A Cognitive Approach to Sketch Understanding

Ronald W. Ferguson

Intelligent Systems Group
College of Computing
Georgia Institute of Technology
801 Atlantic Avenue
Atlanta, GA 30332
rwf@cc.gatech.edu

Kenneth D. Forbus

Qualitative Reasoning Group
Computer Science Department
Northwestern University
1890 Maple Avenue
Evanston, IL 60201
forbus@northwestern.edu

Abstract

Sketching is an interactive process of communication, using drawings and linguistic information to convey spatial and conceptual material. Our work on a computational model of sketching has the goal of both explaining human sketching and creating software that can be a human-like partner in sketching interactions. This focus has lead us to explore a very different set of tradeoffs from those typically chosen in multimodal interfaces. We highlight some results of our approach, including research performed using GeoRep, our diagrammatic reasoning architecture, and sKEA, a multi-modal sketching tool used for knowledge acquisition in spatial domains.

Introduction

Sketching is an interactive process of communication, using drawings and linguistic information to convey spatial and conceptual material. Whether done on whiteboards or on the backs of envelopes, sketches are seldom precise. Although human visual processing is extremely powerful, people's artistic skills are highly variable, especially in interactive time. Consequently, sketches often are ambiguous, both in what the sketched entities represent and what spatial relationships are important. Such ambiguities are clarified in human sketching by a variety of means, including labels, speech, and deictic references. Through interaction, the process of sketching (more than just the sketch itself) can convey substantial conceptual material.

We are working on a computational model of sketching [13], with the goal of both explaining human sketching and creating software that can be a human-like partner in sketching interactions. We focus on rich visual and conceptual understanding of sketched material, using results from cognitive science on the nature of qualitative modeling, analogical reasoning [16], and visual processing to drive the development of our software.

This focus has led us to explore a very different set of tradeoffs from what is typically chosen in multimodal interfaces. In many systems, such as Quickset [3] and ASSIST [1], sketch understanding is equated with glyph recognition, where each glyph is classified using a predefined glyph set given for that domain. Such systems can be very useful for specific tasks in carefully defined domains, and studying the use of multiple modalities to overcome recognition limitations is an important problem (e.g. [18, 20, 24]). However, we believe that the glyph classification approach misses a key aspect of sketching: the deep, shared understanding of what is depicted that is built up between the participants. To capture sketching's power as a communication medium, it is crucial that sketching software utilize aspects of this shared understanding whenever possible.

Creating software capable of shared rich conceptual and visual understanding is a daunting goal. Ideally, the software's visual processing capabilities need to be as powerful and as flexible as those of its human partners. Similarly, its understanding of what is depicted needs to be consistent with human conceptual structures and mental models. Human-level sketch understanding remains a distant goal, but fortunately there appear to be intermediate points that are both immediately useful and represent progress towards that goal. In the rest of this paper we highlight some of our results using this approach.

Rich visual understanding

Our work on visual understanding is guided by results from human visual processing. In our work we have focused on the intermediate level of processing where a significant role is played by visual structure.

Visual structure is heavily utilized in visual recognition and inference, as shown in a number of psychological studies. Nonaccidental properties [25], such as segment contiguity and parallelism, provide structure that can be used to recognize objects in varying orientations. Curvature minima points along a figure boundary have also been shown to be effective in visual descriptions [17]. These and other known characteristics of visual structure can be consolidated into spatial representation frameworks that can be empirically tested [2, 21, 22].

Visual structure is also central in more complex visual reasoning tasks. Although the precise mechanisms by which visual structure is utilized in such tasks are still

unknown, systems in *qualitative spatial reasoning* [4, 12, 15] have shown the broad utility of structural spatial vocabularies (called *place vocabularies*) in a number of domains. So, while these systems are not designed as models of perception, they do show that visual structure can form an integral part of domain-based reasoners. The success of qualitative spatial reasoners also suggests that visual structure might need to be elaborated or specialized to a task. Qualitative spatial reasoning systems typically use domain-specific place vocabularies, and it has been conjectured that no single place vocabulary suffices for all qualitative spatial reasoning tasks [11].

In visual reasoning tasks, even unambiguous spatial relations can have different meanings depending on the domain. The connectivity of segments, for example, can indicate joined paths in a wiring diagram, but might merely indicate an overlay in a different context. In one sketch, the characteristics of an irregular boundary might be important (e.g., finding a peninsula along a coastline), while in another the boundary characteristics are unimportant (e.g., the boundaries of a thought balloon).

GeoRep [9] is a diagrammatic reasoner that can model how visual structure is utilized within particular domains (see Figure 1). The input is a vector graphics file, drawn using a system such as Xfig [23]. The output is a relational description of the diagram in a high-level place vocabulary for a specific domain.

GeoRep works by linking low-level structural relations with a high-level place vocabulary. For the input vector graphics file, GeoRep creates a set of low-level spatial relations for a diagram. Then higher-level characteristics are inferred via domain-specific rules. At the low level, GeoRep models visual structure. At the high level, GeoRep acts like a qualitative spatial reasoner.

GeoRep has been used in performance systems and also in cognitive modeling efforts. One focus of our research has been to model the role of qualitative visual structure in symmetry perception.

MAGI [6, 7] is a model of symmetry and regularity detection that has been used to explain a variety of phenomena in human symmetry perception. MAGI shows how visual structure may be used to detect qualitative symmetries and repetitions based on the regularity in systems of visual relations.

Our hypothesis of the central importance of qualitative representations in human visual understanding has lead to a prediction about human symmetry perception that has been experimentally verified. As predicted by the MAGI model, human subjects more quickly or accurately judged figures as asymmetric when these figures contained mismatched qualitative visual relations (such as a vertex on one side that was missing on the other) than when the figures contained otherwise equivalent quantitative asymmetries (such as differences between each side's total area) [8].

Multimodal interfaces

Sketch understanding often involves more than understanding spatial relationships in isolation, even within a particular domain context. In many cases, inferred spatial relationships must be combined with other information sources.

Sketches are often part of a multimodal interaction, where speech and deictic references are a critical component of what is conveyed. Glyphs in a sketch may be rapidly drawn and then classified via a few quick verbal comments rather than carefully drawn in detail. For such sketches, glyph classification is often extremely difficult—

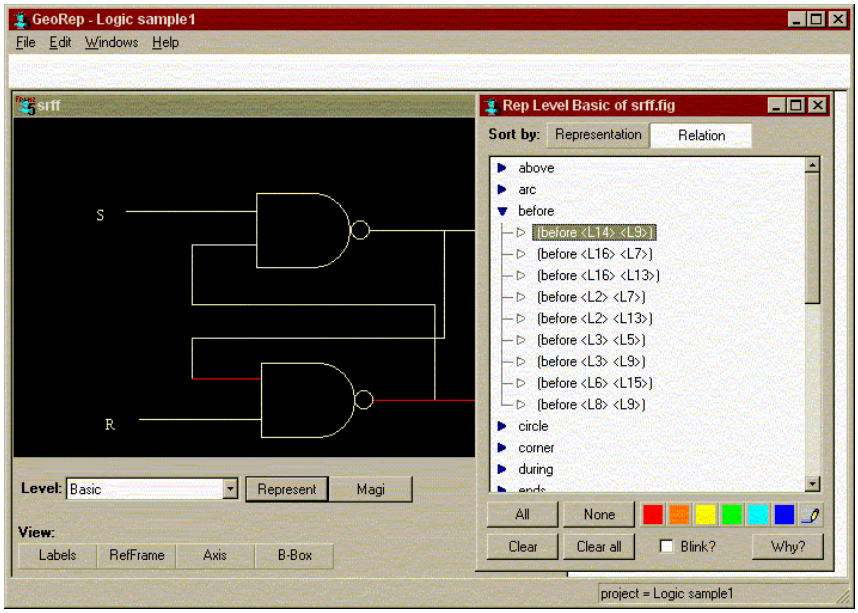

Figure 1: The GeoRep spatial reasoning system

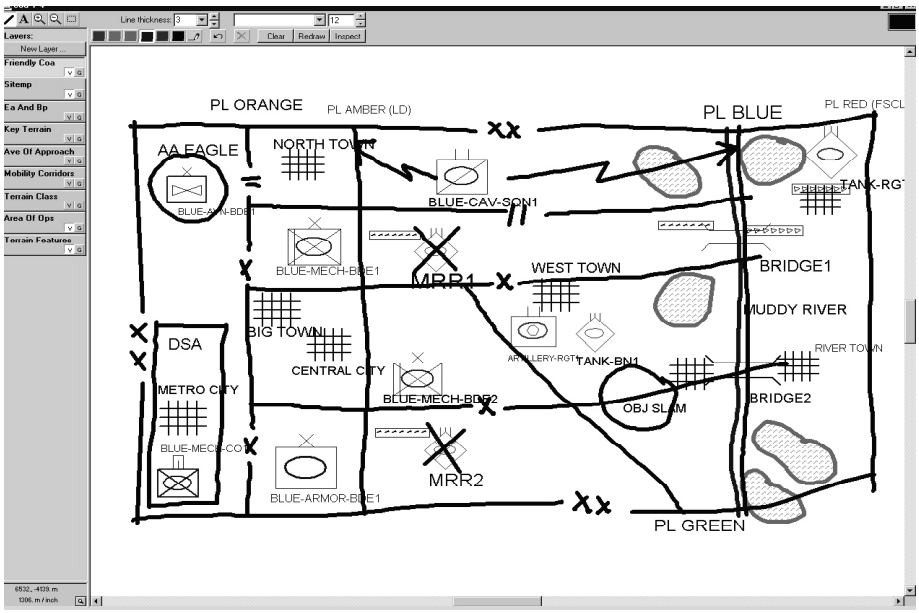

Figure 2: Course-of-Action diagram drawn with nuSketch.

for machines and humans alike—without the context provided by the accompanying multimodal elements.

In consideration of this, we developed the *nuSketch* framework for multimodal interface systems. In nuSketch-based systems, we avoid object recognition entirely. Instead other means, such as speech recognition or glyph "palettes", are used to indicate what is being drawn. When working with the glyph palette, prototype glyphs can be dragged and positioned by the user. When using speech recognition, a single glyph may be sketched only as a boundary and then identified via a few spoken words. In both cases, nuSketch stores the glyph categorization for later reasoning, and it may also redraw the glyph if necessary.

Thus the user provides the conceptual interpretation of the glyph ink explicitly. The ink, though not utilized to categorize the glyph, is used to compute the glyph's visual properties. It is also used to calculate spatial relationships between glyphs, which in turn can help determine spatial and conceptual relationships among the things the glyphs represent. In other words, while this input method omits some visual information, the ability to quickly categorize glyphs allows nuSketch to produce a rich spatially-oriented representation.

In our experience (and that of our users), nuSketch's mode of interaction is relatively natural and requires minimal training. When using the glyph palette, users have little difficulty understanding how to select and drag glyphs from the palette. For the mode that utilizes speech recognition, the task of remembering the command vocabulary and (for a few glyph types) specialized stroke sequences imposes an additional memory load on the user. Once learned, however, interaction with the system appears to be significantly easier and more natural than with vector-based drawing systems such as Xfig.

The nuSketch interface also has the ability to divide the sketch into *visual layers*. Each visual layer represents one level of representation, and may have its own grammar for ink and speech recognition. The visual salience of each layer can be controlled by highlighting, removing, or simply graying out individual layers.

We have created two prototypes using this framework. We summarize each in turn.

COA Creator/Tactical Decision Coach

Military course of action (COA) diagrams are battle plans, using a graphical language to indicate terrain, how the terrain should be thought of for communication purposes, and what tasks should be done by each unit. Figure 2 illustrates an example plan. COA diagrams utilize a relatively small but standardized set of hand-drawn symbols [5], and are utilized in planning at many different levels in the chain of command.

The COA Creator is based on the nuSketch framework. In addition to the multimodal grammar, which describes the available lexicon and glyph types, the COA Creator also uses an extensive knowledge base covering the COA domain. This knowledge base was created in the DARPA HPKB program, using the Cyc Knowledge Base [19] as a starting point. For example, to add a river, the user tells the system "add river" while drawing the river on the screen. The COA Creator then stores the resulting glyphs and links it to a representation of an instance of the knowledge base's *river* category.

The COA Creator's rich conceptual understanding of military tasks is complemented by geographic reasoning

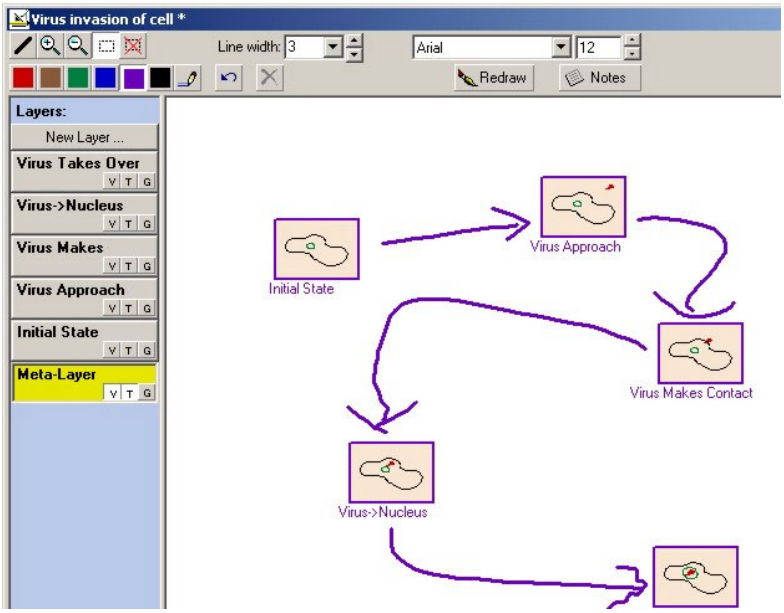

Figure 3: sKEA's metalayer enables relationships between subsketches to be expressed via sketching.

capabilities, which use GeoRep to reason about the spatial relationships indicated in the sketch [10]. Queries concerning spatial relationships can be combined with conceptual inferences generated through the knowledge base. These capabilities allow it to answer relatively sophisticated spatial queries about the sketch, and to illustrate its answers by direct reference to the sketch itself. For example, the COA Creator can answer whether there are any minefields on the paths between the friendly units and their current military objectives, and can highlight any such situations in the sketch.

In an experiment at the Battle Command Battle Lab in Ft. Leavenworth in November 2000, the combination of the COA Creator for input with the Fox/CADET system to generate detailed plans from its output was found to be more effective in generating COA diagrams than traditional methods.

Currently, the COA creator is being extended and embedded into a training system, the Tactical Decision Coach. The TDC is designed to provide feedback to students on tactical problems. Given a problem, they use the COA software to draw their solution. This solution is compared (using our cognitive model of analogical matching) to expert solutions to provide feedback.

The TDC prototype is still under development, but formative feedback from military personnel has been encouraging.

Sketching for knowledge capture

People often communicate new ideas via sketching, so sketching is a natural modality for creating knowledge bases. Knowledge capture is problematic using traditional multimodal techniques, since the system designers cannot know in detail what the experts will be telling it in advance, and hence cannot do the vocabulary and grammar engineering and extensive data gathering required to train today's statistical recognizers.

In our work on the *sketching Knowledge Entry Associate* (sKEA), we take a radically different approach, using clever interface design to solve the segmentation and identification problems.

sKEA [14] is based on several novel ideas. We use *blob semantics* for glyphs to provide a manageable visual/conceptual integration, and *qualitative spatial representations* to link visual and conceptual levels of understanding. The insight is that many sketches convey information through the relationships between glyphs, rather than the internal structure of the glyphs themselves.

Our one nod to glyph recognition is using arrows to express binary relationships between the things that glyphs represent. Even in this case, however, the recognition assumptions are minimal—sKEA assumes that arrows are drawn as two or three strokes, the longest one being the body of the arrow and the rest the head. This simple convention places no constraint on, for instance, the shape of the body or of the head, and is easy to learn without training. SKEA assumes that the relationship holds between the entities represented by the closest glyphs to the head and tail that satisfy the argument restrictions of the predicate. Arrows significantly increase sKEA's expressive power, since through reification arbitrary relationships can be expressed.

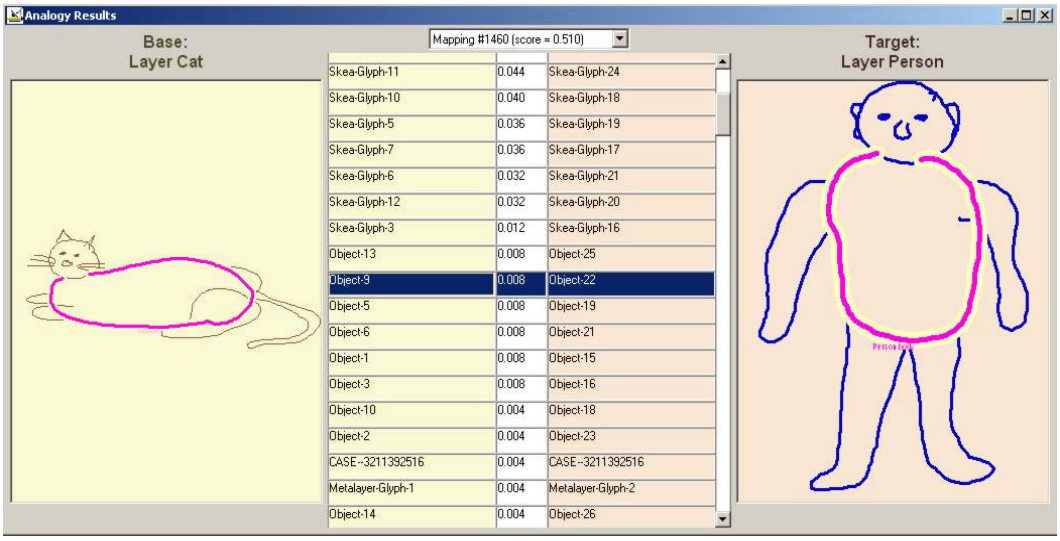

Figure 4: sKEA supports combined visual/conceptual analogies

Sketches often consist of *subsketches* representing different steps in a sequence or perspectives; we organize such subsketches into visual layers and enable relationships between subsketches to be indicated using arrows drawn on a *metalayer* (Figure 3).

Since sKEA's conceptual vocabulary is as broad as its knowledge base (currently a subset of Cyc), it can be used to express a wide variety of ideas. sKEA has been used to express content as diverse as biological sequences, structural descriptions, spatial layouts and geographical maps, and concept maps. sKEA also provides the ability to do analogies involving sketches (Figure 4).

Conclusion

In this paper, we have shown several different techniques that allow sketching systems to represent and reason about the contents of sketches and diagrams with little or no explicit glyph recognition. Some of this ability comes from using a multimodal sketching interface, which allows us to classify glyphs without a visual recognition algorithm, and does so in a way that is natural to the user. An equally important part of this ability, however, is the way that glyphs are interpreted in the context of much more extensive conceptual knowledge about the domain, which allows us to take relatively simple spatial characteristics—boundaries, location, and certain types of polarity—and impose deeper, richer interpretations both for individual glyphs and sets of glyphs.

Future goals include work on characterizing additional spatial properties of sketched glyphs. We are also looking at how regularity and other top-down influences might enable more extensive characterization of sketched glyphs.

Acknowledgements

This research is supported by the DARPA High Performance Knowledge Bases, Command Post of the Future, and Rapid Knowledge Formation programs.

References

1. Alvarado, C. and Davis, R. Resolving ambiguities to create a natural sketch based interface. in *Proceedings of IJCAI-2001*, 2001.
2. Biederman, I. Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94* (2). 115-147.
3. Cohen, P., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L. and Clow, J. QuickSet: Multimodal interaction for distributed applications. in *Proceedings of the Fifth Annual International Multimodal Conference*, Seattle, 1997, 31-40.
4. Cohn, A.G. Qualitative spatial representation and reasoning techniques. in Brewka, G., Habel, C. and Nebel, B. eds. *Proceedings of KI-97*, Springer-Verlag, Freiburg, Germany, 1997, 1-30.
5. Department of Defense. Department of Defense Interface Standard: Common Warfighting Symbology, United States Army, 1999.
6. Ferguson, R.W. MAGI: Analogy-based encoding using symmetry and regularity. in Ram, A. and Eiselt, K. eds. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum Associates, Atlanta, GA, 1994, 283-288.
7. Ferguson, R.W. Symmetry: An Analysis of Cognitive and Diagrammatic Characteristics *Department of Computer Science*, Northwestern University, Evanston, Illinois, 2001.
8. Ferguson, R.W., Aminoff, A. and Gentner, D. Modeling qualitative differences in symmetry

judgments. in *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1996.
9. Ferguson, R.W. and Forbus, K.D. GeoRep: A flexible tool for spatial representation of line drawings. in *Proceedings of the 18th National Conference on Artificial Intelligence*, AAAI Press, Austin, Texas, 2000.
10. Ferguson, R.W., Rasch, R.A.J., Turmel, W. and Forbus, K.D. Qualitative spatial interpretation of Course-of-Action diagrams. in *Proceedings of the 14th International Workshop on Qualitative Reasoning*, Morelia, Mexico, 2000.
11. Forbus, K.D. Qualitative reasoning about space and motion. in Gentner, D. and Stevens, A. eds. *Mental Models*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1983, 53-73.
12. Forbus, K.D. Qualitative spatial reasoning: Framework and frontiers. in Glasgow, J., Narayanan, N.H. and Chandrasekaran, B. eds. *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, The AAAI Press/The MIT Press, Menlo Park, CA, 1995, 183-202.
13. Forbus, K.D., Ferguson, R.W. and Usher, J.M. Towards a computational model of sketching. in *Proceedings of the International Conference on Intelligent User Interfaces*, Sante Fe, New Mexico, 2000.
14. Forbus, K.D. and Usher, J. Sketching for knowledge capture: A progress report *IUI-2002*, 2002.
15. Freksa, C. and Rohrig, R. Dimensions of qualitative spatial reasoning. in Carrete, N.P. and Singh, M.G. eds. *Proc of the IMACS Workshop on Qualitative Reasoning and Decision Technologies*, CIMNE, Barcelona, 1993, 483-492.
16. Gentner, D. and Markman, A.B. Structure mapping in analogy and similarity. *American Psychologist*, *52*. 45-56.
17. Hoffman, D.D. and Richards, W.A. Parts of recognition. *Cognition*, *18* (1-3). 65-96.
18. Landay, J.A. and Myers, B.A. Interactive Sketching for the Early Stages of User Interface Design. in *Proceedings of CHI '95: Human Factors in Computing Systems*, Denver, CO, 1995, 43-50.
19. Lenat, D.B. and Guha, R.V. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley, Reading, MA, 1990.
20. Mankoff, J., Abowd, G.D. and Hudson, S.E. OOPS: A toolkit supporting mediation techniques for resolving ambiguity in recognition-based interfaces. *Computers and Graphics*, *24* (6). 819-834.
21. Palmer, S.E. Hierarchical structure in perceptual representation. *Cognitive Psychology*, *9*. 441-474.
22. Palmer, S.E. Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. in Norman, D.A. and Rumelhart, D.E. eds. *Explorations in Cognition*, W. H. Freeman and Company, San Francisco, 1975, 279-307.
23. Smith, B.V. XFig, 2002. http://www.xfig.org
24. Stahovich, T.F., Davis, R. and Shrobe, H. Generating multiple new designs from a sketch. *Artificial Intelligence*, *104* (1-2). 211-264.
25. Lowe, D. *Perceptual Organization and Visual Recognition*. Kluwer Academic Press, Boston, 1985.