Perceptually Based Learning of Shape Descriptions for Sketch Recognition

Olya Veselova and Randall Davis

Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052
MIT CSAIL, 32 Vassar St., Cambridge, MA, 02139
olyav@microsoft.com, davis@ai.mit.edu

Abstract

We are interested in enabling a generic sketch recognition system that would allow more natural interaction with design tools in various domains, such as mechanical engineering, military planning, logic design, etc. We would like to teach the system the symbols for a particular domain by simply drawing an example of each one – as easy as it is to teach a person. Studies in cognitive science suggest that, when shown a symbol, people attend preferentially to certain geometric features. Relying on such biases, we built a system capable of learning descriptions of hand-drawn symbols from a single example. The generalization power is derived from a qualitative vocabulary reflecting human perceptual categories and a focus on perceptually relevant global properties of the symbol. Our user study shows that the system agrees with the subjects' majority classification about as often as any individual subject did.

1 Introduction

We are interested in creating a generic sketch understanding mechanism that will allow more natural interaction with design tools in a variety of domains (e.g., mechanical engineering, military planning, logic design, etc.). This paper presents a learning system that is part of this larger effort. Our goal is to make teaching the system new shapes as natural as possible, ideally as easy as it is to teach new symbols to another person. Usually it is enough for people to see a symbol once to make a reasonable decision whether new drawings are instances of it, even without knowing its meaning. For example, would you recognize Figure 1b as an instance of the symbol in 1a?

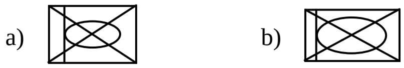


Figure 1. a) Original symbol, b) New drawing.

For most people the answer is "yes" because 1b contains the same features they paid attention to in 1a. And people pay unequal attention to different features, so their recognition is undeterred by the differences between 1a and 1b. This is the goal of any learning system. Common approaches to learning (e.g., neural nets, SVMs) require training on numerous examples of a symbol seeking to "average out" the unimportant differences. We sought a representation and a generalization mechanism that would reduce the number of required training examples, preferably, to as few as one. The difficulty is knowing which features people perceive as relevant in that one example. Consider the symbol in Figure 2a:

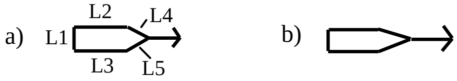


Figure 2. Symbol from military planning.

In 2a lines L1, L4 and L5 are the same length. But people mostly notice only the equality of L4 and L5. Hence, they can accept 2b as the same symbol even though it violates the constraint L1=L4. The system's goal is to do the same.

We turned to studies of human perception to understand what features people find relevant in a geometrical configuration, and created heuristics attempting to capture these perceptual biases. We used results from Goldmeier's perceptual similarity studies (Goldmeier 1972), Arnheim's book on art and visual perception (Arnheim 1974), and grouping principles described by the gestalt psychologists (Wertheimer 1923). Using these heuristics we built a system capable of learning descriptions of hand-drawn symbols from a single example. The description is phrased in terms of geometric primitives (lines and ovals) and constraints between them (connects, parallel, above, vertical, longer, etc.) The generalization power is derived from a qualitative vocabulary reflecting human perceptual categories and a focus on perceptually relevant global properties of the symbol – tension lines (defined below), obstruction, and grouping. Note that we focus on generalizing over variations that are not due to low-level noise and that would occur even with perfect lines, rectangles, ovals, and snap-to-grid (as in Figures 1 and 2). This alone presents a substantial challenge[1].

The recognition engine that will use our generated descriptions is still under development in our group, so we could not evaluate the final recognition accuracy. Instead, we measured how often our system agreed with people's perceptual judgments on near-perfect drawings. We asked several people to decide whether a given variation should be recognized as an instance of a new unfamiliar symbol and checked whether the description produced by the system would cause the same classification. On the whole

[1] The symbolic descriptions produced by our system will be used in a larger recognition engine with its own mechanism for dealing with low-level noise; see (Alvarado and Davis 2002).

data set, the system agreed with the majority vote 77% of the time. In comparison, a subject chosen randomly would agree with the majority 82% of the time. For cases with strong majority (>80% of the people voted the same) the system achieved 83% agreement (and a random person would get 91%). Section 4 describes the study in detail.

Because of several limitations (section 5) our system can't yet describe the full range of symbols in the domains of interest. We believe our preliminary results still show that using a representation and feature ranking built on knowledge about human perception adds considerable generalization power, allowing learning from one example.

2 Related Work

Of the large body of work supporting free-hand sketching, most relevant to this paper are systems attempting to recognize (i.e. give a categorical label to) the input. We are interested in two aspects of these systems: the representation and the learning mechanism. The choice of representation affects the descriptive power and the initial level of generalization. For learning, if the recognizers are trained (rather than created by hand), it is interesting to see what generalization techniques allow some systems to use fewer examples than others.

Some systems, like GRANDMA (Rubine 1991) or CALI (Fonseca, Pimentel, and Jorge 2002) use a set of global aggregate features (sum of angles, or properties of the bounding box, convex hull, etc.) to represent single-stroke gestures or simple shapes, and require over 50 training examples. Using only global features, like the ones in these systems, is not sufficient for capturing the detail of more complex symbols that we are interested in (e.g., Figure 1). Instead we needed to make explicit the properties of the individual parts of the symbol.

In several systems, the representation combines primitive shapes and higher level qualitative spatial relationships between them, similar to the ones in our system (Landay and Meyers 2001), (Gross and Do 2000), (Shilman et al. 2002). However, even if the low level shape recognizers can be trained, there is no mechanism for automatically learning which of the higher level relationships are important. They have to be picked out by the user from a fairly large list recorded by the system. In contrast, we focus on identifying the relevant subset automatically.

GeoRep (Feguson and Forbus 1999) produces descriptions of perfect line drawings in terms of geometric primitives and qualitative spatial relations between them. To generalize, it records only constraints between proximal elements, assuming only those constraints are visually important. Similarly, our system prefers local interactions. However, we found that interaction may be relevant even if two primitives are far away, as long as they are not separated visually by other primitives. We also take into account global alignments and grouping. We show that these are important to better capture perceptual relevancy.

(Calhoun et al. 2002) presents a system that is most similar to ours. It uses a semantic network to represent primitives (lines and arcs) and constraints between them (parallelism; angle between primitives, and intersections with their relative location). The system needs only a few examples to train each recognizer. The training filters out the relationships and properties appearing with low frequency. Different weights are assigned to different kinds of errors for recognition matching, reflecting different perceptual importance. These weights play the same role as the default relevance scores in our system. Yet our system has mechanisms to further adjust these scores, because we observed that the same type of constraint may have different perceptual importance depending on the global configuration of primitives.

In summary, similar to ours, several previous systems have used qualitative constraints for initial generalization. Our contribution is in ranking them and guiding further generalization by including heuristics about perception of the overall configuration of primitives in a symbol.

3 Approach

Our goal is to produce a symbolic description of a drawn symbol adequate for recognition, i.e. capturing just the relevant features. The input to the system is the user's strokes, segmented into simple geometric primitives (lines or ovals) (Sezgin, Stahovich, and Davis 2001). The output description includes these primitives and a set of geometric constraints between them. This section shows how the vocabulary of these constraints was determined and how the system decides which of all the constraints to keep. We also illustrate its operation on the symbol below:

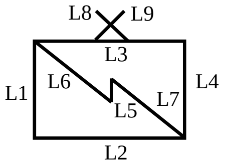


Figure 3. Symbol from military planning.

For this illustrative example, the system has previously learned the descriptions of the cross (L8, L9) and the rectangle (L1-L4), so it can identify them in the drawing.

3. 1 Qualitative Vocabulary

We have gained many insights from Goldmeier's work on perceived similarity of shapes (Goldmeier 1972), which shows how people are strongly biased to notice certain geometric properties and ignore others. Goldmeier identifies the relevant and irrelevant features by exploring their effect on perceived similarity. Figure 4 illustrates a typical experiment. Consider the shape in Figure 4a and ask yourself which of 4b and 4c is more similar to 4a.

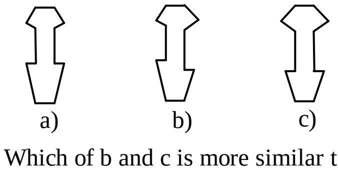


Figure 4. Which of b and c is more similar to a?

The majority of subjects chose c. Note that the left side of b is exactly the same as a. Even though in c all the lengths and angles are slightly changed, it is considered more similar because of preserved symmetry.

Goldmeier found that people attend to properties he calls *singularities*, i.e. special cases in the space of geometric configurations in the sense that small variations in them make a qualitative difference. One example is verticality: a vertical line rotated a little is no longer vertical. Symmetry is another example[1]. Goldmeier also showed that people are much less sensitive to variations in nonsingular properties, perceiving them as essentially the same state (Goldmeier 1982, p. 44). Thus we can reduce the description vocabulary to a few qualitative states that represent singularities and lump non-singular values together. For example, it is enough to describe a line as horizontal, vertical, or positively or negatively sloped.

Goldmeier's work mentions some singularities explicitly, like symmetry, parallelism, horizontality, verticality, and straightness. We have picked the rest of the vocabulary terms using our own introspection and relying on Goldmeier's definition of singularities as the "more regular, better, more unique" (Goldmeier 1982, p. 44), and as properties a change in which significantly alters the perception of the symbol. The table on the right lists the constraints we use.

In an input symbol we find all constraints expressible in this vocabulary, allowing for a small level of noise (like almost vertical or almost connected). For Figure 3, for example, the system identifies 92 constraints (in addition to those in the previously learned cross and rectangle). Sample constraints include: connects (L5 L6), parallel (L6 L8), vertical (L5), and above-centered (L3 L5).

3.2 Default Relevance Ranking

A second set of Goldmeier's experiments demonstrates that singular properties can have different perceptual importance, as illustrated below. In both figures, subjects were again asked which of b and c is more similar to a.

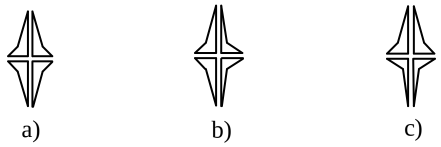


Figure 5. Which of b and c is more similar to a?

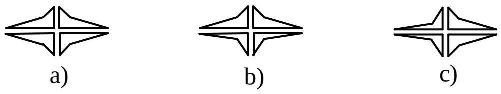


Figure 6. Which of b and c is more similar to a?

In Figure 5, the majority of subjects chose c, while in Figure 6 the choice was b, even though the shapes in Figure 6 are simply rotated versions of Figure 5. In both cases, the viewers preferred the vertical axis of symmetry.

Goldmeier presents several similar experiments. However, they are not sufficient, for ranking all the constraints in our system. As a result, we had to use our own introspective analysis to find the relative perceptual importance of different constraints. We examined common symbols in several domains (military planning, mechanical engineering, etc.) and determined which symbol properties allowed the most variation without a large perceptual change to the symbol. The table below shows supported constraints in order of decreasing relevance, with singular constraints shown in bold.

Constraints	Score
Connects (for line endpoints)	1.0
Meets (i.e. T-intersection), **intersects**, **tangent**. Inside, **inside-centered**	0.95
Touches, **overlaps (for ovals)**	0.9
Horizontal, vertical (for lines)	0.8
Pos- and neg-slope; above, below, right, left, upper-right, upper-left, lower-right, lower-left; **above-, and right-centered**; **parallel, perpendicular**	0.7
Horizontal, vertical (for ovals); elongated, **non-elongated; same-length, same-size**	0.6
longer, larger	0.55

The scores reflect the default relative relevance, and were chosen to be spread in the upper half of the [0;1] interval. They are adjusted by the mechanisms described below.

3.3 Adjusting Relevance Scores

Goldmeier argues that the saliency of a given property depends on the overall configuration of the primitives in a shape. Observations we made in the course of this work helped us create heuristics for adjusting default relevance scores based on three global properties: obstruction, tension lines, and grouping. We describe these heuristics in the order that they are applied by the system.

Obstruction. In symbols with many parts attention seems to be focused on local interactions. Consider Figure 7a.

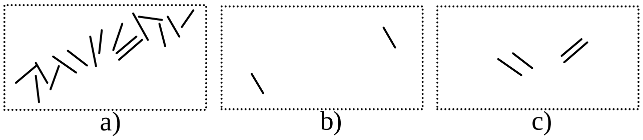


Figure 7. b and c are parts of the pattern in a.

Figures 7b and 7c are parts of the pattern in 7a, yet the obvious parallelism of the lines in b is not noticeable when looking at a. The lines isolated in c, however, are more obviously parallel even in the original context in a. It is easier to pay attention to the local interaction of the lines in 2b because there are no other lines separating them. We try to approximate this effect by the notion of obstruction, which is measured by the number of geometric primitives between a given pair: $O(p_1, p_2)$. The system decreases the relevance of relative orientation, length, position, and size constraints by $0.15*O(p_1, p_2)*r$, where r is the current

[1] We do not yet handle symmetry (or any *n*-ary constraints), but we use the example because it illustrates the larger point so simply and convincingly.

relevance score[1]. For Figure 3, for example, this heuristic causes a large decrease of the relevance for the constraints like “parallel (L8 L6)” and “longer (L9 L5),” ensuring the system would accept the variation in Figure 12a.

Tension Lines. Arnheim argues in his work that people attend to regular alignments of geometric primitives, particularly horizontal and vertical alignments (Arnheim 1974). In Figure 8a the circle is perceived to be “out of balance,” while placing it on one of the dashed lines in 8b would create a more “stable” configuration:

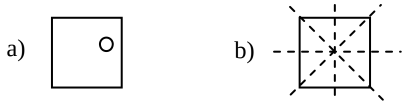


Figure 8. Regular alignments.

The alignments of corners of the square and the centers of its sides form a kind of perceptual grid that other elements are “pulled” toward. In our system, we call these alignments *tension lines*, defined in terms of alignments of line endpoints and midpoints. The system identifies a tension line wherever at least two such line points align horizontally or vertically (we do not yet support diagonal alignments). The system increases by 0.5*(1 – *r*) the current relevance *r* of relative length, position, and orientation constraints that would break the tension lines if violated. For example, the relevance of the relative vertical positions of the lines L8, L9, L3 and L5 in Figure 3 is increased because their midpoints lie on a vertical tension line. This ensures that the example in Figure 12c would be rejected.
Note that this heuristic boosts some constraints that make parts of the symbol horizontally or vertically symmetrical, thus implicitly helping enforce some symmetry constraints.

Grouping. Finally, we also use observations of perceptual bias from the Gestalt psychologists, who noted that people tend to combine individual elements into a greater whole, grouping them by proximity, similarity, etc. (Wertheimer 1923). For example, Figure 9a is perceived as two rows of circles, rather than six individual circles. Properties of a row as a whole are perceptually more important than properties of its components. People do not tend to notice the vertical alignment of the circles in from each of the rows the way we do in Figure 9b.

Figure 9. Perceptual grouping.

The current implementation supports only two grouping principles: connectedness, determined by segmenting the symbol into connected components, and familiarity of shape, determined by looking for previously learned symbols as subparts of the new drawing. Figure 10 gives the group hierarchy for the symbol in Figure 3.

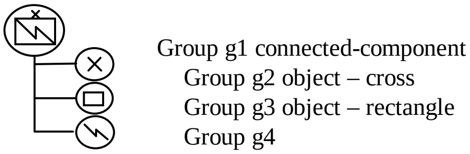
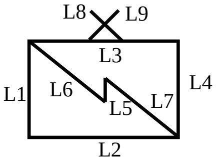


Figure 10. Group hierarchy of the symbol.

The system decreases the relevance *r* of relative orientation, length, position, and size constraints between primitives that belong to different groups by subtracting 0.4**r*. For example, it decreased the relevance of “same-length: (L6 L1) (L7 L4),” hence the variation in Figure 12b would be accepted according to the description.
After all the scores have been adjusted, only constraints ranking above 0.5 are included in the final description. For the example symbol, from 92 constraints initially found by the system only 34 remain in the description. Note that the description in Figure 11 refers to the previously learned descriptions for the cross and the rectangle, which are shown in the box.

GROUP HIERARCHY:
Group g1 connected-component: L1-L9
Group g2 object - cross: L9 L8
Group g3 object - rectangle: L1-L4
Group g4: L5 L6 L7



CONSTRAINTS
vertical: (L5)
neg-slope: (L7) (L6)
connects: (L5 L6) (L5 L7) (L4 L7) (L3 L6) (L7 L2) (L6 L1)
meets: (L9 L3) (L8 L3)
above-centered: (L5 L2) (L3 L5) (L9 L2) (L9 L3) (L9 L5) (L8 L2) (L8 L3) (L8 L5)
right-centered: (L5 L1)(L4 L5)
right: (L5 L6) (L7 L5)
upper-right: (L4 L7) (L3 L6) (L7 L2) (L6 L1)
parallel: (L7 L6)
longer: (L3 L6) (L3 L8) (L3 L9) (L2 L7) (L7 L5) (L6 L5)

CROSS:
intersects: (L8 L9)
pos-slope: (L9)
neg-slope: (L8)
same-length: (L8 L9)

RECTANGLE:
horizontal: (L3) (L2)
vertical: (L1) (L4)
above-centered: (L3 L2)
right-centered: (L4 L1)
upper-left: (L1 L2) (L3 L4)
upper-right: (L3 L1)(L4 L2)
same-length:(L3 L2) (L4 L1)
longer: (L2 L4) (L2 L1) (L3 L4) (L3 L1)

Figure 11. Final description for the symbol in Figure 3.

Figures 12 shows the variations that would be accepted and rejected respectively, according to this description:

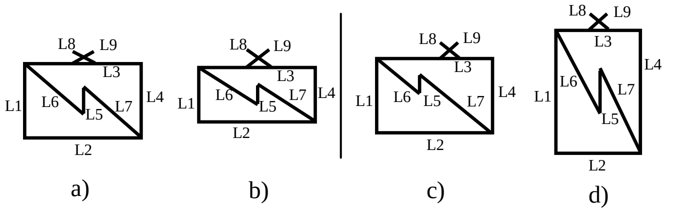


Figure 12. a), b) Variations that fit the system’s description. c), d) Variations that contradict the description.

[1] The constants used in this and other formulas were determined empirically by examining resulting descriptions for symbols in various domains.

4 User Study and Evaluation

The ideal evaluation of the system would be to use the produced descriptions in a sketch recognition engine and test the recognition accuracy. However, the recognition engine in our group is still under development. As an alternative, we tested whether the system's descriptions would produce the same classifications of shapes as people would.

In our study 33 subjects were shown 9 unfamiliar symbols and 20 variations of each symbol. The variations were constructed using the system's descriptions produced for the 9 symbols by randomly picking a property from the description and varying it randomly to a large or a small degree. Half of the 20 variations were chosen to agree with the description and half contradicted it. This means that the description would cause a recognition engine to recognize half of these variations as the learned symbol. The subjects had to choose whether each variation should be recognized as the original symbol, even though on many examples the choice was somewhat difficult (e.g. like it would be for variations in Figure 12 on the original symbol in Figure 3). We forced a binary choice since ultimately this is the task that the system will have to face. Before voting on each variation, the subjects could look at the original symbol as long as they liked.

Given that our system relies only on geometric information, to ensure parity, we selected symbols from military planning, a domain most likely unfamiliar to the subjects and in which the symbols have little resemblance to the physical objects they stand for, preventing test subjects from relying on knowledge about the things represented (Figure 13).

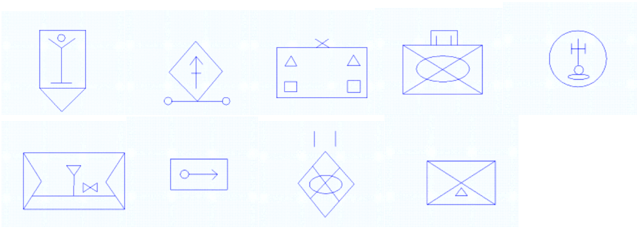


Figure 13. Symbols used for evaluation.

For each variation, we recorded the majority answer and the percentage of people who gave that answer (majority percentage). For almost 40% of the variations the subjects had high agreement – the majority percentage was above 90%. On more than half of data set the majority percentage was higher than 80%. For cases with highly divided opinions, it makes less sense to evaluate the performance of the system (i.e., level of agreement with people) since people did not agree with each other. Hence, we report the results for both the complete data set and for the variations with high agreement (i.e., with majority >= 80% and majority >= 90%).

Chart 1 shows the evaluation results. We measured the proportion of times that the system agreed with the majority answer, reported for different data subsets. For example, for the subset of the variations where the majority percentage exceeded 80%, the system agreed with the majority vote 83% of the time. Note that the baseline performance is 50%: the system would agree with people half of the time if it guessed randomly.

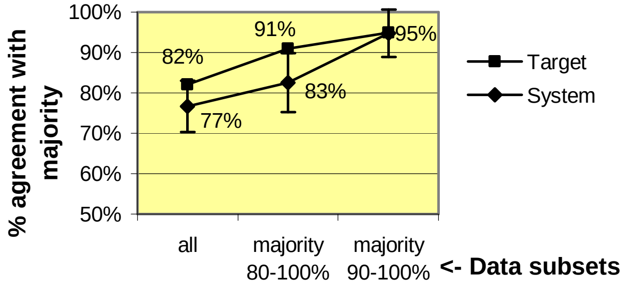
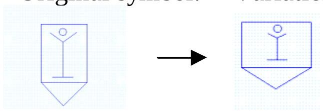


Chart 1. Evaluation results.

The "Target" marker shows the level of agreement that a person randomly selected from the subjects would achieve. Since we have focused on replicating people's perceptual biases, we should not expect the system to do better than this. The results show that system captured enough information about the symbol to perform significantly above chance level and to approach close to the target.

It is interesting to look at the disagreements. Most of the system's errors were false negatives, i.e. the system rejected a variation that most people would still recognize as the symbol. The issue was that people seemed to pay less attention to individual detail (aspect ratio, precise position, etc.) of the composing shapes in the symbol than the system biases accounted for.

Original symbol: Variation:

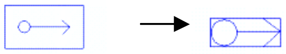


Majority vote: yes 89%
System: no

Figure 14. Example of a false negative error.

Most false positives stemmed from the lack of global symmetry detection and a lack of apparently perceptually relevant "must not" constraints for properties like touching, connection, or intersection.

Original symbol: Variation:



Majority vote: no 72%
System: yes

Figure 15. Example of a false positive error.

5 Limitations and Future Work

There are several limitations in the current system. The system currently supports only symbols composed of lines and ovals, and should eventually incorporate arcs and curves. This will require exploring what properties are perceptually salient for arcs and curves.

Our qualitative vocabulary lumps non-singular property values into one term and does not capture extreme degrees. For both a and b in Figure 16, for example, the system would describe the relative size of the circles as "larger o1 o2," making the descriptions identical. But these symbols appear quite different to human observers.

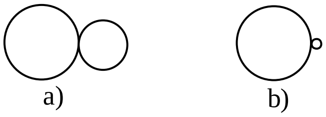


Figure 16. Perceptually different symbols.

As noted, the system doesn't handle "must not" constraints. Perhaps a possible solution is to treat the absence of the most perceptually relevant constraints (connects, touches, meets, etc.) in the description as required must-not's.

The system uses only pairwise constraints. We need to include support for constraints that involve several primitives, like symmetry, interval equality, or alignment of multiple elements. Without them it is impossible to represent configurations like the one in Figure 17.

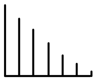


Figure 17. Symbol requiring alignment and interval equality.

The system records a fixed set of geometric primitives, so it is bound to overconstrain symbols that can have an arbitrary number of some elements:

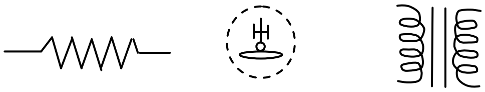


Figure 18. Symbols with an arbitrary number of primitives.

Learning such configurations presents two challenges. The system has to be able to, first, identify a group of repeated components and, second, decide whether an arbitrary number of them was intended. Goldmeier's studies provide some hints on how this may be done. He has shown that when the repeated elements are small compared to the size of the symbol and there is a large number of them, people start perceiving them as material rather than form and become insensitive to the variation in number of such components. The difficult task is defining quantitatively the terms "small relative to the symbol size" and "large number of elements".

While these limitations show that the system is incomplete, the general approach of relying on perceptual biases seems sound, as it suggests potential solutions to some of the limitations.

6 Contributions

We have presented a system for learning shape descriptions. It guides the generalization from a single example using built-in knowledge of observations about human perception. The main sources of the generalization power are a qualitative vocabulary reflecting perceptual singularities, different relevance scores for different constraints, and score adjustment heuristics based on perceptually relevant global properties of the symbol.

Our initial implementation shows that it is possible to produce descriptions of structurally complex symbols from a single example. Our user study has shown that the system agreed with the majority perceptual judgment almost as well as a randomly chosen person would.

Future work on the system includes improving its descriptive ability by adding support for curves and parts with an arbitrary number of elements and by extending the vocabulary to support higher-level constraints like symmetry, interval equality, and multiple alignments. We believe that further exploration of perceptual biases may provide clues on how to achieve these extensions.

7 Acknowledgements

We thank sponsors of MIT Project Oxygen and Microsoft iCampus Project for funding of this research.

References

Alvarado, C. and Davis, R. 2002. A Framework for Multi-Domain Sketch Recognition. Proc. of AAAI 2002 Spring Symposium: Sketch Understanding Workshop.

Arnheim, R. 1974. *Art and Visual Perception.* University of California Press.

Calhoun, T. F, Stahovich, T., Kurtoglu, L., and Kara, B., 2002. Recognizing multi-stroke symbols. Proc. of AAAI 2002 Spring Symposium: Sketch Understanding Workshop.

Ferguson, R. W. and Forbus, K. D., 1999. GeoRep: A Flexible Tool for Spatial Representation of Line Drawings. Qualitative Reasoning Workshop.

Fonseca, M. J., Pimentel, C., and Jorge, J. A. 2002. CALI: An Online Recognizer for Calligraphic Interfaces. Proc. of AAAI 2002 Spring Symposium: Sketch Understanding Workshop.

Goldmeier, E., 1972. Similarity in Visually Perceived Forms. *Psychological Issues*, Vol. 8, No. 1.

Goldmeier, E, 1982. *The Memory Trace: its formation and its fate.*

Gross, M. D. and Do, E. Y.-L., 2000. Drawing on the back of an Envelope: a framework for interacting with application programs by freehand drawings. Computers & Graphics 24(6): 835-849.

Landay, J. A. and Meyers, B. A., 2001. Sketching Interfaces: Toward More Human Interface Design. IEEE Computer. 34(3): pp. 56-64.

Rubine, D, 1991. Specifying Gestures by Example. Computer Graphics, Vol. 25, No. 4.

Sezgin, M., Stahovich, T., and Davis, R. 2001. Sketch Based Interfaces: Early Processing for Sketch Understanding. Proceedings of PUI.

Shliman, M., Pasula, H. Russel, S., and Newton, R. 2002. Statistical Visual Language Models for Ink Parsing. Proc. of AAAI 2002 Spring Symposium: Sketch Understanding Workshop.

Wertheimer, M., 1923. Translation in W. D. Ellis (ed.) *A Source Book of Gestalt Psychology*. New York: H. B. J., 1938.