Communicating semantic part information in 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

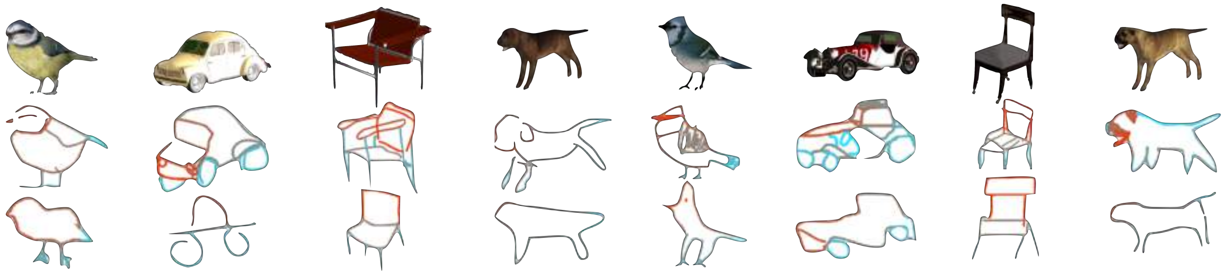

Figure 1: Objects used in communication game with example drawings of each object, where stroke color indicates different parts.

Abstract

We effortlessly grasp the correspondence between a drawing of an object and that physical object in the world, even when the drawing is far from realistic. How are visual object concepts organized such that we can both recognize these abstract correspondences and also exploit them when communicating them to others in a drawing? Here we consider the notion that object concepts are compositional, such that we readily decompose both objects and drawings of objects into a common set of semantically meaningful parts. To investigate this, we developed a web-based platform to densely annotate drawings of real-world objects with part information. Our dataset contained both detailed and sparser drawings produced in different communicative contexts. 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 during production; and (4) both more detailed and sparser 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 sparser drawings. Taken together, our results support the notion that people flexibly deploy their abstract understanding of the compositional part structure of objects 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; compositionality; objects and categories

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). How are visual object concepts organized such that they can robustly encode such abstract correspondences? Here we explore the notion that perceiving these correspondences is supported by our ability to decompose both objects and drawings into a common set of semantically meaningful parts (Biederman & Ju, 1988).

Recent advances in computational neuroscience have provided an unprecedentedly clear view into the algorithms used by the brain to extract semantic information from raw visual inputs, including drawings, exemplified by modern deep learning approaches (Yamins et al., 2014). Nevertheless, a major gap remains in adapting such deep learning models to emulate the structure and flexibility of human semantic knowledge (Lake, Ullman, Tenenbaum, & Gershman, 2017). A promising approach to closing this gap may be to exploit the parsimony and interpretability of structured representations that reflect how visual concepts are organized in the mind (Battaglia et al., 2018).

However, pursuit of this strategy relies upon a thorough empirical understanding of this conceptual organization and how people express this knowledge in natural behavior. We aim 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. This approach departs from the conventional strategy for inferring the organization of visual object concepts, which entails eliciting 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 and combine these elements freely, yielding high-dimensional information about how people organize and deploy visual semantic knowledge under a natural task objective.

The goal of this paper is to shed light on the correspondence between visual semantic knowledge and the procedure by which people robustly convey this knowledge in their drawings. This paper advances recent work (Fan, Yamins, & Turk-Browne, 2018) investigating how drawings convey semantic information in three ways: *first*, we introduce a new dataset of drawings with dense part annotations, allowing an explicit focus on compositional part structure, *second*, we explore the link between this semantic structure and the temporal dynamics of drawing production, and *third*, we examine how different aspects of visual semantic knowledge is expressed in different contexts.

Methods

Toward this goal, we developed a web-based platform (de Leeuw, 2015) to collect dense semantic annotations of the stroke elements in drawings of real-world objects (Fig. 1).

Communicative drawing dataset

We obtained 1195 drawings of 32 real-world objects from a recent experimental dataset in which pairs of participants played an interactive drawing-based reference game (Fan, Hawkins, et al., 2018).[1] Object stimuli were photorealistic 3D renderings belonging to one of four basic-level categories (i.e., bird, car, chair, dog), each of which contained eight exemplars. On each trial of the experiment, participants were presented with a shared context containing four of these objects. One participant (the sketcher) was privately cued to draw a target object so that the other participant (the viewer) could pick it out from the set of distractors. Across trials, the similarity of the distractors to the target was manipulated, yielding two types of communicative contexts: *close contexts*, in which all four objects belonged to the same basic-level category, and *far contexts*, in which objects belonged to different basic-level categories. 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.

Prior work analyzing the semantic properties of drawing data has used a raster image representation (e.g., `*.png`), an expedient format for applying modern convolutional neural network architectures (Fan, Yamins, & Turk-Browne, 2018; Sangkloy et al., 2016; Yu et al., 2017). However, this representation is severely limited for our goal of characterizing how semantic structure manifests during drawing production. Specifically, it was critical to encode each drawing using a vector image format that preserves the inherently sequential and contour-based nature of drawing production (i.e., `*.svg`). Thus, each drawing in our dataset is represented as a sequence of individual strokes, where each stroke consists of a sequence of sub-stroke elements. These sub-stroke elements are parameterized as cubic Bézier *splines*. This format provides a compact representation preserving the sequence in which each element was produced.

Semantic part annotation

We crowdsourced semantic annotations for every spline in every stroke of the drawings from this dataset.

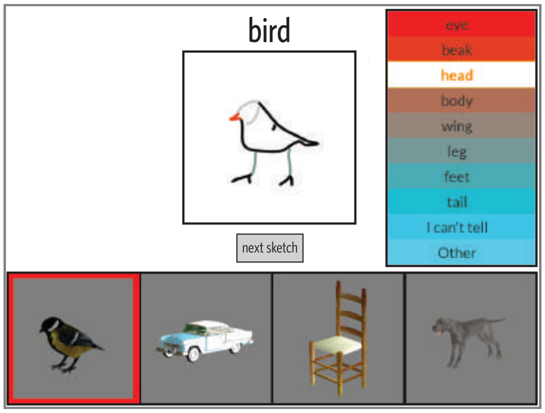

Figure 2: Drawing annotation interface. Participants selected sub-stroke elements (splines) and tagged them with part labels.

Participants 326 participants were recruited via Amazon Mechanical Turk (AMT) and provided informed consent in accordance with the IRB. Participants were given a base compensation of $0.35, plus $0.002 for every spline they annotated and $0.02 for every drawing they annotated completely.

Task procedure Each participant was presented with a sequence of 10 drawings that were randomly sampled from the communicative drawing dataset (Fig. 2). Their goal was to tag each spline with a label corresponding to the part it represented (e.g., seat, leg, back for a chair). To facilitate consistent tagging, participants were provided with a menu of common part labels that were associated with each basic-level category (Table 1). Participants could also generate their own part label if they believed none of the common labels applied. To give participants full information about the original communicative context, we showed the drawing with the same array of four objects that the original sketcher had viewed, with the target object highlighted in red.

Data preprocessing To reduce bias due to missing data, we restricted our analyses to annotation trials in which the drawing was completely annotated (i.e., all splines were tagged). Moreover, to ensure the reliability of our annotations, we only examined drawings that were annotated by three distinct participants. Our final preprocessing step was to incorporate manual part labels provided by participants (i.e. labels not in the menu). Some of these labels were valid, but were synonymous with or subsumed by other more frequently occurring labels. For example, some strokes that represented subparts of the leg of a chair were labeled as 'leg support', 'foot', and 'strut.' In order to ensure that parts of drawings were segmented at a consistent level of granularity, we manually

category	**part labels**
bird	eye, beak, head, body, wing, leg, feet, tail
car	bumper, headlight, hood, windshield, window, body, door, trunk, wheel
chair	backrest, armrest, seat, leg
dog	eye, mouth, ear, head, neck, body, leg, paw, tail

Table 1: Part labels provided to annotators.

[1] All materials and data will be made available following the completion of the review process.

constructed a part dictionary to map these less-frequent part labels to one of the common part labels. After applying these preprocessing steps, our annotated dataset consisted of 864 drawings that had been annotated exactly three times each, using a set of 24 unique part labels.

Results

How well do viewers agree on what strokes mean?

Before proceeding to use our annotations to examine the semantic structure of drawing production, we conducted a basic check of inter-annotator consistency. How well did different annotators agree on what each spline in a drawing represented? We found that 95.6% of all splines received the same label by at least two of the three annotators, and 67.8% of all splines received the same label by all three annotators. This shows that the way viewers interpret which part each stroke represents is highly systematic, providing validation for our general approach. Further, it suggests that sketchers may have exploited this systematicity to produce strokes they expect viewers to interpret consistently. In subsequent analyses, we collapsed over inter-annotator variation: we assigned the modal label to splines to which at least two annotators had given the same label; for the remaining 4.4% of splines, we sampled one of the three labels provided.

How do strokes correspond to parts of objects?

When composing a recognizable drawing of a real-world object, how do sketchers decide what information to convey with each stroke? A natural possibility is that their actions closely correspond to the part structure of that object. Concretely, we hypothesized that most strokes in our dataset would *not* cross part boundaries: that all splines within a given stroke would be assigned the same part label. Conversely, because depictions of parts can be arbitrarily detailed, and some parts re-occur throughout an object (e.g., multiple legs on a bird, chair, or dog), we often expected to find more than one stroke per part (Fig. 3A).

To evaluate the first hypothesis, we computed the number of unique part labels across all splines within each stroke. We found that for 81.6% of the strokes in our dataset there was only one part label; the remaining 18.4% of strokes were associated with two or more labels (Fig. 3B). In other words, most strokes represented exactly one part, but in a minority of cases they spanned multiple parts (e.g., a single stroke connecting the head and body of a bird, or an armrest and leg of a chair). We were concerned, however, that these proportions were inflated by strokes with very few splines[2]. To address this concern, we constructed a null model controlling for the number of splines. Part labels were randomly sampled from the full list of parts in the drawing such that each spline was equally likely to represent any part regardless of the stroke it belonged to. In simulations from this null model, only 55% of strokes corresponded to a unique part while 45% of strokes spanned multiple parts. Thus, strokes in our dataset were much more likely to fall within part boundaries than would be expected by composing strokes at random.

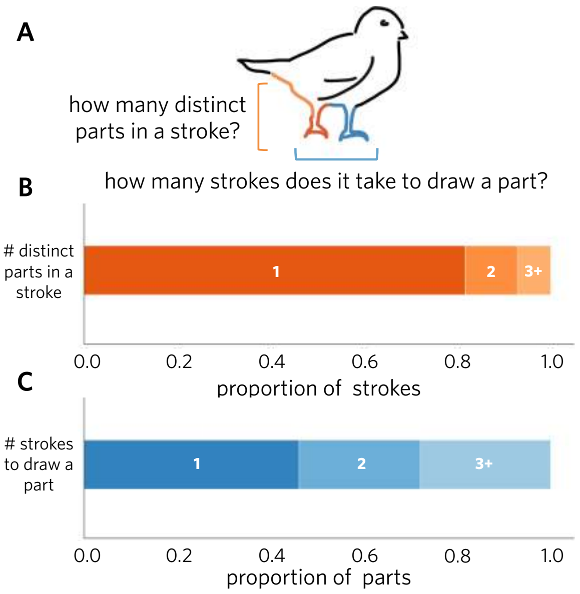

Figure 3: (A) To investigate the correspondence between strokes and part labels, we estimated number of unique part labels assigned to different splines within the same stroke, as well as number of different strokes used to draw each part. (B) Distribution over number of unique part labels within a stroke. (C) Distribution over number of strokes used to draw a part.

To evaluate the second hypothesis, we computed the number of strokes that were used to represent each part of an object (Fig. 3C). We found that 46.1% of parts were depicted using exactly one stroke, 26.0% using exactly two strokes, 11.3% using exactly three strokes, and 16.6% using four or more strokes. Thus, nearly half the time, a single action was sufficient to depict an entire object part. However, the remaining 53.9% of parts required more than one stroke to depict, which would be expected for those parts that consisted of multiple disconnected subparts within an object (e.g., wheels of a car, paws of a dog).

Together, these findings suggest that the information people convey with each stroke systematically corresponds to the part structure of objects. We suspected, however, that these properties may vary across communicative contexts. Indeed, strokes spanning multiple parts were slightly more common in drawings produced in far contexts (19.4%, CI: [17.9%, 20.9%]) than close contexts (17.6%, CI: [16.1%, 18.8%][3]). And the proportion of parts requiring more than one stroke was slightly higher for close drawings (55.8%, CI: [53.7%, 58.6%]) than far drawings (52.0%, CI: [49.9%, 54.6%]). These differences suggested that sketchers were somewhat more likely to use a single stroke to represent multiple spatially contiguous parts in a context that did not require them

[2] The modal number of splines per stroke (20% of cases) was 1, but there was a long tail; the mean number was 2.6.

[3] 95% confidence intervals were estimated via stratified bootstrap resampling (N=1000 iterations) of drawings within each object-context combination.

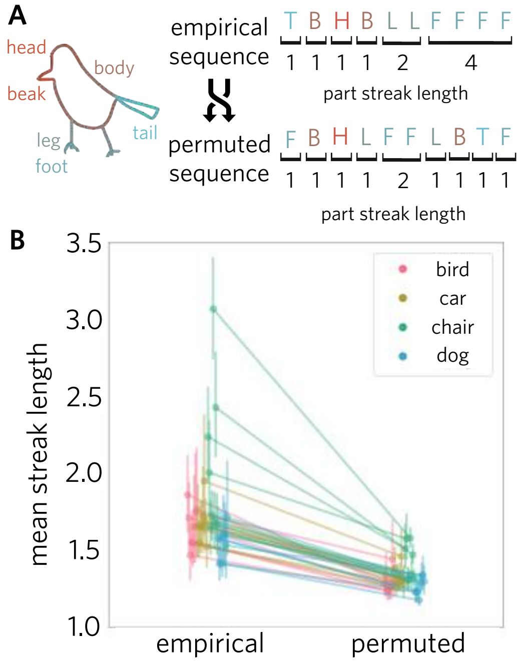

Figure 4: (A) Analysis of sequence in which strokes depicting each part were drawn. (B) Comparison of mean length of streaks consisting of strokes that depict the same part with null distribution of permuted stroke sequences.

to produce a highly detailed drawing and may have included more detail per part to distinguish the target object from similar distractors.

Do strokes representing the same part tend to be produced in succession?

In the previous section we discovered that slightly more than half of the parts in our dataset were depicted using multiple strokes. This result raised the question: to what extent are these multiple strokes depicting the same part being drawn in succession, or being interleaved among strokes corresponding to other parts? In other words, how does the *temporal sequence* of stroke production reflect knowledge of part structure?

To investigate this question, we estimated the mean length of stroke "streaks" depicting the same part. First, we collapsed across the spline annotations examined in the previous section and represented each stroke by the modal part label assigned to its splines. We then encoded each drawing as a sequence of strokes in the order they were produced by the sketcher, and defined a "streak" as a cluster of more than one stroke annotated with the same part label occurring in succession[4]. For example, in the drawing shown in Fig. 4A, two 'leg' strokes were placed before moving on to the 'foot', giving a streak of length 2. Finally, we averaged these streak length values over every drawing in the dataset to obtain our statistic.

To evaluate whether the empirical stroke sequences were more structured than expected from placing strokes at random, we constructed a null model to serve as a baseline. For this null model, we permuted the stroke sequence such that the raw number of strokes in each part was preserved, but any structure in their order was disrupted (Fig. 4A). We then generated a null distribution of streak lengths for each drawing by repeating this permutation procedure 1000 times and measuring the mean streak length for each permutation. Finally, we obtained a z-score for each drawing by computing where the empirical streak length fell in the permuted streak length distribution. A drawing with a z-score near 0 had a streak length that was commonly obtained by placing strokes in a random order, while a drawing with a higher z-score is more structured than expected under the null.

We found that the empirical streak length was reliably higher for all objects than that of the permuted sequences (z-score: 2.07, CI: [1.90, 2.23]; Fig. 4B), and higher for the close drawings (z-score: 2.58; CI: [2.26, 2.90]) than far drawings (z-score: 1.56; CI: [1.38, 1.74]). The lower streak length for far drawings is consistent with their lower stroke count overall—when only one or two strokes are used per part, there is a ceiling on possible streak lengths. However, when sketchers do use multiple strokes to convey a single part (i.e., because there are multiple subparts, or to add more detail), they tend to draw these in succession before moving on to a different part. These results suggest more broadly that the procedure by which people convey semantic information in drawings is organized by the part structure within objects.

How is part information emphasized in different communicative contexts?

Our findings so far bear on how the composition of communicative drawings of objects reflects the sketcher's semantic knowledge of the (part) composition of those objects. A key consequence of such semantically organized part knowledge is that it naturally supports flexible expression across different communicative contexts. For example, when communicating about a chair in a far context containing objects from other basic-level categories, sketchers may include only the essential information to indicate the presence of certain parts that distinguish it at the category level. On the other hand, when communicating about that same chair in a close context containing other, perceptually similar chairs, sketchers may emphasize aspects of parts that distinguish it at the object level (e.g., the number of back slats), by applying more strokes and/or more ink in each stroke.

We hypothesized that sketchers emphasize part information to preserve relevant distinctions in context. To explore this possibility, we asked the following questions: (1) How similarly is object-specific part information emphasized in close and far communicative contexts? (2) How do differences in how part information is emphasized in different contexts affect the discriminability of drawings?

[4] We excluded 78 out of the 864 drawings where this measure was not well-defined, i.e. sketches that contained only one stroke or part label, or that contained fewer than two strokes sharing the same part label.

To investigate these questions, we represented each drawing by a *part-feature vector* that combined information about: (a) how many strokes and (b) how much total ink was expended on each part of that object. In order to represent all drawings in our dataset using a common feature representation, we combined part labels across categories, yielding a set of 24 unique part labels to which any stroke could be assigned. Each part-feature vector thus consisted of 48 elements: 24 of these represented the number of strokes allocated to each part, and the remaining 24 represented the total arc length of all strokes allocated to each part. Before further analysis, we z-scored values within each feature dimension in order to map stroke count and arc length measurements to the same unit-variance scale. Because our primary goal was to understand differences between objects and contexts, we then collapsed across drawings within each object-context combination, yielding 64 average part-feature vectors (i.e., 32 objects x 2 context conditions).

Similar part information emphasized across different communicative contexts In order to investigate to what extent similar object-specific part information is emphasized in different communicative contexts, we computed the matrix of Pearson correlations between part-feature vectors. Formally, this entailed computing: $R_{ij} = {cov(\vec{r}_i,\vec{r}_j)}/{\sqrt{var(\vec{r}_i)\cdot var(\vec{r}_j)}}$, where $\vec{r}_i$ and $\vec{r}_j$ are the mean part-feature vectors for the ith and jth object-context combinations, respectively.

While close and far drawings of an object differ in their overall amount of detail, we hypothesized that they would still emphasize part information in similar ways. Specifically, insofar as similar, object-specific part information is emphasized in both close and far drawings of the same object, we predicted higher correlations between close and far part-feature vectors for the *same* object than for close and far part-feature vectors of *different* objects. Consistent with this, we found strong correlations between the part vectors of close and far drawings of the same object (within-object: $r =$ 0.740, CI: [0.726, 0.753]), which were significantly stronger than close and far drawings of *different* objects (between-object: $r = 0.653$, CI: [0.646, 0.659]). These results suggest that close and far drawings emphasize similar parts and that these patterns of emphasis were to some extent object-specific (Fig. 5B).

Detailed drawings are more distinct from each other than sparser drawings The above findings show that close drawings exhibited similar patterns of emphasis on different parts as their far counterparts. However, given that close drawings contain greater emphasis on these parts overall than far drawings (i.e., contain more and longer strokes), how are these additional strokes being spent?

We hypothesized that the additional part information provided in close sketches enhances the *discriminability* of close drawings from one another, relative to far ones, by increasing the feature distance between objects that otherwise share many perceptual properties. To evaluate this possibility, we computed the mean correlation distance (i.e., $1-r$) between the part-feature vectors of *close* drawings of objects in a given category. We compared this value with the distance between *far* drawings of exactly the same objects. We found that close drawings were reliably more distant from one another than far drawings were (close similarity: $r = 0.67$, CI: [0.65,69]; far similarity: $r = 0.73$, CI: [0.72,0.75]), suggesting that sketchers discern which part information is most diagnostic of the target object among highly similar distractors, and emphasize this distinctive information accordingly (Fig. 5A & C).

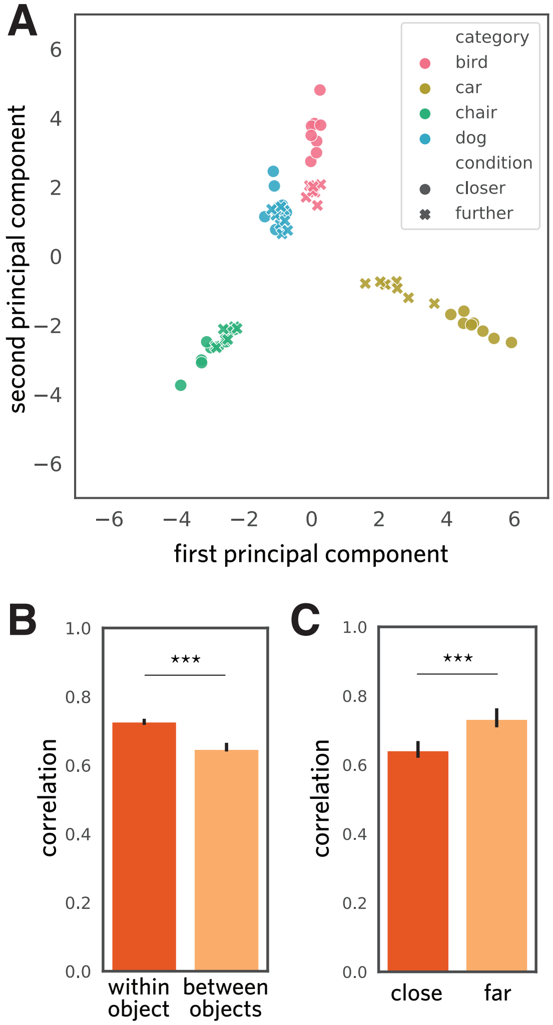

Figure 5: (A) Visualization of the distribution of mean feature vectors for each object-condition combination, projected onto the top two principal components. (B) Feature vectors of close and far drawings of the same object were more similar to each other than those of close and far drawings of different objects within a category. (C) Far drawings were more similar to each other than close drawings were to each other. *** $p < 0.001$.

Discussion

In this paper, we explored how the way people compose communicative drawings of objects reflects their semantic knowledge about what objects are composed of. To accomplish this, we developed a novel web-based platform to collect dense semantic annotations of the stroke elements in drawings of real-world objects that were produced in different communicative contexts. 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.

These findings are resonant with classic and recent work that has argued for the importance of compositionality in human perception and cognition in general (Biederman, 1987; Battaglia et al., 2018; Lake et al., 2017), and for visual production in particular (Lake, Salakhutdinov, & Tenenbaum, 2015). However, unlike prior work which focused on the production of abstract symbols (Lake et al., 2015), we consider the challenge of how people transform perceptually grounded representations of real-world objects into procedures for producing figurative drawings that communicate not only what they see and know about them, but also what is relevant.

Our work is also related to recent progress in the development of computational models of drawing production (Ha & Eck, 2017; Li, Lin, Mech, Yumer, & Ramanan, 2019). While results from these efforts have been galvanizing, the development of principled metrics by which to rigorously evaluate how well they emulate human drawing behavior has not kept pace. By interrogating in detail how humans encode semantic information into their drawings, and flexibly adjust their production behavior in different contexts, this paper presents a first step towards such a set of behavioral metrics. Having such metrics is important because they would enhance our ability to distinguish between generative models, and thereby help advance further model development.

In future work, we plan to extend our analysis of how different part information is expressed in drawings beyond simple effort cost measures (i.e., number of strokes, amount of ink) to encompass content and style information (e.g., the shape of a drawn bird's wing, caricaturization of the curvature of a chair leg). We expect that this will entail augmenting current vision models with the requisite compositional semantic part knowledge to parse drawings in a more human-like way. More broadly, achieving this synthesis will lead to both more robust artificial intelligence as well as more unified theories of human cognition and behavior.

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

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, *20*(1), 38–64.

de Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*, 1–12.

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.

Fan, J., Yamins, D., & Turk-Browne, N. (2018). Common object representations for visual production and recognition. *Cognitive Science*.

Ha, D., & Eck, D. (2017). A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

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*.

Li, M., Lin, Z., Mech, R., Yumer, E., & Ramanan, D. (2019). Photo-sketching: Inferring contour drawings from images. *arXiv preprint arXiv:1901.00542*.

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.