**Placeholder Title - Sketching as a window into our semantic knowledge of objects**

Kushin Mukherjee

Vassar College

**Placeholder Abstract -**

**Introduction**

Drawings predate written language as the oldest explicit record of human thought. Figurative cave paintings discovered in a cave in Borneo are dated to be at least 40,000 years-old and are the oldest remaining instances of prehistoric human art (Sample, 2018). Due to poor documentation, it is difficult to ascertain why these early drawings were made. One possible explanation is that they were made to document important animals, who were a source sustenance and clothing. It is also possible that these paintings were purely decorative, made to celebrate hunting hauls. Regardless of intent, given the figurative nature of these paintings, it is difficult to imagine that they were made for purely aesthetic purposes without any regard for the content they represented. The fact that cattle were depicted in these paintings is significant insofar they must have held meaning for these early communities. Thus, much as it is for modern day paintings, communication was at the heart of these prehistoric painting endeavors. Given the prevalence of graphical communication from ancient times, it is worth investigating how we communicate via drawings and sketches, and what perceptual and cognitive mechanisms allow us to do so almost effortlessly.

Drawing, an ostensibly simple task, is loaded with difficult questions, which cognitive scientists have tried to tackle over the last few decades. A fundamental problem is one of choice. Given that most graphical productions are necessarily sparser in terms of visual detail compared to the objects or scenes they represent, how do people know what to include in their sketches and graphs in order to communicate effectively. Efficacy here has to do with an unambiguous conveyance of the object, scene, or idea we wish to communicate to others. One way to mitigate ambiguity is to of course have drawings correspond to the original target as much as possible. In the case of maps, another form of graphical production, their utility lies in having a close-correspondence to the ground-truth geographical information. There are also a wider subclass of production problems including how to convey luminescence, depth, relative-scale, and texture, but these can be grouped together under the correspondence problem.

On the recognition side, people should be able to see and interpret drawings in much the same manner as the original target. This calls for some flexibility in how our visual system creates mid to high level representations of visual objects and scenes. It has been argued that observational sketches that people make correspond to the lowest 2½ D representation of visual surfaces in Marr’s model of the visual system. Additionally, it is the 3D representation from the same model that is supposedly utilized for drawing from memory. Caricatures, cartoon-like drawings of people, seem to be just as effective in conveying the identity of a person as a digital photograph. So, we do seem to be able to contend with some level of abstraction in perception. This loosens the need for an extremely tight correspondence between real-world object and graphical production. However, the question remains as to what mechanisms facilitate this flexibility. And given that sketching is a skill that can be honed, what correspondence is there between production and perception?

There are two main modes of artistic drawing, as hinted from the earlier comment about the Marr model— articulating perception and drawing from memory. [ Want to expand this section a bit. Not sure if this is the best place to do so.]

[The following sections provide some scaffolding, helping to build up to a rationale for the experiment]

**Graph and Diagram-based communication**

Correspondence of drawings to ground-truth objects as mentioned earlier is less of a general problem than it is a requisite of certain kinds of graphical productions. A flowchart of steps needed to complete some task or a diagram of a football play drawn on a chalkboard can get by with symbols that are both abstract and, to some degree, arbitrary. Here, space on a canvas is allocated to create these different symbols and Tversky, Zacks, Lee, and Heiser (2000) al refer to this as a categorical use of space. Maps on the other hand utilize pictorial space to represent real-world space. Hence some meaningful correspondence must exist between the two, and it does in the form of map scales — ratios of ground distance to map distance. This is a stricter, ratio-based use of space. However, specific communicative requirements determine to what extent space is strictly represented. Often in creating directional maps, people have little regard for spatial accuracy, instead opting to focus on intersections and turns. They may also choose to represent landmarks on such maps. This is more akin to the categorical use of space, with ordinal arrangement of turns and landmarks. Subways and train routes often choose to apply such a graphical representation.

Besides the manner in which space is allocated, the form in which said space is constructed is equally important in communication. Dreyfuss (1984) in his authoritative guide to international graphic symbols, categorizes symbols based on their graphic forms. He states 14 such forms such as circles, lines, blobs, and squares. These simple forms appear to be useful for the flexibility their representational ambiguity grants them. A line can represent a connection in a circuit, a road, a barrier between two areas, or in the case of simplest of stick-figures, a torso. If viewers of a graphical production have adequate context, these simple forms can be exploited for effective communication. The varied use of such simplified forms is clearly not prescriptive as the average person is not familiar with Dreyfuss’s works. In general, we are able to extract meaningful components from ideas and objects we wish to represent and produce sketches using these simplified graphical forms as stand-ins for these components.

The precise manner in which we map concepts to drawings is mediated by the form and context. Certain forms or shapes are more conducive to representing certain kinds of objects and concepts and the specific context they are drawn in help reduce their representational ambiguity.

Perhaps one of the most versatile schematic forms is the arrow. It is noteworthy due to its ability to dynamize otherwise static drawings. Unlike modern day animations, individual graphical productions are purely spatial representations, containing little temporal information. Ordinal arrangement of forms is one way to convey time, but it is a fairly limited one. Arrows help viewers infer the causal chain of events in a diagram by augmenting the form of a diagram with directional cues. Arrows can indicate direction of movement, increase or decrease in a quantity, the sequence of events, and causal outcomes of actions to name a few. Even without a mathematical background one can easily discern the difference in magnitude of two vectors, represented as arrows, due to our ability to discern differences in lengths. Much like paralinguistic cues supplement linguistic communication, versatile forms such as arrows help better facilitate communication through graphical productions. The use of arrows speaks to the metaphorical strength of graphical productions. [I want to expand this section a bit after I read Lakoff’s metaphor book]

Graphical productions have a wide series of applications in communication due to their versatility in representing objects, scenes, and processes through systematic usage of space and forms. Spoken language has been described as a seminal part of human cognitive development because it conferred us the ability to talk about events beyond those that are currently occurring. We can describe events that have already happened and speculate about events that may occur in the future. Similarly, in the case of graphical productions, they provide a way for us to describing elements and relationships that aren’t necessarily visuo-spatial (Tversky et al, 2000).

But why do these strategies of making graphical productions come so easily to us? How is it that even with no formal training, toddlers are able to partition space on a canvas to create a sketch of an object?

Given that the end goal of diagrams and sketches is to communicate ideas visually, the next section focuses on what kinds of representations may guide our perceptual systems.

[May expand this section to talk a bit about how graphs are used for conveying quantitative information]

**Visual Representations**

There is immense variation in the kinds of visual input the human perceptual systems are subjected to. Yet, from behavioral experiments it can be seen that people are really good at sorting objects into categories, even novel objects.? Beginning with Marr’s computational level of analysis, we ask what kind of representational model would best explain such a phenomenon? A compelling candidate is Biederman’s (1987) recognition-by-components (RBC) model. According to this theory, perceptual recognition of objects is facilitated through segmenting the retinal image at points of surface concavity and subsequent approximation of those segments into simple conic sections called geons. These geons are constrained in number, (<36) and they facilitate a matching-like system where the perceived combinations of geons are matched to internal models of geon combinations to facilitate recognition. This is a compelling model because it flexibly accounts for partial-object recognition, situations where the original object is partly occluded. The statistical regularities of co-occurrence of certain geons of the object would suffice in classifying such objects. The approximation of segments into geons helps account for ‘noisy’ visual input and does not call for a precise quantitative understanding of the spatial properties of visual objects.

[Further expansion of this section with modern updates to the Biederman model and commentaries on this approach. Also descriptions of common representations for production and recognition, Fan et al 2018/Robinson and Rolls 2015/Barsalou 1999]

**Language and Sketching**

Graphical productions are more transparent than written language in their ability to convey concepts and ideas. Whereas one needs at least a rudimentary understanding of any script to comprehend a paragraph written in said script, a study by Hochberg & Brooks (1962) showed that even when a child has no prior experience viewing pictures, they can easily identify simple contour drawings. There is a certain universality that the graphical production possesses, which linguists have had a difficult job describing in the case of language. That being said, there do appear to be some meaningful parallels between language and graphical productions in their development of and structure. And they share the functionality of being communicative media.

Much like there is a critical period for language acquisition, after which it becomes difficult for adults to acquire secondary and tertiary languages, Kindler & Darras (1997) describe how artistic development also stagnates after the age of 11, accounting for certain kinds of socio-cultural factors. This stagnation is not a robust effect, failing to manifest in Japanese culture (Cohn, 2007) likely due to the ubiquity of manga culture in popular media. Manga accounts for one third of all printed materials in Japan, and the common practice of copying drawings from manga may account for this difference (Toku, 2001). In general, spoken and written language are our main modes of communication, resulting in ‘practice’ and development of our language skills over time. In contrast, aside from formal drawing classes, people have limited opportunities to practice drawing. Cohn (2012) argues that development of drawing follows a similar trajectory to that of language. Language is a referential system because it maps systematic groups of sounds and written symbols to express concepts and ideas. Similarly, for graphical productions, the strokes on a canvas or the engravings on a stone tablet are to some degree referential too, though they may appear to be iconic. In fact, what most people consider to be prototypical drawings of objects are highly conventionalized to specific cultures or contexts.

Such systematicity is akin to how languages become hyper-conventionalized over several generations of speakers speaking them. Therefore, people encode some degree of propositional knowledge in their viewing of graphical productions, which in turn inform their own productions. Such a view precludes a purely perceptually-deterministic view of even observational drawing. Much like we have a lexicon of words that are combined in innumerable ways to convey distinct, novel ideas, so are conventionalized drawing schemas. These schemas are combined and modified in production to create unique sketches and diagrams. Cohn’s (2012) description of a cognitive model of drawing relies on such a lexicon of schemas, as well as a Marr-inspired perceptual system and a production script to facilitate the task of drawing. Whereas the lexicon consists of primary graphemes such as lines, curves, and dots, it also contains composite graphemes such as body parts and prototypical versions of common houses. The production script links the cognitive mechanisms to motor actuation, and accounts for how perception and lexical items influence the sequence in which graphemes are put to surface. The takeaway here is that such a cognitive model relies on graphemes that are built up through referential conventions, much like the structure of language changes through iterated convention formation.

The greatest differences between language and drawing lies in comprehension. Whereas language must be linearly processed, be it written or spoken language, drawings are spatially preserved. Though they must be produced in a temporally sequential manner as with language, it is rare in nature that comprehension involves the viewer being witness to the temporal unfurling of the individual strokes that constitute the drawing. Secondly, while language and drawing may both be referential systems, spoken language uses inherently arbitrary phonetic symbols while written language uses indexicals to refer to these symbols. There are some exceptions to the latter, like in pictorial writing systems. But these too undergo severe conventionalization to the point that modern incarnations of pictorial scripts like Chinese have symbols that are more indexical than iconic. However, due to the partially iconic nature of drawings, purely perceptual knowledge can be leveraged to interpret them. This is in contrast to language, which requires an understanding of conventions and grammar for comprehension. While an understanding of (grapheme) conventions aids in visual comprehension, it is not as strict as a requirement as it is in the case of language. Drawings are, thus, an extremely versatile vehicle of communication.

Goldin-Meadow (2003) described core resilient features of language as well as fragile features that are built around this core through development. Cohen (2012) says that some of these core features may not be domain-specific to language, and that drawing shares some of these more biologically based features. One feature highlighted was the coarse mapping of form to meaning. This accounts for our people’s capacity to draw basic shapes and figures despite a lack of formal training. Such training imparts fragile language-like qualities to graphical production, which may override the core features. Some modality-specific techniques include the use of strokes to denote luminescence and contrast, as well as the relative partitioning of the canvas space into different objects of varying shapes to imply perspective.

The implications of drawing such equivalences between graphical production and language, a fundamental aspect of human cognition, are in broadening how skills such as sketching are viewed. Instead of being thought of as a highly specialized skill that only artists have access to, sketching should be viewed as a core aspect of human cognition, which is heavily reliant on our perceptual abilities and can be nurtured much like language. And much like language is studied to gain further insight into how knowledge is structured in the human mind, so should drawing.

**Modern Models of Vison** \* Might not include this depending on how lit review goes

**The Current Experiment**

**Why?**

[This section needs major edits; placeholder material is an edited version of CogSci paper abstract. Feedback would still be incredibly helpful. Rationale would describe how such annotation approaches have been used in computer vision to create powerful segmentation algorithms like MASK R-CNN, and how this approach can help test some of the RBC statements.]

Drawing is a versatile tool for communication, spanning detailed renderings and simple sketches. Even the same object can be drawn in different ways, depending on the context.

How do people decide how to draw in order to be understood?

Here we investigate the semantic structure of drawings as a window into how people deploy both perceptual information and conceptual knowledge to produce communicatively effective drawings in context.

We analyzed a dataset containing drawings of real-world objects that were produced in different semantic contexts, and contained both detailed and simpler sketches of each object.

We explored the hypothesis that during visual communication, people spontaneously decompose visual objects into semantically meaningful parts (e.g., chairs consist of legs, seat, and back), resulting in a tight correspondence between the organization of this semantic part knowledge and the procedure people use to sketch an object.

For example, if someone aims to produce a recognizable sketch of a chair, they produce strokes that represent individually meaningful parts, e.g., seat, armrest, legs.

To investigate this, we developed a web-based platform to collect dense semantic annotations of the stroke elements in each drawing.

**Methods**

**Dataset**

We required sketches of the common objects created under different contexts to begin to understand the semantics that underlie sketch production.

So, we obtained sketch data from a two-player 'Pictionary'-style reference game experiment.

In this experiment, a 'sketcher' aimed to produce sketches of target objects to distinguish them from three distractor objects.

A 'viewer' 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.

We obtained 1198 sketches for the annotation task. 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. The mean payout for the task was $\_\_.

**Annotation Procedure**

To collect fine-grained annotations of our sketches, we implemented a web-based Javascript annotation tool. We structured this tool as an annotation game where participants labeled sketches to create colorful annotated sketches. Each participant annotated 10 sketches in this annotation game.

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

**Analysis**

Inter-Annotator

How much agreement is there of what a spline represents between different annotators? That is, do different people see the same parts in abstract-ish sketches?

Relationship between strokes and parts

Was there a need for us to collect data at the spline-level, or is it so that the individual strokes people make correspond to singular parts?

Do people need to make multiple strokes to represent a single part/instance of a part?

Stroke sequence organization

Is there any semantically meaningful manner in which people organize their strokes when sketching? Do people draw all the stroke belonging to the same part in a 'streak'?

Modulation between communicative contexts

where I would report analysis of the sketch vector features (num strokes, arc length)

e.g., when the far sketches are more abstract, how does that manifest in this feature representation?

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.

**Results**

**References**

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

Cohn N. (2012). Explaining ‘I can’t draw’: Parallels between the structure and development of language and drawing. *Human Development, 55*. 167-192.

Dreyfuss, H. (1984). *Symbol sourcebook: An authoritative guide to international graphic symbols.* N. Y.: John Wiley & Sons.

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

Hochberg, J., & Brooks, V. (1962). Pictorial recognition as an unlearned ability: A study of one child’s performance. *American Journal of Psychology, 75*, 624-628.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Toku, M. (2001). What is manga? The influence of pop culture in adolescent art. *Journal of National Art Educatio*n, *54,* 11–17.

Tversky, B., Zacks, J., Lee, P., & Heiser, J. (2000). Lines, blobs, crosses and arrows: Diagrammatic communication with schematic figures. *Proceedings of the 1st international conference on diagrammatic representation and inference.* 221-230

Tversky, B. and Lee, P. U. (1998). How space structures language. In C. Freksa, C. Habel, and K. F. Wender (Editors), *Spatial cognition: An interdisciplinary approach to representation and processing of spatial knowledge*. 157-175.