

Semantic structure in communicative drawings

Anonymous Authors

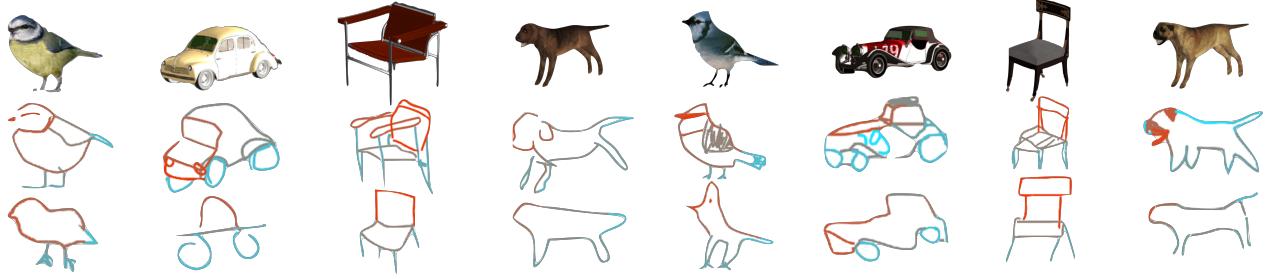


Figure 1: Drawings of 3D objects produced in communication task. Different stroke colors correspond to different object parts.

Abstract

The ability to represent semantically meaningful structure in our environment is a powerful aspect of human visual perception and cognition. As a testament to this ability, we effortlessly grasp the correspondence between a drawing of a particular object and that physical object in the world, even if the drawing is far from realistic. How are visual object concepts organized such that they can robustly encode such abstract correspondences? Here we explore the hypothesis that this is in part because we readily decompose both objects and drawings into a common set of semantically meaningful parts. Towards this end, we developed a web-based platform to densely annotate the semantic attributes of drawings of real-world objects produced in different contexts, allowing us to examine the semantic structure of more-detailed and sparser drawings of the same object. We found that: 1, people are highly consistent in how they interpret what individual strokes represent; 2, single strokes tend to correspond to single parts; 3, strokes representing the same part tend to be clustered in time; and 4, detailed and sparse drawings of the same object emphasized similar part information, although 5, detailed drawings of different objects tend to be more distinct from one another than simpler ones. Taken together, our results support the notion that people deploy their abstract understanding of the compositional part structure of objects in order to select actions to communicate relevant information about them in context. More broadly, they highlight the importance of structured knowledge for understanding how pictorial representations convey meaning.

Keywords: sketch understanding; perceptual organization; visual production; object representation; compositionality

Introduction

When we open our eyes, we do not experience a meaningless array of photons — instead, we parse the world into people, objects, and their relationships. The ability to represent semantically meaningful structure in our environment is a core aspect of human visual perception and cognition (Navon, 1977). As a testament to this ability, we effortlessly grasp the correspondence between a drawing of a particular object and that physical object in the world, even if the drawing is far from realistic (Eitz, Hays, & Alexa, 2012; J. E. Fan, Yamins, & Turk-Browne, 2018). How are visual object concepts organized such that they can robustly encode such abstract correspondences? Here we explore the hypothesis that this is in part because we readily decompose both objects and

drawings into a common set of semantically meaningful parts (Biederman & Ju, 1988).

Recent advances in computational neuroscience and artificial intelligence have provided an unprecedentedly clear view into the algorithms used by the brain to extract semantic information from raw visual inputs, exemplified by modern deep learning approaches (Yamins et al., 2014). Nevertheless, a major gap remains in elucidating how the feature representations learned by deep learning models can be adapted to emulate the structure and flexibility of human visual semantic knowledge (Lake, Ullman, Tenenbaum, & Gershman, 2017). A promising approach to closing this gap may be to combine the learning capacity of deep neural networks with the parsimony and interpretability of structured representations that reflect how visual concepts are organized in the human mind (Battaglia et al., 2018). Pursuing this strategy relies on a thorough understanding of this conceptual organization and how this organization enables behavioral flexibility.

The goal of this paper is to contribute to this understanding by probing the expression of visual semantic knowledge in a naturalistic setting that exposes both its structure and flexibility: visual communication via drawing.¹ This approach departs from the conventional strategy for inferring the organization of visual object concepts from behavior, which relies upon tasks that elicit judgments about visual inputs, usually with respect to experimenter-defined dimensions. By contrast, visual communication tasks permit participants to include any elements they consider relevant to their goals and combine these elements freely, yielding high-dimensional information about how visual semantic knowledge is organized and deployed under a naturalistic task objective.

Our aim in probing the semantic structure of communicative drawings is to shed light on how the semantic organization of visual object representations supports their flexible expression across contexts. Our

¹All materials and data are available at https://github.com/cogtoolslab/semantic_parts.

approach advances recent work (J. E. Fan et al., 2018; Long, Fan, & Frank, 2018) that has investigated the production of object drawings to communicate in two ways: first, an explicit focus on compositional semantic structure in drawings, and second, the examination of flexibility in how visual semantic knowledge is expressed in different semantic contexts.

Towards this end, we developed a web-based platform to densely annotate drawings of real-world objects produced in different semantic contexts, including detailed and simpler sketches of each object. Overall, we found that: (1) people are highly consistent in how they interpret what individual strokes represent; (2) single strokes tend to correspond to single parts; (3) strokes representing the same part tend to be clustered in time; and (4) detailed and sparse drawings of the same object emphasized similar part information, although (5) detailed drawings of different objects tend to be more distinct from one another than simpler ones. Taken together, our results support the notion that people deploy their abstract understanding of the compositional part structure of objects in order to select actions to communicate relevant information about them in context.

Methods

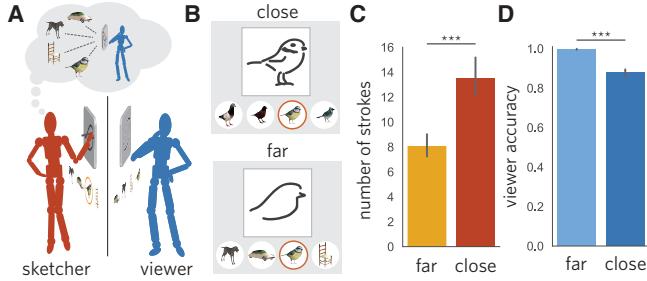


Figure 2: (A) Participants were paired in an online environment to play a drawing-based reference game in which the sketcher aimed to draw a target object such that a viewer could distinguish it from three distractor objects. (B) In close contexts, the target and distractors all belonged to the same basic-level category; in far contexts, the target and distractors belonged to different basic-level categories. (C) Sketchers used fewer strokes in the far condition, while producing sketches that were accurately recognized by the viewer in both conditions.

Drawing dataset

We obtained 1195 drawings of 32 real-world objects from a recent experimental dataset in which participants were paired in an online environment to play a drawing-based reference game (J. Fan, Hawkins, Wu, & Goodman, 2018). Objects belonged to one of four basic-level categories (i.e., bird, car, chair, dog), each of which contained eight exemplars (Fig. 3A). On each trial of this reference-game experiment, both participants were presented with a shared context containing an array of photorealistic 3D renderings of four objects. One participant (i.e., the sketcher) aimed to draw one of these objects – the target – so that the other participant (i.e.,

the viewer) could pick it out from a set of distractor objects (Fig. 4A). Across trials, the similarity of the distractors to the target was manipulated, yielding two types of semantic context: close contexts, where the target and distractors all belonged to the same basic-level category, and far contexts, where the target and distractors belonged to different basic-level categories (Fig. 4B). This context manipulation led sketchers to produce simpler drawings containing fewer strokes and less ink on far trials than on close trials, while still achieving high recognition accuracy in both types of context (Fig. 4C, Fig. 3B&C).

Prior work analyzing the semantic properties of such drawing data have represented them as raster images (e.g., *.png), an expedient format for applying modern convolutional neural network architectures (J. E. Fan et al., 2018; Sangkloy, Burnell, Ham, & Hays, 2016; Yu et al., 2017). However, a key limitation of treating a drawing like an image is that one loses information about the inherently sequential and contour-based nature of drawing production. Because our goal is to characterize the fine-grained semantic organization of drawings, it was thus crucial for our purposes to represent each drawing instead using a vector image format (i.e., *.svg).

Each drawing in our dataset is represented as a sequence of individual strokes, where each stroke consists of a sequence of sub-stroke elements, known as splines. These splines are parameterized as cubic Bezier curve segments, which are uniquely defined by four points: the initial point, the final point, and two control points that control the spline's curvature. This data format provides a relatively compact representation of each drawing compared with a rasterized image, while still providing sufficient expressivity to provide an accurate representation.

Semantic annotation

In the present study, we developed a novel web-based platform to crowdsource semantic annotations for every spline of every stroke of every sketch in our dataset.

Participants 326 participants were recruited via Amazon Mechanical Turk (AMT). For this experiment, participants provided informed consent in accordance with the Stanford University IRB. Participants were provided with a base compensation of \$0.35, plus \$0.002 for every sub-stroke element they annotated and \$0.02 for every sketch they annotated completely.

Annotation task [jefan: Would be good to put a cleaned up version of the annotation task interface around here.]

In each annotation session, participants were presented with 10 drawings that were randomly sampled from the reference-game dataset. Each trial, one of these sketches appeared in the center of the display, above the same array of four objects that the sketcher had viewed, with one of these objects highlighted as the target. Thus the annotator had full information about which object the sketcher had intended to depict, as well as the identity of the distractors. The goal of the annotator was to tag each spline with a label

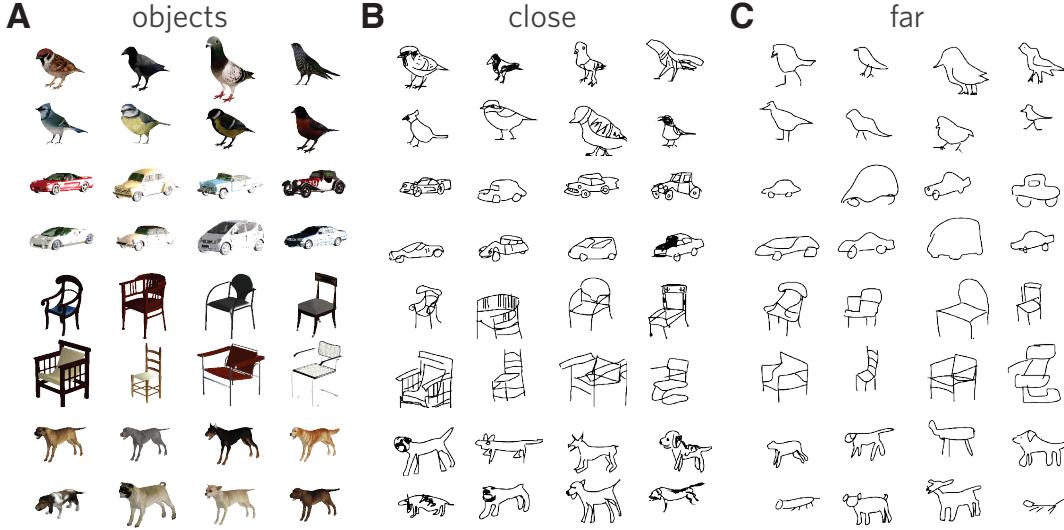


Figure 3: (A) Target objects. (B) Example object drawings produced in a close context. (C) Example drawings produced in a far context.

corresponding to the part it represented (e.g., seat, leg, back for a chair). To facilitate this, participants were provided with a menu of common part labels that were tailored to each basic-level category represented in our dataset. [jefan: It would be good to have a table containing the part labels that were provided for each category.]

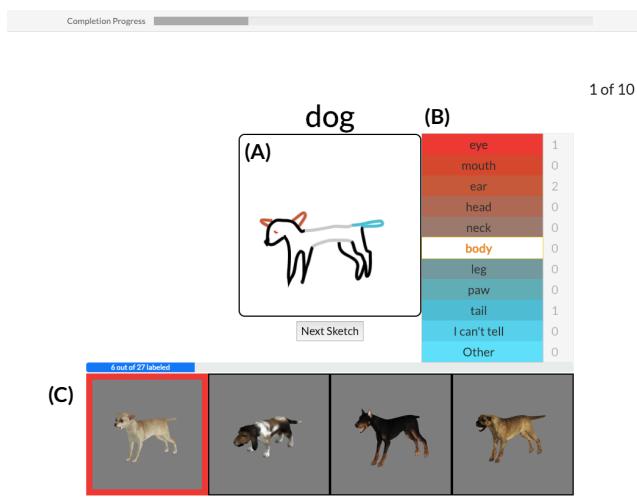


Figure 4: Sketch annotation Tool. (A) Participants were provided with the category label of the sketch whose parts they were to label. Labeled strokes changed color to match the label's color in the menu. (B) The accompanying part menu labels varied depending on the category of the sketch. (C) Images that the sketcher distinguished between during the reference game were provided for fuller context.

Category	Labels
Bird	eye, beak, head, body, wing, leg, feet, tail, other
Car	bumper, headlight, hood, windshield, window, body, door, trunk, wheel, other
Chair	backrest, armrest, seat, leg, other
Dog	eye, mouth, ear, head, neck, body, leg, paw, tail, other

Table 1: The four sketch categories in our dataset and the accompanying part labels we provided for sketches of each category.

However, participants were also free to generate their own part label if none of the common labels applied. In total, we collected 3608 annotation trials of 1195 unique sketches. [jefan: What happened to the other 3 drawings? See top of ‘Drawing dataset’ subsection.]

Inclusion criteria Because one of our central goals was to understand the relative emphasis that sketchers placed on different part information, we restricted our analyses to annotation trials in which the drawing was completely annotated (i.e., all splines were tagged). Moreover, in order to be able to examine inter-annotator consistency in how drawings were annotated, we only examined drawings that were annotated by at least three distinct participants. [jefan: By the way, are we including drawings that were annotated more than three times? We could, of course, and just ignore the 4th and onward... Also, have we ever checked whether the annotations were coming from distinct participants? Not

sure, this requires comparing `workerId`, which we do not save to the group dataframe.] Some of the custom part labels provided by participants were valid, but at a finer grain than or synonymous to other more frequently occurring labels. For example, in the case of chair sketches, strokes that represented legs were sometimes labeled as "leg support", "foot", and "strut". We mapped these labels to the more frequently occurring label "leg". In order to characterize the semantic structure of drawings at a consistent granularity, we also manually constructed a part dictionary to map these overly fine-grained part labels to one of the common part labels, where appropriate. After applying these inclusion criteria, our annotated dataset consisted of 764 drawings that had been annotated exactly 3 times each, using a set of 24 unique part labels.

Results

How well do different people agree on what strokes represent?

Given that our goal was to create an annotated dataset of sketches created under different contexts, we required that the annotations we collected through our interface be reliable. In order to assess this reliability, we looked at whether different annotators saw the same parts in these abstract sketches of objects. Specifically, we looked at inter-annotator reliability in spline labels between participants for each spline in our dataset. Reliability was measured in terms of 'agreement' on spline labels. For example, a 3/3 agreement score for a given spline meant that each of the 3 annotators applied the same label to that spline. We found that 67.85% of splines in our dataset had 3/3 inter-annotator agreement, 27.77% of splines had 2/3 agreement, and 4.38% of splines had no agreement, which means that each participant applied a different label for each of those splines. For the purposes of analysis, we set the modal label for each spline as its true label.

What is the relationship between the parts of an object and the strokes in a drawing?

People's hierarchical organization of visual concepts, such object category membership being determined by its constituent parts, allows for robust recognition of objects in the real world. We were interested in whether people might employ similar abstractions in producing sketches of such objects as well. Since an individual stroke correspond to a person's decision to make a mark on the canvas, we looked at the relationship between strokes and the parts that they represented in our dataset. We explored 3 possible stroke-to-part relationships: A) Singular strokes correspond to singular parts, B) Singular strokes are used to convey multiple parts, that is, strokes cross semantic boundaries, and C) Multiple strokes are required to convey a single part.

We compared A) and B) by looking at within-stroke label agreement for spline labels for all strokes in our dataset. High agreement among all the splines in a given stroke would be indicative of that being used to represent a single part. On the

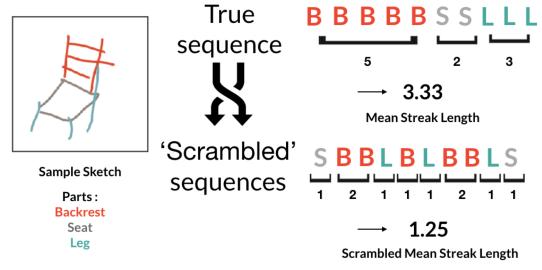


Figure 5: Analyzing stroke sequences in sketches. We first coded each stroke in our dataset in terms of its part label. Whenever multiple consecutive strokes that the sketcher had made shared a part-label, we counted the number of such strokes and termed it as a streak length for that part. This process was repeated for every stroke in a given sketch, after which we averaged over all the streak lengths for every part to obtain a single mean streak length value for every sketch in our dataset. To test whether strokes of the same part were reliably being drawn in bunches we scrambled the order of strokes in each sketch and calculating a 'scrambled' mean streak length. We repeated this scrambling process 1000 times to get a distribution of scrambled mean streak lengths for every sketch.

other hand, low agreement would indicate that stroke crosses semantic boundaries and is used to represent different parts. We found that splines contained in 76.85% of strokes in our dataset shared a single label, 12.75% of strokes contained 2 labels, and less than 11% of strokes contained 3 or more labels. People, in general, tended to use their strokes to draw a single part while only sometimes utilizing a single stroke to represent multiple parts.

We also compared A) and C) by looking at the average number of strokes used to represent specific parts within a given category of sketches. A high average number of strokes for a given part would indicate that multiple strokes are utilized to draw that part, whereas a low average would indicate that a single or few strokes might suffice in depicting that part. Figure ?? shows these part-specific stroke averages by object category. [kushin: I feel the above summary, including the figure caption for figure 2 is a little inconclusive. Thoughts on how to remedy this?]

To what extent are strokes representing the same part produced in succession?

Since individual strokes seemed to mostly correspond to singular part labels, we can view strokes as the building blocks for sketches much like words are the building blocks for sentences. Under this view of stroke organization, we looked at whether there was any meaningful temporal organization of strokes in terms of their part labels. If there was any such organization, any variation between the context conditions would also highlight a difference in how parts are mapped onto strokes under different communicative needs. This investigation was done through a permutation test approach where we created distributions of scrambled stroke sequences to test whether in the true sequence strokes

of the same part were preferentially grouped together. Figure 5 outlines the procedure we undertook for this analysis. 74 sketches were excluded from this analysis because a) they consisted of only a single stroke, b) all the constituent strokes shared the same part label, or c) each stroke in the sketch had a unique label, making the permutation procedure not feasible. We calculated the z-score of the true sequence of every sketch relative to their scrambled distributions. The higher this statistic was, the greater the amount of grouping of similar parts was relative to if the strokes were organized in a random fashion.

The mean z-score for sketches in the close condition was 2.58 (95 CI: 2.26, 2.90) and 1.56 (95 CI: 1.38, 1.74) for the far condition.

Modulation between communicative contexts

[jefan: where we would report analysis of the sketch part features (num strokes, arc length) e.g., when the far sketches are more abstract, how does that manifest in this feature representation? like, are they more similar to each other, more like "bird" and lacking object-specific details? a way of measuring this is that the centroid (euclidean norm, magnitude of the vector) is closer to the origin for far vs. close, and also that the RMSD to centroid of far sketches is smaller than for close sketches....]

Discussion

Acknowledgments

Tables

Figures

References

- Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... others (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.
- Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, 20(1), 38–64.
- Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Trans. Graph.*, 31(4), 44–1.
- Fan, J., Hawkins, R., Wu, M., & Goodman, N. (2018). Modeling contextual flexibility in visual communication. *Journal of Vision*, 18(10), 1045–1045.
- Fan, J. E., Yamins, D. L. K., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science*, 0(0).
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.
- Long, B., Fan, J., & Frank, M. (2018). Drawings as a window into the development of object category representations. *Journal of Vision*, 18(10), 398–398.
- Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3), 353–383.
- Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4), 119.
- 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.
- Yu, Q., Yang, Y., Liu, F., Song, Y.-Z., Xiang, T., & Hospedales, T. M. (2017). Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, 122(3), 411–425.