Characterizing Cross-Person and Cross-Cultural Variability in Meanings Through Millions of Sketches

Molly Lewis
mollyllewis@gmail.com
Department of Psychology
Carnegie Mellon University

Anjali Balamurugan
XXX
XXX
Carnegie Mellon University

Bin Zheng
XXX
XXX
Carnegie Mellon University

Gary Lupyan
lupyan@wisc.edu
Department of Psychology
University of Wisconsin-Madison

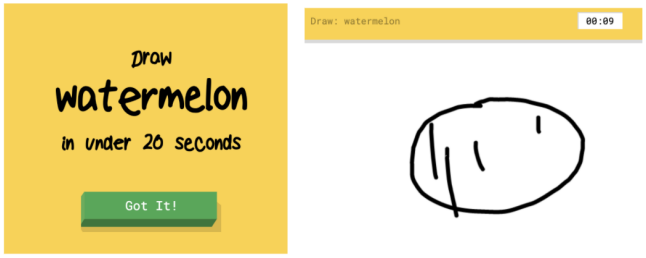

Figure 1: Screenshots of the Quick, Draw! App (https://quickdraw.withgoogle.com/). Participants arefirst cued with a word (e.g. "watermelon"; left), and then asked to sketch the corresponding object in under 20 seconds (right).

Abstract

Include no author information in the initial submission, to facilitate blind review. The abstract should be one paragraph, indented 1/8 inch on both sides, in 9˜point font with single spacing. The heading 'Abstract' should be 10˜point, bold, centered, with one line of space below it. This one-paragraph abstract section is required only for standard six page proceedings papers. Following the abstract should be a blank line, followed by the header 'Keywords' and a list of descriptive keywords separated by semicolons, all in 9˜point font, as shown below.

Keywords: Add your choice of indexing terms or keywords; kindly use a semi-colon; between each term.

Introduction

Study 1: Estimating drawing similarity

To quantify the similarity between two arbitrary drawings, we collected human judgments of the similarity for a sample of drawing pairs.

Methods

Participants We recruited 331 participants through Amazon Mechanical Turk and an undergraduate subject pool. We excluded 64 participants who missed an attention check question (see procedure below). Our final sample of included 267 participants.

Stimuli Drawings were taken from the Quick, Draw! dataset collected by Google (https://github.com/googlecreativelab/quickdraw-dataset). The drawings were collected through an online app (https://quickdraw.withgoogle.com/) in which participants were cued with a word (e.g., "watermelon") and asked to sketch the corresponding object in under 20 seconds (Figure 1). As participants sketched, a neural net trained on other participants' drawings made guesses about the cue word. Once the neural net guessed correctly, the app progressed to the next word cue. Each participant completed up to 6 drawings per session. Each drawing is represented as a X x X binary matrix. The quickdraw dataset contains over XX drawings collected from participants worldwide.

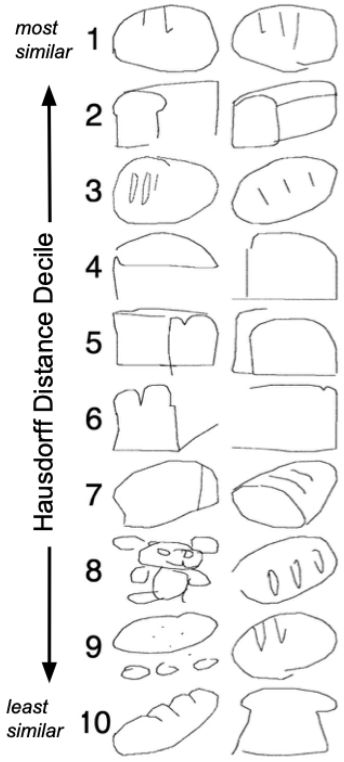

Figure 2: Example stimuli pairs for the cue "bread" sampled from each hausdorff distance decile (1 = most similiar; 10 = least similar).

For the current study, we sampled 1,000 drawing pairs for each of five word cues: "tree," "bread," "chair," "house" and "bird." In order to include a range of drawing similarities in our stimuli, we quantified the similarity between drawings in a pair using a computational measure of visual image similar-

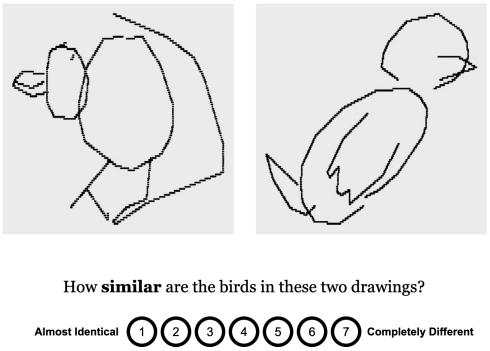

Figure 3: Screenshot of the norming procedure in Study 1. Participants were presented with pairs of drawings from the Quick, Draw! dataset and asked to make judgments about their visual similarity.

ity commonly used in machine vision, called hausdorff distance (Huttenlocher, Klanderman, & Rucklidge, 1993; Taha & Hanbury, 2015). Informally, hausdorff distance quantifies the similarity of two images by treating each image as a set of x-y coordinates, and calculating the Euclidean norm between each point in one image to the closest point in the other. The hausdorff distance is the maximum of these pairwise distances (the distance between the most mismatched points). We calculated hausdorff distance for each drawing pair and then sampled 20 drawing pairs from each hausdorff distance decile (see Figure 2). Our final stimuli list included 200 drawing pairs for each of the 5 target cues .

Procedure Participants were instructed to rate how similar pairs of drawings were to each other on a 7-pt Likert scale, ranging from "almost identical" to "completely different" (Figure 3). Each participant rated a sample of 50 drawing pairs from a single cue word. As an attention check, we also included two additional trials where the two drawings were identical to each other. Participants were excluded from the final sample if they responded 3 or higher on the Likert scale for either of these two trials. Each drawing pair was rated by 13.34 participants on average (SD = 7.04).

Results

Log Hausdorff distance was moderately positively correlated with human judgments of visual dissimilarity ($r(998) = 0.39$, $p < .0001$; Figure 4), accounting for 15.15% percent of the variance in human judgments.

We next tried to better predict human similarity judgment using additional computational measures of similarity. We examined three new measures: Log average Hausdorff distance [AHD; Taha & Hanbury (2015)], Euclidean distance (ED) and Mahalanobis distance (MD). Log average Hausdorff distance is similar to the Hausdorff distance metric described above, but is less sensitive to outliers. Average Hausdorff distance is calculated by taking the Euclidean norm between each point in one image to the closest point in the other,

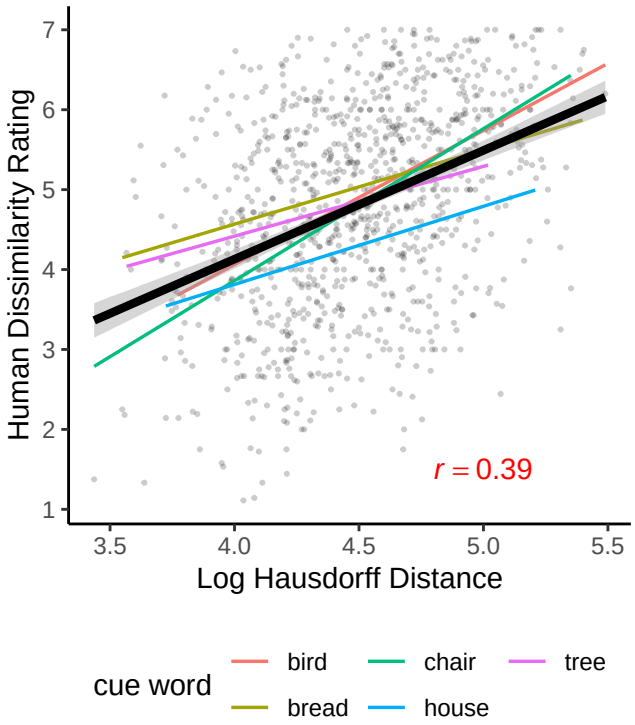

Figure 4: Relationship between human judgments of drawing similarity and drawing similarity estimated from a computational measure, log hausdorff distance. Each point corresponds to a drawing pair (N = 1,000). The color lines show the best fit for each of the five individual cue words; black line shows the best fit for all drawing pairs and corresponding standard error.

and then taking the average across all point pairs and log transforming. Euclidean distance is calculated as the average pairwise Euclidean distance between all points. Mahalanobis distance [@] is similar to Euclidean distance, but takes into account the correlation of points in the drawings.

All three distance measures were correlated with human similarity judgments, (AHD-ED: $r(998) = 0.6$, $p < .0001$; AHD-MD: $r(998) = 0.44$, $p < .0001$; ED-MD: $r(998) = 0.24$, $p < .0001$), and with each other (AHD: $r(998) = 0.34$, $p < .0001$; ED: $r(998) = 0.22$, $p < .0001$; MD: $r(998) = 0.22$, $p < .0001$). We next fit an additive linear model predicting human judgments with each of these three predictors. This model accounted for '24% of the variance in human judgments (see Table 1 for model parameters). Figure Figure 5) shows a 2D multi-dimensional scaling solution of the predicted human similarity ratings for a sample of one hundred "bird" drawings. In sum, Study 1

Study 2: Cross-person meaning variability

items: which items are more variable across people?

word predictors of variability (Concreteness, Frequency, Semantic category, AoA)

joy plots of items with high and log variability

	Estimate	SE	t-value	Pr(>|t|)
(Intercept)	-3.52	0.67	-5.25	<.001
Log Avg. Haus.	1.19	0.12	9.61	<.001
Mahalanobis	3.78	0.39	9.80	<.001
Euclidean	-0.02	0.00	-7.63	<.001

Table 1: Parameters of an additive linear model predicting human similarity judgment of 1,000 drawing pairs in Study 1 from three computational similarity measures. Log Avg. Haus. = Log average Hausdorff distance.

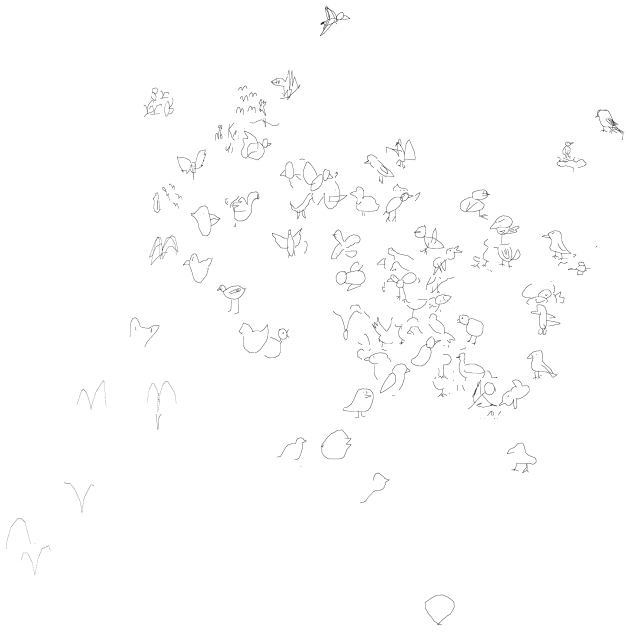

Figure 5: Multi-dimensional scaling solution of pairwise similarity of 100 bird drawings judged in Study 1. Similarity is estimated from as the predicted values from a model predicting human judgements with three computational similarity measures (log average Hausdorff distance, Mahalnobis distance, and Euclidean distance).

countries - across items, which countries have the most variability?

country predictors of variability

prototype fig.

Study 3: Cross-cultural meaning variability

predictors of cross-cultural similarity (Geographical distance, Cultural distance (dspace), weather, Language distance, semantic alignment?)

interactions with item? (with embedding models?)

Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

References

10 Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(9), 850–863.

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, *15*(1), 29.