Building a Gold Standard for Perceptual Sketch Similarity

S. Cakmak and T. M. Sezgin

Koc University, Istanbul, Turkey

Abstract

Similarity is among the most basic concepts studied in psychology. Yet, there is no unique way of assessing similarity of two objects. In the sketch recognition domain, many tasks such as classification, detection or clustering require measuring the level of similarity between sketches. In this paper, we propose a carefully designed experiment setup to construct a gold standard for measuring the similarity of sketches. Our setup is based on table scaling, and allows efficient construction of a measure of similarity for large datasets containing hundreds of sketches in reasonable time scales. We report the results of an experiment involving a total of 9 unique assessors, and 8 groups of sketches, each containing 300 drawings. The results show high interrater agreement between the assessors, which makes the constructed gold standard trustworthy.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [User Interfaces]: Evaluation/methodology—Interaction styles (e.g., commands, menus, forms, direct manipulation) H.1.2 [User/Machine Systems]: Human factors—Human information processing

1. Introduction and Related Work

Perceptual similarity is the subjective similarity between two stimuli as perceived by the observer. Our purpose in this experiment is to examine the issue of sketch similarity from the perspective of the human observers in order to build a gold standard. We aim to use this gold standard in the evaluation phase of the sketch clustering methods that we develop. Different behavioral methods for acquiring similarities are available in the literature. The most preferred methods are pairwise similarity judgments, perceptual confusion tasks, free sorting, single arrangement, and multi arrangement [KM12]. In pairwise similarity judgments, each pair of items is presented in isolation and the subject rates the similarity on a scale. In confusion tasks, subjects are presented with two similar items and asked whether the items are the same or different. The probability of the confusion between these two items is measured simultaneously. These two methods are very slow since $(n^2-n)/2$ separate judgments are required, where n is the number of items. In free sorting, subjects are instructed to place items into groups. This method actually suffers from graded similarity estimates for individuals. Single arrangement method has been proposed to overcome this problem, where the subject arranges items in $2D$ by considering the distances between the items which reflect dissimilarities [Gol94]. However, spatially arranging 300 sketch cards at the same time is not possible both on a table and on a computer screen in our case. Even worse, if we chose the multi arrangement method, multiple item subsets would be arranged iteratively, causing the experiment to last for hours.

Due to the time complexity of other methods, we decided to follow free sorting method in the experiment. Free sorting method is also consistent with the famous Gestalt principle of perceptual organization, which states that similar things will tend to be grouped together by humans [Köh70]. In order to overcome the lack of graded similarity estimates problem, we averaged the scores over all participants and acquired the graded similarity estimates. Throughout the experiment, we assumed that the frequency with which two items are placed in the same group is proportional to their perceptual similarity.

Another important issue while designing the experiment was to decide the experiment environment. Due to the restricted size of any possible display, a computer interface would be a limitation for the participants. Rather than designing a computer interface for the experiment, we prepared a more realistic and a more interactive environment. Our experiment design was inspired from table scaling experiment, which was previously used in the perceptual image similarity context [RFS*98]. We took Gurcuoglu's sketch ranking experiment as an example and modified it to meet our needs [Gur14].

2. Method and Results

In order to make the experiment conditions more realistic, we conducted the experiment with the use of printed versions of sketches. Two human assessors were invited to a previously prepared meeting room at the same time, where a suitable table was present to spread the printed sketch cards (Figure 1). A sample sketch card used in

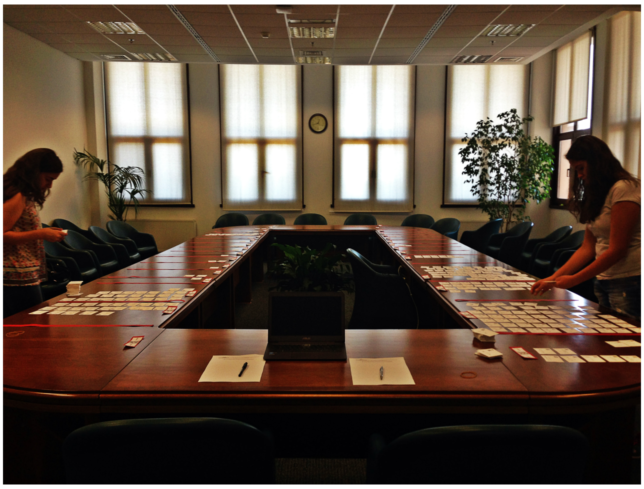

Figure 1: *A snapshot from the experiment: Two assessors grouping perceptually similar sketches*

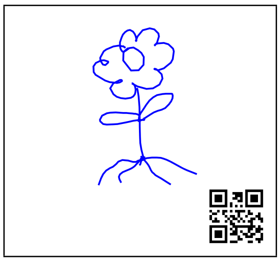

Figure 2: *A sample sketch card*

the experiment is presented in Figure 2. Human assessors were provided with 300 sketches and they were asked to gather perceptually similar sketches into groups within a limited time period. Assessors did not have any prior knowledge about the provided sketches. After the assessors finished the grouping task, experiment conductor recorded the group information of the sketch cards with the help of a QR code scanner program.

Each assessor participated in the experiment for 8 different sketch categories on different days and each sketch category was grouped by 9 different assessors. Sketch categories consist of geometric shapes and everyday objects. There were 4 female and 5 male assessors participating in the experiment, whose ages vary between 16-24. In total, perceptual similarity data for 8x300 = 2400 individual sketches were efficiently obtained (approximately 40-45 minutes by session). After obtaining the data from 9 different assessors for each category, we validated the consistency of assessors within themselves to make sure that the acquired data is not coherent by chance. For this purpose, BCubed extrinsic clustering comparing metric was used and consistency score among assessors was calculated [AGAV09]. Figure 3 shows the BCubed FScores among participants for all sketch categories. BCubed FScores with higher values on the half left of Figure 3 prove that assessors are more consistent within themselves when compared with lower values coming from random assignments.

After proving the general consistency among assessors, for each category, $\binom{300}{2}$ pairs of sketches were examined and the number of assessors claiming this pair as similar was identified. By this means, a similarity matrix consisting of the perceptual similarity scores per sketch category was constructed. Figure 4 illustrates the perceptual similarity matrix. Thanks to this perceptual similarity matrix obtained from the common judgments of 9 assessors, a gold standard was built. This gold standard will be used while evaluating our clustering algorithms' success on these sketches.

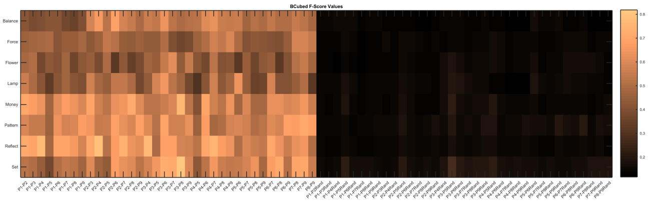

Figure 3: *BCubed FScore for all sketch categories (On the left: Participant to Participant - On the right: Participant to Random)*

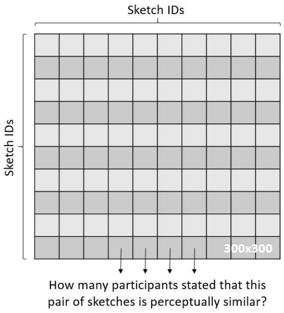
Sketch IDs
Sketch IDs
300x300
How many participants stated that this pair of sketches is perceptually similar?

Figure 4: *Illustration of perceptual similarity matrix*

3. Acknowledgements

The authors gratefully acknowledge the support and funding of TUBITAK under grant number 113E059.

References

[AGAV09] AMIGÓ E., GONZALO J., ARTILES J., VERDEJO F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval 12*, 4 (2009), 461–486. 2

[Gol94] GOLDSTONE R.: An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, and Computers 26*, 4 (1994), 381–386. 1

[Gur14] GURCUOGLU B.: *Learning people's perception of messiness for hand-drawn sketches*. Master's thesis, Koc University, Istanbul, Turkey, 2014. 1

[KM12] KRIEGESKORTE N., MUR. M.: Inverse mds: inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology 3*, 245 (2012), 1–13. 1

[Köh70] KÖHLER W.: *Gestalt psychology: An introduction to new concepts in modern psychology*. WW Norton & Company, 1970. 1

[RFS*98] ROGOWITZ B. E., FRESE T., SMITH J. R., BOUMAN C. A., KALIN E.: Perceptual image similarity experiments. *In Photonics West'98 Electronic Imaging* (1998), 576–590. 1