Semantic structure in communicative drawings

Kushin Mukherjee[1], Robert X. D. Hawkins[2], Judith E. Fan[2,3]
[1]Department of Cognitive Science, Vassar College,
[2]Department of Psychology, Stanford University,
[3]Department of Psychology, University of California, San Diego

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; pragmatics

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.

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 efficiency and interpretability of structured representations that reflect how visual concepts are organized in the human mind (Battaglia et al., 2018). Yet pursuing this strategy clearly 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 *production* of visual semantic knowledge in a naturalistic setting that exposes both its structure and flexibility. This focus on production departs from the conventional strategy for inferring the organization of visual concepts from behavior, which relies upon tasks that evaluate *comprehension* of visual inputs, usually with respect to experimenter-defined dimensions. By contrast, visual production tasks permit participants to include any elements they consider relevant to their goals and compose these elements into a structured whole, yielding high-dimensional information about how visual semantic knowledge is organized and deployed.

Here we investigate one of the most basic forms of visual production, the communication of object concepts via drawing. Specifically, our goal is to understand how the semantic structure of such communicative drawings may shed light on how the structure of high-level object object representations supports their flexible expression across contexts. Our approach builds on prior work that has investigated the production of object drawings to communicate (Fan, Yamins, & Turk-Browne, 2018) in two ways: first, an explicit focus on interpretable semantic structure in drawings, and second, the manipulation of semantic context to investigate flexibility in how perceptual and semantic knowledge is expressed.

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 during visual communication, people readily deploy their abstract understanding of the semantic structure of objects in order to select actions that effectively communicate about them.

Methods

Dataset

In order to investigate

We obtained 1198 sketches for the annotation task.

In this experiment, one participant (the sketcher) aimed to produce sketches of target objects to distinguish them from three distractor objects. had to guess which of the 4 images

the sketch represented. The targets and distractors were chosen from a set 32 real-world objects belonging to 4 basic-level categories: cars, chairs, dogs, and birds. Each category had 8 distinct exemplars. There were 2 main context conditions in this experiment - close and far. In the close condition, the target image and the distractors belonged to the same basic-level category. In the far condition, the target and each of the distractors belonged to a different basic-level category.

These sketches were represented as scalable vector graphics (SVG) images. The strokes that participants made on the canvas when creating the sketch can be represented as a concatenated string of cubic Bezier curves. Thus, the final sketch can be represented by a list of such concatenated strings, each of which corresponds to an event of the participant placing their drawing instrument on the canvas, making some marks on the canvas, and lifting the instrument off of the canvas. We were interested in collecting fine-grained annotations of these strokes, so we split strokes into sub-stroke elements, which we called splines. A single spline was equivalent to a single cubic Bezier curve, i.e., a Bezier curve with two fixed end points and two control points to control curvature. We had participants in our annotation task label each sketch's constituent splines.

Participants

We recruited a total of 326 participants via Amazon Mechanical Turk (AMT). For this experiment, participants provided informed consent in accordance with the Stanford University IRB. Participants were paid a base amount of $0.35 and were given an additional bonus of $0.002 for every stroke they annotated. In addition to this, they were given a $0.02 bonus for every sketch for which they labeled all strokes.

Annotation Procedure

To collect fine-grained annotations of our sketches, we implemented a web-based Javascript annotation tool. Each participant annotated 10 sketches. We provided participants with a sketch to be annotated on a canvas as well as a category-specific menu of labels, which they were encouraged to use for the annotation task. We also provided them with the option of entering their own labels through a free-response box. The original set of images the sketcher had to discriminate between were shown to help the annotator better understand the contents of the sketch. Labeling was done by clicking on individual splines or clicking and dragging across multiple splines to highlight them before assigning them a label. Participants were encouraged to conduct their labeling of strokes in bouts they were to highlight all the strokes corresponding to a single instance of a part before selecting a label from the menu. Participants could do the task at their own pace and continue to a subsequent sketch whenever they felt they were ready. They could choose to continue to the next trial without labeling every stroke in a sketch, but they would lose out on the completion bonus as well as the amount they would have earned for labeling the remaining strokes. In total, we collected 3608 annotations of 1195 unique sketches.

Preprocessing

After collecting annotations, we filtered out any sketches that didn't have all of its constituent splines labeled. This left us with 3319 annotations of 1190 unique sketches. Since there was some variability in the number of times each sketch in our dataset was annotated, we selected those sketches that had been annotated exactly 3 times. This left us with 764 unique sketches, each of which had been annotated 3 times. We also created unique dictionaries for each object category that mapped participant-generated labels to the most frequently occurring labels in our dataset. This helped reduce the total number of unique labels in our dataset from 228 to 24.

Given that our goal was to created 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.

Results

Relationship between strokes and parts

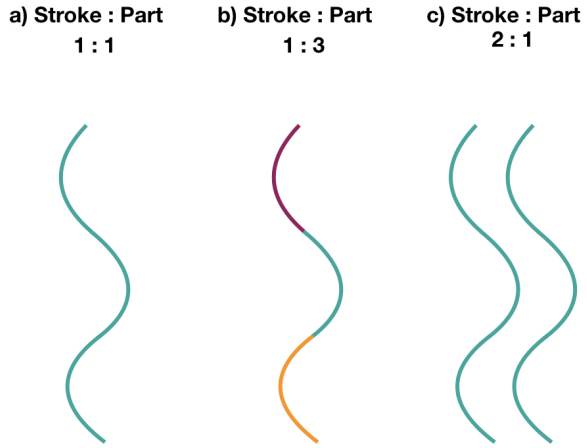
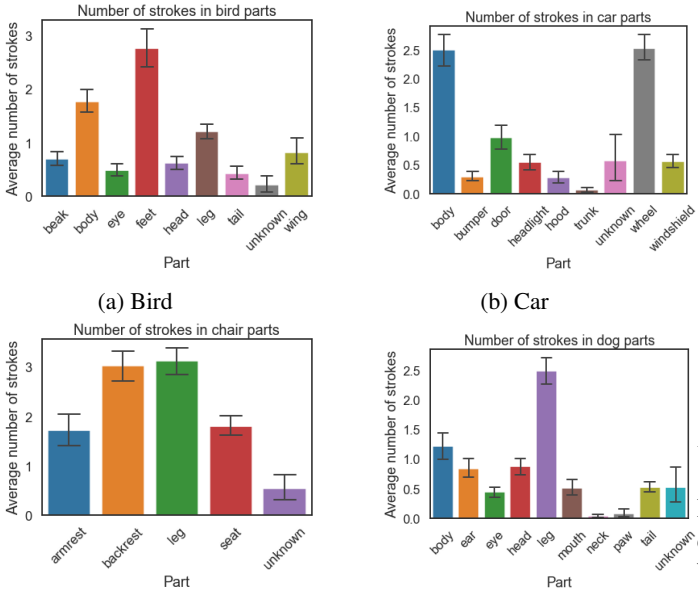


Figure 1: The three part-stroke relationships we explored. Each squiggle represents a hypothetical stroke. Different colors indicate different part-labels.

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



(a) Bird (b) Car

(c) Chair (d) Dog

Figure 2: Average number of strokes used to draw each part per category. While participants use multiple strokes to represent some parts, other parts are sometimes expressed using single strokes.

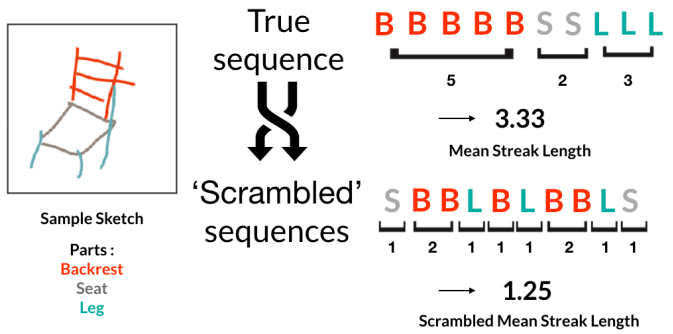


Figure 3: Analyzing stroke sequences in sketches. We first coded each stroke in our dataset in terms of its part label. Whenever multiple consecutive strokes that the sketcher had made shared a part-label, we counted the number of such strokes and termed it as a streak length for that part. This process 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.

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 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 2 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?]

Stroke sequence organization

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 3 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

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*.

Fan, J. E., Yamins, D. L. K., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science*, *0*(0). Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12676` doi: 10.1111/cogs.12676

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*.

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.