additional labels ("house", "horse"). Figure 3B shows recognition d' results as a function of view type. We found that views of the shoe, clock, car, teapot, and coffee machine were recognized more accurately when they were sampled from the modes in our KDEs ($p < 0.001$ in all cases, following a Bonferroni correction for multiple comparisons). Overall, views sampled from the modes were associated with improved classification accuracy ($p < 0.0001$).

Discussion

We found the use of transmission chains to be particularly sensitive to characterizing shared 3D memory biases. The biases we observed for a small set of bilaterally symmetric everyday objects are highly systematic and not identical to known canonical views. They are strongly diagrammatic views of the sides, top and bottom, or front and back faces of the objects. In this respect, they resemble the bimodal characteristics of the distributions of online images estimated by (Mezuman & Weiss, 2012), although the diagrammatic aspects of these views are more reminiscent of well-known biases in visual inspection of 3D objects Perrett et al. (1992);

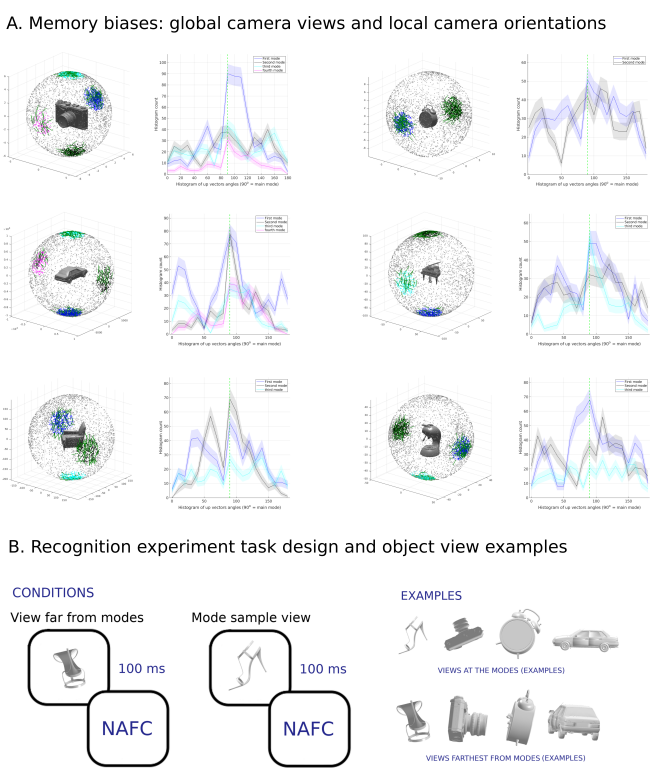

Figure 3: Recognition experiments. A. Memory biases for global camera views and local camera orientations for grayscale objects. B. Recognition experiment task design and view examples. For each object, participants were presented with a view sampled either from one of the perspective modes obtained from the final chain iteration, or from a point farthest from one of the final modes sampled from the uniform seed distribution, for 100 ms. They were then asked to select the correct object name from a list of possibilities. C. Recognition d' results for each of the objects, and for all the objects. Results show that in most cases, participants were more likely to select the correct object label when the view shown was sampled from one of the modal views sampled form the final chain iteration. Error bars correspond to 1000 bootstrapped samples of the data, with replacement. We used the Bonferroni correction for multiple comparisons.

Pereira et al. (2010). However, we did observe differences between objects, with some object representations containing four distinct modes, and others containing fewer (the clock). In addition to finding clear biases in camera locations, we also observed that camera views tended to be consistently oriented upright for the side views, but not for top or bottom views. Finally, we determined that categorization accuracy was higher for views that were sampled from the modes of the distributions we estimated, when compared to views sampled from regions that were farthest from the modes. This suggests that 2D memory representations of 3D objects are informative for recognition.

Finally, using this tool to uncover memory priors for objects that are not bilaterally symmetric, and with different geometries could help determine what factors are responsible for shaping biases in 3D memory representations. Our current findings do not appear to be altogether consistent with statistical priors (the "frequency hypothesis"), since diagrammatic views (especially of the bottom of objects like cars, teapots, and pianos) are not views of these objects that are typically experienced. However, they may be due to variable angular discrimination accuracy, which may be increased for sides that are aligned with the first principal component axes of the objects, and decreased for the shorter sides. Our approach provides a powerful tool for estimating detailed distributions of biases in 3D memory, and can provide an empirical basis for spurring novel theoretical insights on the nature of these representations.

Acknowledgments

This work was funded in part by National Science Foundation grant SPRF-IBSS-1408652 to T.L.G. and J.W.S. and DARPA Cooperative Agreement D17AC00004 to T.L.G and J.W.S. The contents of this paper does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

References

Bartlett, F. C. (1932). Remembering: An experimental and social study. *Cambridge: Cambridge, UK*.

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

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, *5*(3), 247–260.

Carr, J. C., Beatson, R. K., Cherrie, J. B., Mitchell, T. J., Fright, W. R., McCallum, B. C., & Evans, T. R. (2001). Reconstruction and representation of 3d objects with radial basis functions. In *Proceedings of the 28th annual conference on computer graphics and interactive techniques* (pp. 67–76).

Jacoby, N., & McDermott, J. H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, *27*(3), 359–370.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. In *Proceedings of the national academy of sciences* (Vol. 105, pp. 10681–10686).

Lew, T. F., & Vul, E. (2015). Structured priors in visual working memory revealed through iterated learning. In *Proceedings of the 37th annual meeting of the cognitive science society.*

Lowe, D. (2012). *Perceptual organization and visual recognition* (Vol. 5). Springer Science & Business Media.

Mezuman, E., & Weiss, Y. (2012). Learning about canonical views from internet image collections. In *Advances in neural information processing systems* (pp. 719–727).

Palmer, S., & Rosch, E. (1981). Chase. p.(1981). canonical perspective and the perception of objects. *Attention and performance IX*, 135–151.

Pereira, A. F., James, K. H., Jones, S. S., & Smith, L. B. (2010). Early biases and developmental changes in self-generated object views. *Journal of Vision*, *10*(11), 22–22.

Perez-Sala, X., Igual, L., Escalera, S., & Angulo, C. (2013). Uniform sampling of rotations for discrete and continuous learning of 2d shape models. In *Robotic vision: Technologies for machine learning and vision applications* (pp. 23–42). IGI Global.

Perrett, D. I., Harries, M. H., & Looker, S. (1992). Use of preferential inspection to define the viewing sphere and characteristic views of an arbitrary machined tool part. *Perception*, *21*(4), 497–515.

Rock, I. (1983). The logic of perception. *Vision Science*.

Rosch, E. (1999). Principles of categorization. *Concepts: core readings*, *189*.

Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E. (2015). Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the ieee international conference on computer vision* (pp. 945–953).

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience*, *1*(4), 275.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*(2), 169–193.

Vetter, T., Hurlbert, A., & Poggio, T. (1995). View-based models of 3d object recognition: invariance to imaging transformations. *Cerebral Cortex*, *5*(3), 261–269.

Xu, J., & Griffiths, T. L. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology*, *60*(2), 107–126.