Crowd-Sourcing Design: Sketch Minimization using Crowds for Feedback

David Engel,[a] Verena A. Kottler,[b] Christoph Malisi,[b] Marc Roettig,[c]
Eva-Maria Willing,[d] Sebastian J. Schultheiss,[e]

MIT Center for Collective Intelligence

MIT Center for Collective Intelligence Working Paper No. 2012-01

Spring 2012

MIT Center for Collective Intelligence
Massachusetts Institute of Technology
http://cci.mit.edu

[a] MIT – Sloan School of Management, Cambridge, MA, dengel@mit.edu
[b] Max Planck Institute for Developmental Biology, Tuebingen, Germany
[c] University of Tuebingen Center for Bioinformatics, Tuebingen, Germany
[d] Max Planck Institute for Plant Breeding Research, Cologne, Germany
[e] computomics.com, Molecular Data Analysis, Tuebingen, Germany

Crowd-Sourcing Design: Sketch Minimization using Crowds for Feedback

David Engel
Massachusetts Institute of Technology
Center for Collective Intelligence
Cambridge, MA, USA

Verena A. Kottler
Max Planck Institute
for Developmental Biology
Tuebingen, Germany

Christoph Malisi
Max Planck Institute
for Developmental Biology
Tuebingen, Germany

Marc Roettig
University of Tuebingen
Center for Bioinformatics
Tuebingen, Germany

Eva-Maria Willing
Max Planck Institute
for Plant Breeding Research
Cologne, Germany

Sebastian J. Schultheiss
computomics.com
Molecular Data Analysis
Tuebingen, Germany

Abstract

Design tasks are notoriously difficult, because success is defined by the perception of the target audience, whose feedback is usually not available during design stages. Commonly, design is performed by professionals who have specific domain knowledge (i.e., an intuitive understanding of the implicit requirements of the task) and do not need the feedback of the perception of the viewers during the process. In this paper, we present a novel design methodology for creating minimal sketches of objects that uses an iterative optimization scheme. We define minimality for a sketch via the minimal number of straight line segments required for correct recognition by 75% of naïve viewers. Crowd-sourcing techniques allow us to directly include the perception of the audience in the design process. By joining designers and crowds, we are able to create a human computation system that can efficiently optimize sketches without requiring high levels of domain knowledge (i.e., design skills) from any worker.

Introduction

Graphic and visual design processes are usually handled by highly trained professional individuals or groups that posses an understanding of how the product will be perceived by the audience. Here, we present a novel approach of tackling such design tasks using a human computation pipeline that integrates non-professional individuals and crowd-sourced feedback into a system that can optimize visual designs. We worked on minimizing sketches of object categories but we are confident that the results and the pipeline are applicable to other design tasks as well.

Images convey a wealth of information that can be grasped effortlessly. Sketches, symbols, and icons represent real world objects by simpler pictorial structures and can reduce the objects to a point where they have no visual similarity to the real objects, but still convey the same information as much more complex representations.

A crucial criterion for the design of pictorial structures such as sketches is the ease of recognition by untrained viewers, allowing the relevant information to be easily conveyed. Another important design criterion in the modern world is machine readability. Decades of computer vision research have shown that automatic object recognition of real world images is notoriously difficult. However, sketch-, icon-, and symbol-recognition are much easier to solve by machines, due to the higher amount of structure these entities provide. Automatic recognition of UPC codes (Global Standards One 2010) or of Kanji (Toyoda 2009), which are contemporary instances of symbols representing objects, are good examples.

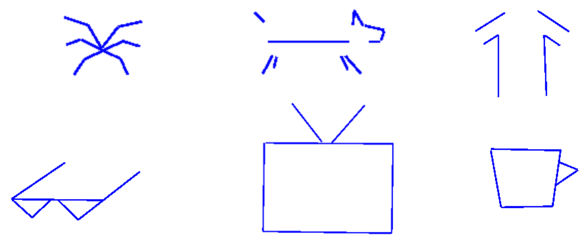

Figure 1: Minimal sketches of six categories. At least 75% of the readers should be able to identify them correctly.

These considerations led us to investigate the concept of minimal sketches, i.e., a sketch that is reduced so far that removing anything from it would lead to severely reduced recognition rates by untrained viewers. Such sketches have the desirable property that they do not contain visual clutter and are therefore effortlessly recognized. Furthermore, it is reasonable to believe that machine vision systems can parse them under a wide variety of viewing conditions. Finding such minimal sketches is a hard task, since it depends critically on the perception of naïve viewers.

The contribution of this paper is twofold: we present the paradigm of minimal sketches and also outline a pipeline to create such sketches that includes the perception of a crowd as an integral part of the design process. In addition, our pipeline, which aims to minimize the requirement for special qualifications for the workers, could be applied to other design tasks.

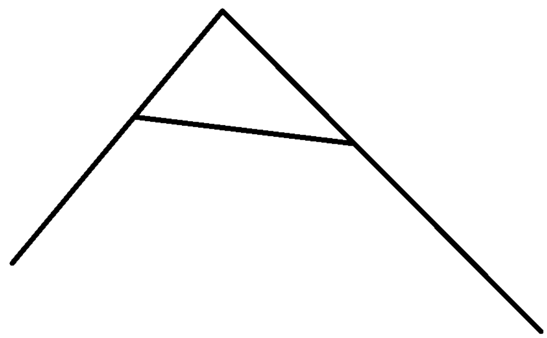

Figure 2: A sketch of a mountain that could be recognized in a multiple-choice question but not in a free-naming task.

Related Work

Considerable work has been done on sketches. The earliest study relevant to this work has been done by (Snodgrass and Vanderwart 1980), who present a set of 260 sketches of common objects and a psychophysical evaluation of their recognizability. These sketches were, however, only optimized for high recognition rates and are far from minimal as defined here. Icon design is a related field and has attracted a lot of attention over the years (Calpin 2001). However, icons differ from sketches in being only symbolic representations rather than a depiction of the object and their design has not been the object of much systematic research, (e.g. (Chen 2003)). Much more research has been devoted to the automatic recognition of sketches using techniques from computer vision and machine learning, cf. e.g.(Mahoney and Fromherz 2002); (Field et al. 2011); (Paulson and Hammond 2008). These approaches show that automatic sketch recognition systems are able to identify sketches correctly even in suboptimal conditions.

Over the last years, the use of crowd-sourcing and micro-task markets such as Amazon Mechanical Turk to harness collective intelligence for complex tasks has seen a surge in popularity. The idea of taking complex tasks and splitting them up in a large number of smaller tasks has led to exciting work in, e.g., computer vision (Wah et al. 2011), graphical perception (Heer and Bostock 2010), and biomedical research (Khatib et al. 2011). Most relevant to this paper is the work done in crowd creativity. (Yu and Nickerson 2011) had crowds combine, refine and evaluate designs of chairs. Beyond that, (Yu and Sakamoto 2011) analyses the crowd design process more closely. Furthermore, (Dow et al. 2010) investigated an advertisement design task and the factors that lead to better and more diverse results.

Perceptually Optimal Sketches

Our goal is to create minimal sketches of object categories. Minimality of sketches is difficult to define, especially under the secondary constraint that we want the resulting sketches to be easily recognized by machines as well. Restricting the drawing tools is a good approach to address both constraints since allowing arbitrarily shaped lines makes it nearly im-

Figure 3: Two minimal sketches of the category "Face" according to our definition. Both have the same number of lines and both will be recognized correctly by at least 75% of viewers.

possible to find a metric that compares the minimality of two sketches (e.g. an Etch-A-Sketch™consists of a single line but can be visually highly complex).

The most basic approach would be drawing single dots and using the number of pixels as a metric for minimality. However, this approach puts a high demand on the designer, makes iterative minimization very difficult, and is not likely to lead to machine recognizable sketches. We opted for the most basic two-dimensional shape, a straight line segment. All lines are of the same color and width. By only allowing lines, which are easily recognizable by machines (Duda and Hart 1972), the resulting sketches have a good chance of being machine-readable. We use the total number of lines as our metric for minimality. On the one hand, we want the sketches to be minimal, but on the other hand the sketches also need still be recognizable by a high percentage of randomly chosen participants. Here, "recognizable" means that the object category we want sketched should be the foremost word chosen to describe the sketch by at least 75% of the viewers.

A simpler approach would consist of a categorization task. In it, the participant would be asked whether the sketch shows a certain object or not, thus steering the thought processes into this direction and priming the visual recognition system. Such a task would allow too few lines in a sketch defined as optimal (see Figure 2).

Our approach was to show the sketch to the participant and ask for a one-word description in a free naming task. With this, we make sure that the viewers really recognize the sketch, rather than favourably interpret an oversimplified abstraction of it when choosing between a small set of possible answers. In the end, we define a sketch as minimal if at least 75% of viewers correctly recognize the object in a free naming task and there is no other recognizable sketch with fewer lines. This definition allows several different sketches to be minimal at the same time (see Figure 3).

Experiment 1: Iterative Sketch Minimization

During the usual design process, the designer has to predict the perception of the target audience, which is a hard and not easily quantifiable task. When designing our experimental setup, we wanted to make the perception of the sketch an integral part of the design process instead of keeping it a

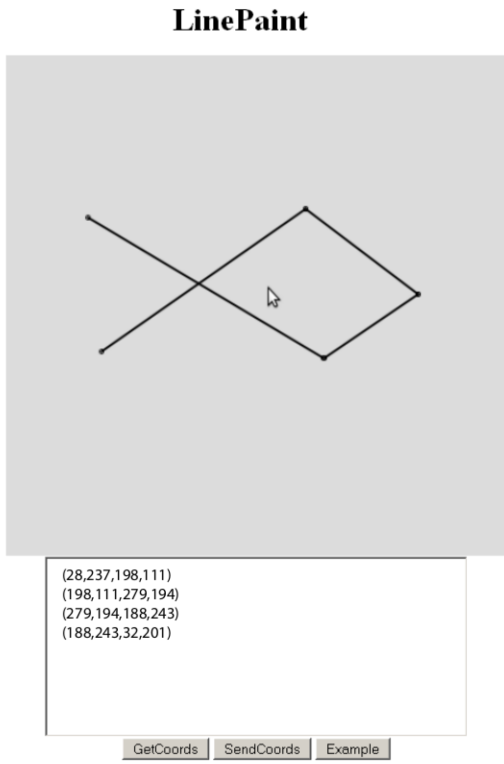

Figure 4: A screenshot of the sketching tool.

separate stage at the end.

When asked to draw a minimal sketch, designers usually cannot judge well how many lines are required without feedback from the perception of the viewers. We therefore developed an iterative scheme, where sketching rounds (in which the designers draw and optimize sketches) and recognition rounds (where a crowd of naïve viewers names the sketches) are interleaved. The categorizations of the viewers are provided as feedback to the designers for the following rounds. Our goal was the development of an automated process where the designer is led by the responses from participants of the recognition rounds. This scheme drastically reduces the requirements on the designers: they no longer have to predict the perception of the audience. As professional designers may be biased by their prediction of the outcome, we opted against using professionals for the experiment.

Experiment design

The core element of the design rounds is a browser-based tool that allows designers to draw sketches that adhere to our constraints. The canvas was deliberately kept as simple as possible and its only capabilities consist of drawing and removing line segments, as well as moving their start and end points. A screenshot of the tool is shown in Figure 4.

The sketches can be exported in a text-based format that lists start and end points of the lines on the canvas. Some designers used this representation to straighten lines or align them to each other. Sketches could also be imported again using this format. We chose 30 common objects representing different categories (see Table 1 for a list of classes) for which we wanted to obtain minimal sketches. These objects cover a large set of classes, most of which do not yet have a symbolic representation that is entrenched in the Western culture.

Book	Letterbox
Window	TV
Car	Lamp
Fish	Fire
Shoe	Tooth
Flower	Key
Coffin	Glasses
Mobile phone	Cup
Sun	Face
Shirt	Bird
Woman	Spider
Computer	Dog
Airplane	Alien
Eiffel Tower	Brain
Vacuum cleaner	Rose

Table 1: Object categories from Experiment 1. The categories were selected to cover a wide array of objects.

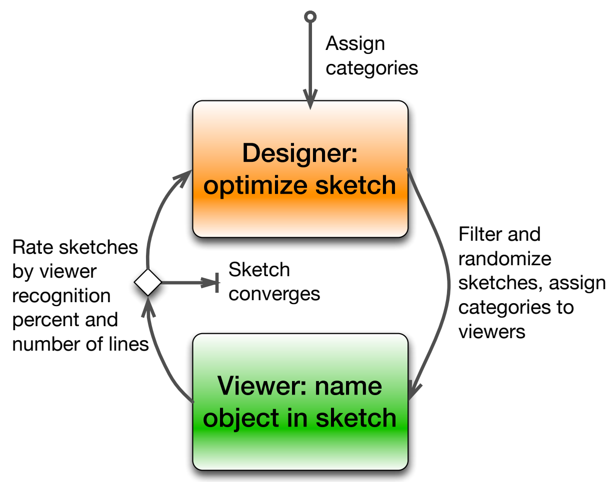

Figure 5: The optimization procedure used to generate the minimized sketches. Designers and a crowd of naïve viewers iteratively create, rate and optimize the sketches.

Using the tool described above, we devised an iterative procedure of interleaved sketching and recognition rounds as depicted in Figure 5. In a sketching round, ten designers created sketches for half of the categories. The categories were randomly assigned to each designer. The resulting sketches were briefly presented to eight naïve viewers in the recognition rounds. After a viewing time of just one second, the participants were asked to name the object that was depicted in the sketch. The naming of a sketch had to consist of one word and had to be a representative identifier for the category.

In order to preserve naïveté, each viewer was presented only once with each category. Since we had ten designers, each drawing half of the categories, we had five sketches per category. In a prefiltering step, we removed the two sketches with the most lines from each category. Therefore, each recognition round required twenty-four viewers (eight viewers for each of the three sketches). The responses from viewers were evaluated first via an automatic script that corrected for spelling errors and synonyms of categories, while also filtering adjectives (e.g., “plane” vs. “airplane” or “flower” vs. “flwer” or “woman” vs. “old woman”). A sketch was considered “correctly classified” if at least six of the eight viewers named it correctly.

We iterated these steps by presenting the designers with feedback about their sketches, i.e. the recognition rates, the one-word responses as well as the sketches and recognition rates for the sketches from the other designers. Given this feedback they were given a chance to improve their sketches, either by reducing the number of lines or by boosting recognition rates with added or altered lines. Because, of the time it takes to gather the feedback for all the sketches the loop was not interactive at this point.

Results

The most stunning result from the first experiment was the small number of rounds it took to optimize the sketches to our criteria. After only two rounds, most of the categories were fully optimized. Most of the designers claimed they could not reduce the number of lines any further without sacrificing recognition rate beyond the allowed cutoff. In some cases, designers had already tried and failed with a lower number of lines used for a sketch. In total, twenty-seven of thirty categories were solved, i.e., one sketch depicting the category was recognized by more than 75% of viewers. For the categories “Alien,” “Brain,” and “Fire,” no minimal sketch could be found, as none of the sketches were recognized by the required percentage of viewers. These categories posed a hard challenge, since they would normally either require curved lines (in the case of “Brain”) or are so unexpected that viewers opted for more common categories even if they did not fit the image that well (e.g., labeling “Alien” as “Ant”). We are, however, confident that these categories could be solved using a higher number of lines.

Figure 6 shows how the number of lines used in a sketch is declined, while the recognition rate increased. We estimated the average number of lines for the third round by asking the sketchers, given their knowledge about previous rounds, how many lines they would use for all the sketches in the next round. The results demonstrate that even naïve designers are capable of correctly interpreting the feedback from viewers and that the presented scheme is able to generate minimal sketches of object categories.

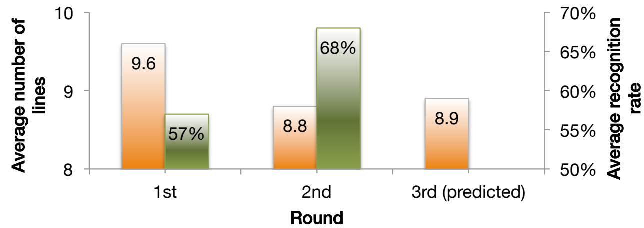

Figure 6: Results from Experiment 1. The orange bars show the average number of lines across all categories, and the green bars show the average recognition rates.

Experiment 2: Crowdsourced Sketch Optimization

In the previous Section, we described a method to obtain minimal, perceptually valid sketches using crowds to provide feedback. In a second experiment, we take the next step and aim to automate as much of the iterative design process as possible. For this, we are using workers from the crowd-sourcing platform Amazon Mechanical Turk (AMT). The nature of micro-task markets such as AMT requires the posted tasks to be fairly small, to be easily explainable and not to require any specialized domain knowledge. Consequently, we are not able to outsource the whole pipeline but only parts of it. Specifically, the creative part in coming up with valid starting points for the optimization requires a considerable amount of creative effort, motivation and design capabilities and does not seem to be easily amendable to AMT. Workers who were asked to create approximately minimal sketches from scratch either used far too many or too few lines and rarely returned useful sketches. However, crowd-sourcing the remainder of the tasks is feasible and leads to a much more favourable trade-off between the time of the designers and the cost of the project.

Experimental design

From the 30 categories in the previous experiment, we selected a subset of ten objects. Our goal was to automatically optimize the sketches for these categories, starting from a single sketch provided by a designer. We therefore selected ten sketches from the first round of the first experiment that were recognized by all viewers as starting points. The ten selected categories are: Book, Computer, Face, Flower, Mailbox, Mobile, Plane, Sun, Tooth and TV. Starting from these initial sketches, we iteratively asked ten AMT workers per sketch to act as viewers and name the presented sketch. Based on these ratings, we asked other workers to optimize the presented sketch. Workers were given the previous sketch as a starting point. If the sketch was recognized correctly by at least seven out of ten viewers, the workers were asked to remove at least one line and, optionally, move the remaining lines to keep the sketch intact. If less than seven viewers correctly named the sketch, we asked the workers to either adjust the existing lines to make the sketch more easily recognizable or add one line. One full round of refinement and rating for all ten sketches usually took about one to two hours.

We iterated this procedure for eight rounds until a state close to convergence had been reached. To counteract some of the problems that could arise during this iterative procedure, we added safeguards after the sketching and the viewing phases. If a sketch was not correctly recognized by at

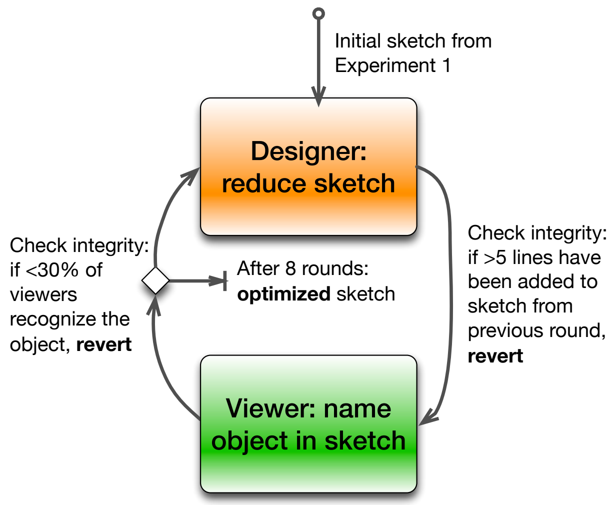

Figure 7: Optimization pipeline of Experiment 2.

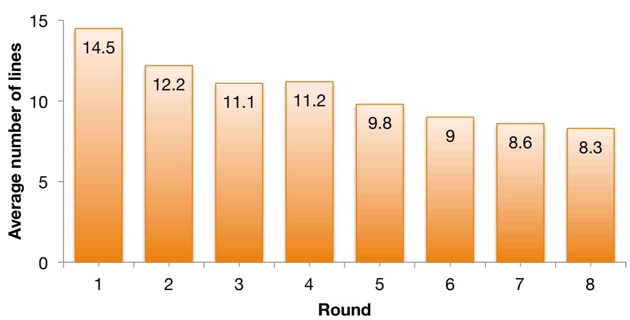

Figure 8: Average number of lines in the eight rounds of Experiment 2. Almost all of the sketches converged at the end of the eighth round (i.e., they were rejected as unrecognizable if any further lines were removed).

least three of ten viewers (according to a script that automatically parsed the answers according to a synonym table), we assumed that the sketcher had reduced it beyond recognition (or, in a few cases, did not use the sketching tool properly and removed all lines from the sketch). As this would provide a poor starting point for the next iteration, we reverted these sketches to their previous version. After the sketching phase, we added another safeguard that ensured sketchers did not add more than five lines per sketching round. Interestingly, some workers were very eager to provide a detailed sketch even though they were informed about the goal of creating a minimal one. In these cases, we also reverted to the sketch from the previous round. Figure 7 shows the iterative pipeline of Experiment 2.

Results

Starting from the original sketches, our automatic pipeline was able to optimize the sketches. Figure 8 depicts the number of lines averaged over all objects for each of the rounds, while Figure 9 displays the average recognition rates.

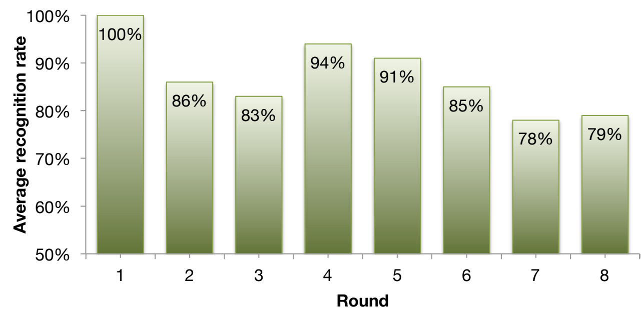

Figure 9: Average recognition rates in the eight rounds of Experiment 2. The initial sketches were assumed to be 100% recognizable.

One can clearly see that the average number of lines converges to a minimum, while the recognition rate stays about the same. After eight rounds, we terminated the sketching, because a minimum had been reached for most sketches, i.e., the sketch was recognizable but after removing one more line it became unrecognizable and was therefore reverted to the previous version. Three examples of the the evolution of the sketches are shown in Figure 10.

Discussion

In this paper, we have shown a novel approach to semi-automatically minimize sketches to a perceptual optimum, using collective intelligence to refine and rate sketches. Design tasks are usually left to individuals or groups that are highly specialized and trained, because the optimization criterion is usually hard to quantify. We have shown that, given a clean definition of the optimization function, even complex design tasks can be crowd-sourced and solved almost automatically using micro-task markets. By including crowds as an unbiased, collective perception step in the design process, we can ensure that our results are not only minimal but also perceptually valid. Furthermore, by restricting the sketches to straight lines, we can ascertain that a machine vision system would also have a good chance to parse the resulting sketches.

It is important to note that we cannot guarantee that the resulting sketches are minimal, since we cannot test all possible sketches. We did not find any sketches with less lines that achieve the required recognition rate. For most situations, this is only a theoretically relevant argument because we can assume that the found local minimum is close to the global one. However, there are some situations where two very different approaches to sketch the category exist (see Figure 11) and it might initially not be clear which variant will lead to a minimal sketch. Therefore, it is important to have a group of designers develop a wide array of possible sketches which can then be optimized using the approach described above.

We see clear applications for other design processes where traditionally instant feedback by crowds and iterative optimization is foregone for intuitive knowledge of highly trained professional designers. As we only propose a semi-

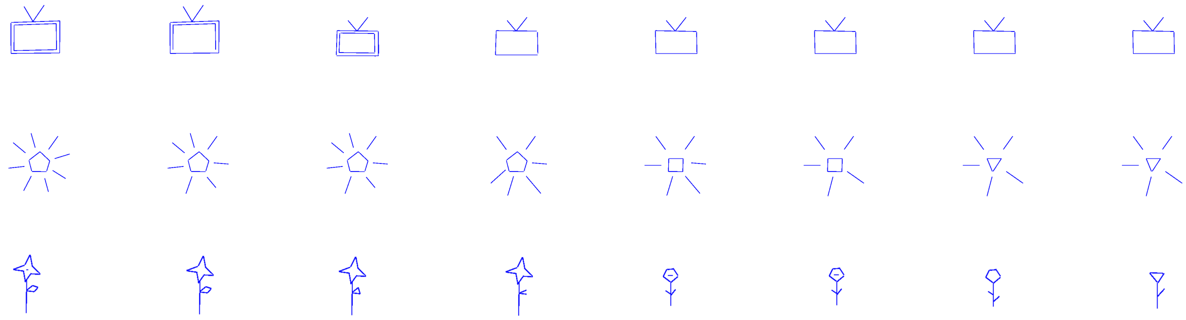

Figure 10: The evolution of categories “TV”, “Sun” and “Flower” throughout the eight rounds of Experiment 2 (from left to right). Transitions where the sketches do not change are situations where the integrity checks were triggered and the drawings were reverted to the previous state.

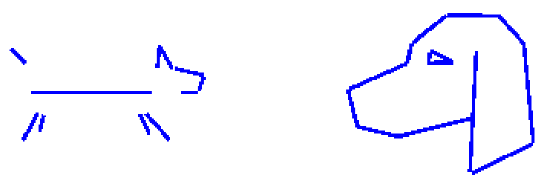

Figure 11: Two alternative non-minimal starting sketches for the category ’Dog’. As we cannot test all possible sketches we can only claim that any sketch is minimal with respect to all the alternatives we have tried so far.

automatic system, we do not aim to replace professional designers but to provide them with a tool that allows them to augment their work. Our system assumes that the rating feedback from AMT is representative for the general populace, which might in some situations not be the case due to cultural differences. For example, a minimal sketch of the category “Woman” could be different in Western and Asian cultures. Such differences could require different optimizations for different cultural contexts; see, for instance, a race effect in (Beaupré and Hess 2006).For the ten categories we optimized using AMT (which has a high ratio of Indian workers), we did not observe any bias along those lines.

Outlook

As a next step, we plan to extend our approach to design applications beyond sketch minimization and icon design. Furthermore, semi-automatic design methods offer the interesting property of containing subtasks that are not easily amendable to crowdsourcing at micro-task markets as AMT (e.g. creating the initial sketches). By analyzing which parts of the task require specialized domain knowledge, we hope to gain an insight into and find a predictor for the optimal trade-off between specialized individuals and crowdsourcing for design tasks.

Acknowledgements

We thank all participants of our experiments for their time and effort. David Engel is supported by NSF grant IIS-0963285. Sebastian Schultheiss is supported by DFG grant RA1894/1-1.

References

Beaupré, M. G., and Hess, U. 2006. An ingroup advantage for confidence in emotion recognition judgments: The moderating effect of familiarity with the expressions of outgroup members. *Personality and Social Psychology Bulletin* 32(1):16.

Calpin, S. 2001. *Icon design: graphic icons in computer interface design*. New York, NY, USA: Watson-Guptill.

Chen, P. P. 2003. Toward a methodology of graphical icon design. In *5th International Symposium on Multimedia Software Engineering*, 120–121. New York, NY, USA: IEEE Press.

Dow, S. P.; Glassco, A.; Kass, J.; Schwarz, M.; Schwartz, D. L.; and Klemmer, S. R. 2010. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Trans. Comput.-Hum. Interact.* 17(4):18:1–18:24.

Dow, S.; Kulkarni, A.; Bunge, B.; Nguyen, T.; Klemmer, S.; and Hartmann, B. 2011. Shepherding the crowd: managing and providing feedback to crowd workers. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, CHI EA ’11, 1669–1674. New York, NY, USA: ACM.

Duda, R. O., and Hart, P. E. 1972. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM* 15(1):11–15.

Engel, D.; Kottler, V. A.; Malisi, C. U.; Röttig, M.; Schultheiss, S. J.; Willing, E. M.; Curio, C.; and Bülthoff, H. H. 2010. Optimizing minimal sketches of visual object categories. *Perception* 39(ECVP Abstract Supplement):11.

Field, M.; Valentine, S.; Mary, S.; Linsey, J.; and Hammond, T. 2011. Sketch recognition algorithms for comparing complex and unpredictable shapes. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Global Standards One. 2010. *GS1 Data Carriers: EAN/UPC BarCodes*. Brussels, Belgium.

Heer, J., and Bostock, M. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, 203–212. ACM.

Khatib, F.; DiMaio, F.; Foldit Contenders Group; Foldit Void Crushers Group; Cooper, S.; Kazmierczyk, M.; Gilski, M.; Krzywda, S.; Zabranska, H.; Pichova, I.; Thompson, J.; Popović, Z.; et al. 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature Structural & Molecular Biology* 18(10):1175–1177.

Mahoney, J. V., and Fromherz, M. P. J. 2002. Three main concerns in sketch recognition and an approach to addressing them. 105–112.

Paulson, B., and Hammond, T. 2008. Marqs: retrieving sketches learned from a single example using a dual-classifier. *Journal on Multimodal User Interfaces* 2(1):3–11.

Snodgrass, J. G., and Vanderwart, M. 1980. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory* 6(2):174.

Toyoda, E. 2009. An analysis of L2 readers' comments on kanji recognition. *Electronic Journal of Foreign Language Teaching* 6(1):5–20.

Wah, C.; Branson, S.; Perona, P.; and Belongie, S. 2011. Multiclass recognition and part localization with humans in the loop. In *International Conference on Computer Vision*.

Yu, L., and Nickerson, J. V. 2011. Cooks or cobblers?: crowd creativity through combination. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, 1393–1402. New York, NY, USA: ACM.

Yu, L., and Sakamoto, Y. 2011. Feature selection in crowd creativity. In *Proceedings of the 6th international conference on Foundations of augmented cognition: directing the future of adaptive systems*, FAC'11, 383–392. Berlin, Heidelberg: Springer-Verlag.