Journal of Graph Algorithms and Applications
http://www.cs.brown.edu/publications/jgaa/
vol. 6, no. 3, pp. 225-254 (2002)

A User Study in Similarity Measures for Graph Drawing

Stina Bridgeman
Department of Computer Science
Colgate University
Hamilton, NY 13346
http://cs.colgate.edu/faculty/stina/
stina@cs.colgate.edu

Roberto Tamassia
Center for Geometric Computing
Department of Computer Science
Brown University
Providence, RI 02912–1910
http://www.cs.brown.edu/~rt/
rt@cs.brown.edu

Abstract

The need for a similarity measure for comparing two drawings of graphs arises in problems such as interactive graph drawing and the indexing or browsing of large sets of graphs. Many applications have been based on intuitive ideas of what makes two drawings look similar — for example, the idea that vertex positions should not change much. In this paper, we formally define several of these intuitive ideas of similarity and present the results of a user study designed to evaluate how well these measures reflect human perception of similarity.

Communicated by Michael Kaufmann: submitted February 2001;revised January 2002 and June 2002.

Research supported in part by the National Science Foundation under grants CCR–9732327, CCR–0098068 and CDA–9703080, and by the U.S. Army Research Office under grant DAAH04-96-1-0013. Work completed while the first author was at Brown University.

1 Introduction

The question of how similar two drawings of graphs are arises in a number of situations. One application is interactive graph drawing, where the graph being visualized changes over time and it is important to preserve the user's "mental map" [12] of the original drawing as much as possible so the user does not need to spend a lot of time relearning the drawing after each update. Having to relearn the drawing can be a significant burden if the graph is updated frequently. Animation can be used to provide a smooth transition between the drawings and can help compensate for greater changes in the drawing, but it is still important to maintain some degree of similarity between the drawings to help the user orient herself to the new drawing. Related to interactive graph drawing is layout adjustment, where an existing drawing is modified so as to improve an aesthetic quality without destroying the user's mental map.

Another application is in indexing or browsing large sets of graphs. An example of a graph browser is contained the SMILE graph multidrawing system of Biedl et. al. [1]. The SMILE system tries to find a good drawing by producing many drawings of the graph and letting the user choose among them, rather than trying to code the properties of a good drawing into the algorithm. The graph browser arranges the drawings so that similar ones are near each other, to help the user navigate the system's responses. Related to this is the idea of using similarities between drawings as a basis for indexing and retrieval. Such a system has applications in character and handwriting recognition, where a written character is transformed into a graph and compared to a database of characters to find the closest match.

Let M be a similarity measure defined so that M's value is always nonnegative and is 0 when the drawings are identical. In order to be useful, M should satisfy three properties:

Rotation: Given drawings D and D', $M(D, D'_\theta)$ should have the minimum value for the angle a user would report as giving the best match, where D'_θ is D' rotated by an angle of θ with respect to its original orientation.

Ordering: Given drawings D, D', and D'', $M(D, D') < M(D, D'')$ if and only if a user would say that D' is more like D than D'' is like D.

Magnitude: Given drawings D, D', and D'', $M(D, D') = \frac{1}{c}M(D, D'')$ if and only if a user would say that D' is c times more like D than D'' is like D.

This paper describes a user study performed in order to evaluate several potential similarity measures with respect to rotation and ordering, and to test a possible method for obtaining data to be used for evaluating measures with respect to magnitude. Data cannot be collected directly for the magnitude part as it can be for rotation and ordering because it is very difficult to assign numerical similarity values to pairs of drawings; Wickelgren [18] observes that it is more difficult to assign numerical values than to judge ordering. As a result, other data must be gathered — for example, response times on a particular task —

with the hope that the data is sufficiently related to the actual similarity values to be useful. This can be partially tested by using the data (e.g., response times) to order the drawings, and determining whether the results are consistent with user responses on the ordering part.

This study improves on our previous work [3] in several ways:

- **More Experimental Data:** A larger pool of users (103 in total) was used for determining the "correct" behavior for the measure.
- **Refined Ordering Part:** Users made only pairwise judgments between drawings rather than being asked to order a larger set.
- **Addressing of Magnitude Criterion:** The previous experiment did not address magnitude at all.
- **More Realistic Drawing Alignment:** The previous drawing alignment method allowed one drawing to be scaled arbitrarily small with respect to the other; the new method keeps the same scale factor for both drawings.
- **Refinement of Measures:** For those measures computed with pairs of points, pairs involving points from the same vertex are skipped.
- **New Measures:** Several new measures have been included.

We describe the experimental setup in Section 2, the measures evaluated in Section 3, the results in Sections 4 and 5, and conclusions and directions for future work in Section 6.

2 Experimental Setup

This study focuses on similarity measures for orthogonal drawings of nearly the same graph. "Nearly the same graph" means that only a small number of vertex and edge insertions and deletions are needed to transform one graph into the other. In this study, the graphs differ by one vertex and two or four edges. The focus on orthogonal drawings is motivated by the availability of an orthogonal drawing algorithm capable of producing many drawings of the same graph, and by the amount of work done on interactive orthogonal drawing algorithms. (See, for example, Biedl and Kaufmann [2], Fößmeier [8], Papakostas, Six, and Tollis [14], and Papakostas and Tollis [15].) Producing multiple drawings of the same graph is important because it can be very difficult to judge if one pair of drawings is more similar than another if the graphs in each pair are different.

2.1 Graphs

The graphs used in the study were generated from a base set of 20 graphs with 30 vertices each, taken from an 11,582-graph test suite. [7] Each of 20 base graphs was first drawn using Giotto [17]. Each of the Giotto-produced base drawings was modified by adding a degree 2 and a degree 4 vertex, for a total of 40 modified

drawings. Each modified drawing is identical to its base drawing except for the new vertex and its adjacent edges, which were placed in a manner intended to mimic how a user might draw them in an editor. Because InteractiveGiotto (used in the next step) preserves edge crossings and bends, routing the edges realistically avoids introducing a large number of extra crossings. Finally, a large number of new drawings were produced for each modified drawing using InteractiveGiotto [4], and four drawings were chosen from this set. The four drawings chosen range from very similar to the base drawing to very different.

2.2 Definition

The experiment consisted of three parts, to address the three evaluation criteria. In all cases, the user was asked to respond as quickly as possible without sacrificing accuracy. To promote prompt responses, each trial timed out after 30 seconds if the user did not respond.

Rotation Part The rotation part directly addresses the rotation criterion. The user is presented with a screen as shown in Figure 1. The one drawing D on the left is the base drawing; the eight drawings $D_1, \ldots, D_8$ on the right are eight different orientations of the same new (InteractiveGiotto-produced) drawing derived from D. The eight orientations consist of rotations by the four multiples of $\pi/2$, with and without an initial flip around the x-axis. For orthogonal drawings, only multiples of $\pi/2$ are meaningful since it is clear that rotation by any other angle is not the correct choice. The vertices are not labelled in any of the drawings to emphasize the layout of the graph over the specifics of vertex names.

The user's task is to choose which of $D_1, \ldots, D_8$ looks most like the base drawing. A "can't decide" button is provided for cases in which the drawings are too different and the user cannot make a choice. The user's choice and the time it took to answer are recorded.

Ordering Part The ordering part directly addresses the ordering criterion. In this part, the user is presented with a screen as shown in Figure 2. The one drawing D on the left is the base drawing; the two drawings D_1 and D_2 on the right are two different (InteractiveGiotto-produced) new drawings of the same modified drawing derived from D.

The user's task is to choose which of D_1 and D_2 looks most like the base drawing. A "can't decide" button is provided for cases in which the drawings are too different and the user cannot make a choice. The user's choice and the time it took to answer are recorded.

Difference Part The difference part addresses the magnitude criterion by gathering response times on a task, with the assumption that a greater degree of similarity between the drawings will help the user complete the task more quickly. The screen presented to the user is shown in Figure 3. The drawing

D on the left is the base drawing; the drawing D_1 on the right is one of the InteractiveGiotto-produced new drawings derived from D.

The user's task is to identify the vertex present in the right drawing that is not in the left drawing. The vertices are labelled with random two-letter names — corresponding vertices in drawings in a single trial have the same name, but the names are different for separate trials using the same base drawing to prevent the user from simply learning the answer. Displaying the vertex names makes the task less difficult, and mimics the scenario where the user is working with a dynamically updated graph where the vertex labels are important.

The user's choice and the time it took to answer are recorded.

2.3 Methodology

The three parts were assigned to students as part of a homework assignment in a second-semester CS course at Brown University. A total of 103 students completed the problem.

Before being assigned the problem, the students had eight lectures on graphs and graph algorithms, including one on graph drawing. They had also been assigned a programming project involving graphs, so they had some familiarity with the subject.

The homework problem made use of an online system which presented the displays shown in Figures 1, 2, and 3. A writeup was presented with the problem explaining how to use the system, and the directions were summarized each time the system was run.

Each of the three parts was split into four runs, so the students would not have to stay focused for too long without a break. The graphs used were divided into 10 batches: the first batch (the practice batch) contained two modified drawings along with their associated new drawings, and each of the other nine batches contained three modified drawings and the associated new drawings. All of the students were assigned the practice batch for the first run of each part, and were randomly assigned three of the other batches for later runs so that each batch was completed by 1/3 of the students. A given student worked with the same batches for all three of the parts. Within each run of the system, the individual trials were presented in a random order and the order of the right-hand drawings in the rotation and ordering parts was chosen randomly.

On average, students spent 6.9 minutes total on all four runs the rotation task (out of 22 minutes allowed), 8.3 minutes on the ordering task (out of 33), and 12.9 minutes on the difference task (out of 22).

After the students completed all of parts, they answered a short questionnaire about their experiences. The questions asked were as follows:

1. **(Ordering and Rotation)** What do you think makes two drawings of nearly the same graph look similar? Are there factors that influenced your decisions? Did you find yourself looking for certain elements of the drawing in order to make your choice?

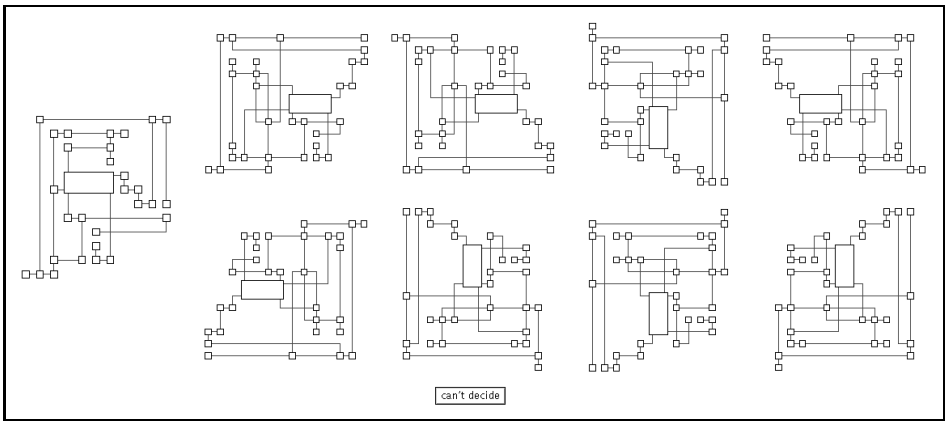


Figure 1: The rotation part.

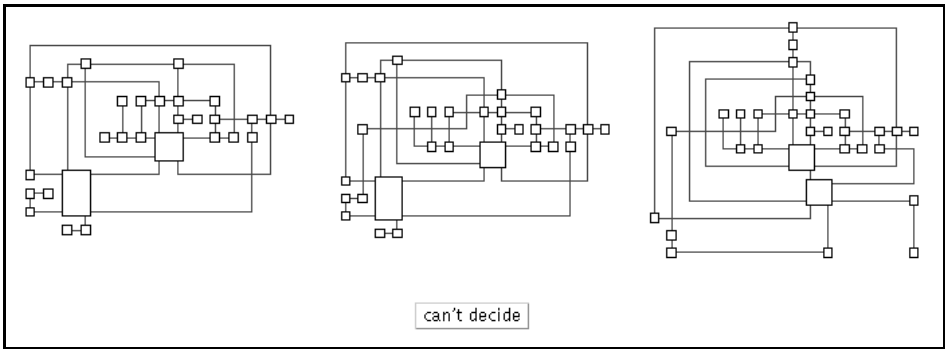


Figure 2: The ordering part.

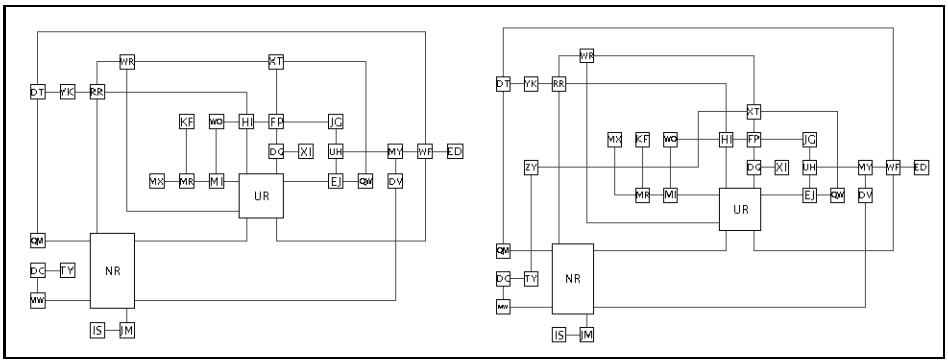


Figure 3: The difference part.

2. **(Difference)** What factors helped you locate the extra vertex more quickly? Did you compare the overall look of the two drawings in order to aid your search, or did you just scan the second drawing?

3. **(All Parts)** As you consider your answers, think about what this means for a graph drawing algorithm that seeks to preserve the look of the drawing. What types of things would it have to take into account?

3 Measures Evaluated

All of the measures evaluated in this study are described below. Most are the same as or similar to those described in [3]; the primary difference is that all of the measures have been scaled so that 0 means the drawings are identical and 1 is the maximum difference. The upper bound is frequently based on the worst-case scenario for point positioning and may not be achievable by an actual drawing algorithm. (For example, the upper bound may only be achieved when all of the vertices are placed on top of each other, an impossible situation with most drawing algorithms.)

3.1 Preliminaries

Corresponding Objects Most of the measures make use of the fact that the graph in the drawings being compared is the same. (If the graphs are not the same, those parts that are different are ignored and only the common subgraphs used.) This means that each vertex and edge of G has a representation in each of the drawings, and it is meaningful to talk about the *corresponding vertex or edge* in one drawing given a vertex or edge in the other drawing.

Point Set Selection All measures except for the shape measures are defined in terms of point sets derived from the edges and vertices of the graph rather than the edges and vertices themselves. Like vertices and edges, each point in one drawing has a corresponding point in the other drawing.

Points can be selected in a variety of ways; inspired by North [13], one point set contains the four corners of each vertex. Inspired by feedback from the study (section 5.2), a second "borders-only" point set was also considered. This is a subset of the "full" point set which contains only those points outside an ellipse centered within the drawing's bounding box and with radii 90% of the width and height of the bounding box, or whose corresponding point in the other drawing meets this criterion.

A change from the previous experiment [3] is that the computation of the nearest neighbor measures skips pairs of points derived from the same vertex. This can have a great effect because a point's nearest neighbor will often be another corner of the same vertex, which does not convey much information about how that vertex relates to other vertices in the drawing. In this study, a point's nearest neighbor is the nearest point from a different vertex. This is not

explicitly written in the definitions below for clarity of notation, but it should be assumed.

Drawing Alignment For the measures involving the comparison of coordinates between drawings, the value of the measure is very dependent on how well the drawings are aligned. Consider two identical drawings, but let the x-coordinate of each point in the second drawing be one bigger than the x-coordinate of the corresponding point in the first drawing. The average distance moved by each point will be reported as 1, even though the drawings actually are the same. Aligning the drawings before comparing the coordinates removes this effect.

In the previous experiment [3], alignment was done by simultaneously adjusting the scale and translation of one drawing with respect to the other so as to minimize the distance squared between corresponding points. This had the effect of potentially reducing one drawing to a very small area if the drawings did not match well. This has been replaced by a new alignment method which separates the determination of the scale and translation factors into two steps. First, the scale factor is set to ensure that the two drawings are drawn to the same scale. Since the drawings are orthogonal drawings, there is a natural underlying grid which can be used to adjust the scale. Once scaled, the translation factor is chosen so as to minimize the distance squared between corresponding points. The new alignment method is intended to better match how a person might try to match up the drawings — it does not seem likely that someone would mentally shrink or enlarge one drawing drastically with respect to the other, but rather would work with the current scale and try to adjust the translation.

Suitability for Ordering vs. Rotation and Ordering Some of the measures do not depend on the relative rotation of one drawing with respect to the other. This means that they fail the rotation test, however, they are included because there may be situations in which the measure is not being used to determine the proper rotation for the drawings. Furthermore, a successful ordering-only measure could be combined with one which is successful at rotation but less so at ordering to obtain a measure which is good at both. Measures suitable for ordering only are marked [order only] below.

Notation In the following, P and P' will always refer to the point sets for drawings D and D', respectively, and $p' \in P'$ will be the corresponding point for $p \in P$ (and vice versa). Let $d(p, q)$ be the Euclidean distance between points p and q.

3.2 Degree of Match

The following measures measure a degree of matching between the point sets by looking at the maximum mismatch between points in one set and points in

another. The motivation for these measures is straightforward — if point sets are being used to represent the drawings, then classical measures of point set similarity can be used to compare the drawings.

Undirected Hausdorff Distance The *undirected Hausdorff distance* is a standard metric for determining the quality of the match between two point sets. It does not take into account the fact that the point sets may be labelled.

$$\text{haus}(P, P') = \frac{1}{\text{UB}} \max \{ \max_{p \in P} \min_{q' \in P'} d(p, q') \,, \max_{p' \in P'} \min_{q \in P} d(p', q) \}$$

UB is the maximum distance between a corner of the bounding box of P and a corner of the bounding box of P'.

Maximum Distance The *maximum distance* is an adaptation of the undirected Hausdorff distance for labelled point sets, and is defined as the maximum distance between two corresponding points:

$$\text{maxdist}(P, P') = \frac{1}{\text{UB}} \max_{p \in P} d(p, p')$$

UB is the maximum distance between a corner of the bounding box of P and a corner of the bounding box of P'.

3.3 Position

These measures are motivated by the idea that the location of the points on the page is important, and points should not move too move far between drawings.

Average Distance *Average distance* is the average distance points move between drawings.

$$\text{dist}(P, P') = \frac{1}{|P|} \sum_{p \in P} d(p, p')$$

Nearest Neighbor Between *Nearest neighbor between* is based on the assumption that a point's original location should be closer to its new position than any other point's new position.

$$\text{nnb}(P, P') = \frac{1}{\text{UB}} \sum_{p \in P} \text{weight}(\text{nearer}(p))$$

where

$$\text{nearer}(p) = \{ q \mid d(p, q') < d(p, p') \,,\, q \in P \,,\, q \neq p \}$$

Unweighted In the unweighted version, the score for p counts only whether or not there are points in P' between p and p'.

$$\begin{aligned} \text{weight}(S) &= \begin{cases} 0 & \text{if } |S| = 0 \\ 1 & \text{otherwise} \end{cases} \\ \text{UB} &= |P| \end{aligned}$$

Weighted In the weighted version, the number of points in P' between p and p' is taken into account.

$$\begin{aligned} \text{weight}(S) &= |S| \\ \text{UB} &= |P|\,(|P|-1) \end{aligned}$$

3.4 Relative Position

These measures are based on the idea that the relative position of points should remain the same. There are two components to relative position — the distance between the points, and the orientation. All of the measures except for average relative distance are concerned with changes in orientation.

Orthogonal Ordering *Orthogonal ordering* measures the change in orientation between pairs of points. Imagine compass roses centered on p and p', with north oriented towards the top of the page. Let θ_q and $\theta_{q'}$ be the directions associated with q and q', respectively.

$$\text{order}(P, P') = \frac{1}{W} \sum_{p,q \in P} \min \left\{ \int_{\theta_q}^{\theta q'} \text{weight}(\theta)\, d\theta, \int_{\theta_{q'}}^{\theta q} \text{weight}(\theta)\, d\theta \right\}$$

Constant-Weighted In the constant-weighted version, all changes of direction are weighted equally.

$$\begin{aligned} \text{weight}(\theta) &= 1 \\ W &= \pi \end{aligned}$$

Linear-Weighted In the linear-weighted version, changes in the north, south, east, west relationships between points are weighted more heavily than changes in direction which do not affect this relationship. The weight function grows linearly with the distance between θ and north, south, east, or west.

$$\begin{aligned} \text{weight}(\theta) &= \begin{cases} \frac{(\theta \bmod \pi/2)}{\pi/4} & \text{if } (\theta \bmod \pi/2) < \pi/4 \\ \frac{\pi/2 - (\theta \bmod \pi/2)}{\pi/4} & \text{otherwise} \end{cases} \\ W &= \pi/2 \end{aligned}$$

Ranking The *ranking* measure considers the relative horizontal and vertical position of the point. This is a component of the similarity measure used in the SMILE graph multidrawing system. [1] Let right(p) and above(p) be the number of points to the right of and above p, respectively.

$$\text{rank}(P, P') = \frac{1}{\text{UB}} \sum_{p \in P} \min\{\, |\, \text{right}(p) - \text{right}(p')\, | + |\, \text{above}(p) - \text{above}(p')\, |\,,\ \text{UB}\,\}$$

where

$$\text{UB} = 1.5\,(|P| - 1)$$

Of note here is that the upper bound is taken as $1.5\,(|P|-1)$ instead of $2\,(|P|-1)$, the actual maximum value occurring when a point moves from one corner of the drawing to the opposite corner. The motivation for this is simply that it scales the measure more satisfactorily.

Average Relative Distance [order only] The *average relative distance* is the average change in distance between pairs of points.

$$\text{rdist}(P, P') = \frac{1}{|P|\,(|P| - 1)} \sum_{p,q \in P} |\, d(p, q) - d(p', q')\,|$$

λ-Matrix [order only] The λ-matrix model is used by Lyons, Meijer, and Rappaport [10] to evaluate cluster-busting algorithms. It is based on the concept of order type used by Goodman and Pollack [9], where two sets of points P and P' have the same order type if, for every triple of points (p,q,r), they are oriented counterclockwise if and only if (p',q',r') are also oriented counterclockwise.

Let $\lambda(p, q)$ be the number of points in P to the left of the directed line from p to q.

$$\text{lambda}(P, P') = \frac{1}{\text{UB}} \sum_{p,q \in P} |\, \lambda(p, q) - \lambda(p', q')\,|$$

where the upper bound for a set of size n is:

$$\text{UB} = n \left\lfloor \frac{(n-1)^2}{2} \right\rfloor$$

3.5 Neighborhood

These measures are guided by the philosophy that each point's neighborhood should be the same in both drawings. The measures do not explicitly take into account the point's absolute position, and considers its position relative to other points only in the sense of keeping nearby points together.

Nearest Neighbor Within [order only] For *nearest neighbor within*, a point's neighborhood is its nearest neighbor. Let $\mathrm{nn}(p)$ be the nearest neighbor of p in the p's point set and $\mathrm{nn}(p)'$ be the corresponding point in P' to $\mathrm{nn}(p)$. Ideally, $\mathrm{nn}(p)'$ should be p''s nearest neighbor.

$$\mathrm{nnw}(P, P') = \frac{1}{\mathrm{UB}} \sum_{p \in P} \mathrm{weight}(\mathrm{nearer}(p))$$

where

$$\mathrm{nearer}(p) = \{\, q \mid d(p', q') < d(p', \mathrm{nn}(p)')\,,\, q \in P\,,\, q \neq p\,,\, q \neq \mathrm{nn}(p)\,\}$$

Unweighted The unweighted version considers only whether or not $\mathrm{nn}(p)'$ is p''s nearest neighbor.

$$\begin{aligned} \mathrm{weight}(S) &= \begin{cases} 0 & \text{if } |S| = 0 \\ 1 & \text{otherwise} \end{cases} \\ \mathrm{UB} &= |P| \end{aligned}$$

Weighted The weighted version takes into account the number of points in P' closer to p' than $\mathrm{nn}(p)'$.

$$\begin{aligned} \mathrm{weight}(S) &= |S| \\ \mathrm{UB} &= |P|\,(|P| - 1) \end{aligned}$$

ϵ-Clustering [order only] *ϵ-clustering* defines the neighborhood for each point to be its ϵ-cluster, the set of points within a distance ϵ, defined as the maximum distance between a point and its nearest neighbor. The ϵ-cluster for each point is guaranteed to contain at least one other point. The measure considers the ratio of the number of points in p's ϵ-cluster in both drawings to the number of points in the ϵ-cluster in at least one of the drawings; ideally, this ratio would be 1 because the same points would be in both clusters.

$$\mathrm{eclus} = 1 - \frac{|S_I|}{|S_U|}$$

where

$$\begin{aligned} \epsilon &= \max_{p \in P} \min_{q \in P, q \neq p} d(p, q) \\ S_I &= \{\, (p, q) \mid p \in P\,,\, q \in \mathrm{clus}(p, P, \epsilon) \text{ and } q' \in \mathrm{clus}(p', P', \epsilon')\,\} \\ S_U &= \{\, (p, q) \mid p \in P\,,\, q \in \mathrm{clus}(p, P, \epsilon) \text{ or } q' \in \mathrm{clus}(p', P', \epsilon')\,\} \\ \mathrm{clus}(p, P, \epsilon) &= \{\, q \mid d(p, q) \leq \epsilon\,,\, q \in P\,,\, q \neq p\,\} \end{aligned}$$

Separation-Based Clustering [order only] In the *separation-based clustering* measure, points are grouped so that each point in a cluster is within some distance δ of another point in the cluster and at least distance δ from any point not in the cluster. The intuition is that the eye naturally groups things based on the surrounding whitespace.

Formally, for every point p in cluster C such that $|C| > 1$, there is a point $q \neq p \in C$ such that $d(p,q) < \delta$, and $d(p,r) > \delta$ for all points $r \notin C$. If C is a single point, only the second condition holds. Let clus(p) be the cluster to which point p belongs.

$$\text{sclus} = 1 - \frac{|S_I|}{|S_U|}$$

where

$$\begin{aligned} S_I &= \{\, (p,q) \mid p,q \in P\,,\ \text{clus}(p) = \text{clus}(q) \text{ and } \text{clus}(p') = \text{clus}(q') \,\} \\ S_U &= \{\, (p,q) \mid p,q \in P\,,\ \text{clus}(p) = \text{clus}(q) \text{ or } \text{clus}(p') = \text{clus}(q') \,\} \end{aligned}$$

3.6 Edges

Shape The *shape* measure treats the edges of the graph as sequences of north, south, east, and west segments and compares these sequences using the edit distance.

$$\text{shape} = \frac{1}{\text{UB}} \sum_{e \in E} \text{edits(e,e')}$$

Regular The edit distance is not normalized for the length of the sequence, and the upper bound is as follows:

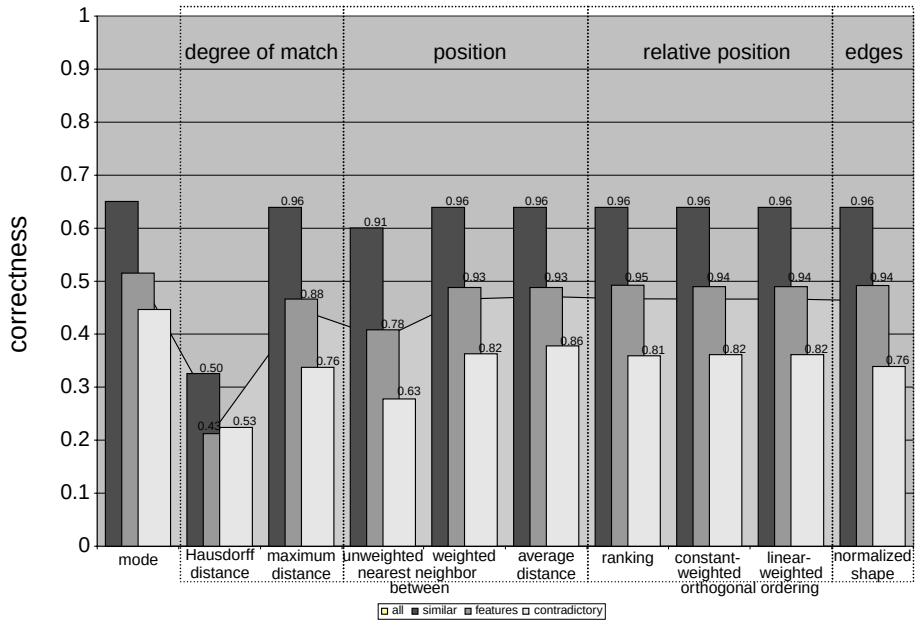
$$\text{UB} = \sum_{e \in E} |\,\text{length}(e) - \text{length}(e')\,| + \min\{\,\text{length}(e)\,,\,\text{length}(e')\,\}$$

Normalized The edit distance is normalized for the length of the sequence using the algorithm of Marzal and Vidal [11], and the upper bound is as follows:

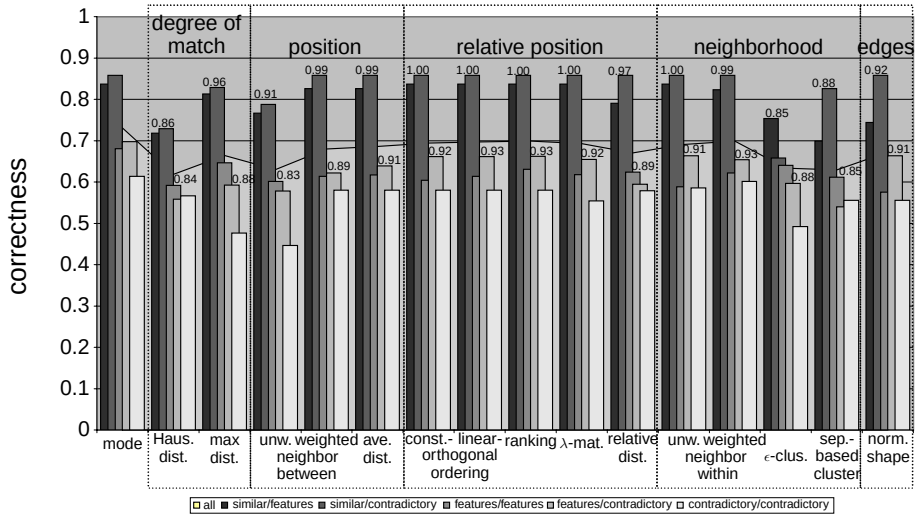
$$\text{UB} = |E|$$

4 Results: Rotation and Ordering

Figures 4(a) and 4(b) show the results for the rotation and ordering tasks, respectively. The measures are grouped according to the categorization of section 3. These results are explained and discussed in the following sections.



(a) Rotation task.



(b) Ordering task.

Figure 4: Average measure correctness for the rotation and ordering tasks. Numbers by columns show average relative correctness for that group; for (b), relative correctness is shown only for the "similar" and "not" trial groups.

4.1 Correctness

The success or failure of the candidate similarity measures is determined by how well they match observed user responses. A measure's "correctness" is the fraction of the time the measure correctly predicted user responses on the tasks.

For the rotation task, let an individual trial T_{rot} be described by the tuple $(D, D_1, \ldots, D_8)$, where D is the base drawing and the D_i are the eight orientations the user must choose between. There are only two drawings that the user must choose between for the ordering task, so an ordering trial T_{order} can be described by the tuple (D, D_1, D_2).

Let T_M be the measure's choice for a rotation or ordering trial T:

$$T_M = \begin{cases} D_k & \text{if } M(D, D_k) < M(D, D_i)\ \forall i \neq k \\ \text{tie} & \text{if } \exists\ j, k \text{ such that } M(D, D_j) = M(D, D_k) \leq M(D, D_i) \\ & \quad \forall i \neq j, k\ (j \neq k) \end{cases}$$

Also, let T_k denote user k's response for trial T:

$$T_k = \begin{cases} D_k & \text{if the user chose drawing } D_k \\ \text{tie} & \text{if the user clicked the "can't decide" button} \end{cases}$$

Note that T_k is only defined if user k was presented with trial T — each trial was completed by about one-third of the users.

Define the correctness of the measure M with respect to user k for T as $C(M, T, k)$:

$$C(M, T, k) = \begin{cases} 1 & \text{if } T_M = T_k \text{ or if } T_k = \text{tie} \\ 0 & \text{otherwise} \end{cases}$$

$C(M, T, k)$ is undefined if T_k is undefined. Any T_M is considered to be correct for those trials where the user's response is "can't decide" because it is assumed that if the user has no preference as to the correct response, she will be happy with any of the choices.

Let $\mathcal{K}$ be the set of users k for which T_k is defined for a particular trial T. Then the correctness $C_{M,T}$ of measure M for a trial T is

$$C_{M,T} = \frac{\sum_{k \in \mathcal{K}} C(M, T, k)}{|\mathcal{K}|}$$

The solid shaded areas behind the columns in Figures 4(a) and 4(b) show the average correctness for each measure over all rotation and all ordering trials, respectively. Only the results for the full point set (for point-based measures) are shown in Figures 4(a) and 4(b); the results for the borders-only point sets are discussed in section 4.4. Also, only results for the normalized shape measure are included because there is little difference in the performance of the two shape measures ($p = .95$). Unless otherwise specified, the performance of different measures over a set of trials is compared using Student's t-test to determine the probability that the measures have the same average correctness for those trials.

Observations For the rotation task, the average correctness over all trials for even the best measures is disappointingly low — below 50%! — meaning that even the best of the tested measures will tend to rotate drawings incorrectly much of the time.

For both rotation and ordering, a striking result is that no one measure stands out as being significantly better than the others. There is little difference between the angle-sensitive relative position measures (constant- and linear-weighted orthogonal ordering, ranking, λ-matrix) and the nearest neighbor within measures; distance, weighted nearest neighbor between, and shape also have similar good performance on the rotation task. On the other hand, several measures stand out as being noticeably worse than the others. Hausdorff distance, the clustering neighborhood measures (ϵ-clustering and separation-based clustering), unweighted nearest neighbor between, and to a lesser degree maximum distance have the poorest performance over all trials. Figures 5 and 6 summarize the significant similarities and differences in measure performance.

4.2 Correctness By Drawing Category

Some trials were easier than others — one might expect that in the rotation task it might be easier to choose the "correct" rotation when the new drawing is very similar to the old, and that in the ordering task it might be easier to pick the most similar drawing of two when one is very similar to the base. As a result, the measures under consideration may perform better or worse in these circumstances.

To evaluate this, the new drawings were separated into categories: *similar* for drawings very close to the corresponding base, *features* for drawings somewhat different from the base but with noticeable recognizable features to help identification, *contradictory* for drawings with recognizable features but where those features contradicted each other (for example, when one feature was rotated with respect to another), and *different* for drawings that are very different from the base.

The trials were also grouped according to the type of drawing(s) involved: rotation trials were considered to be "similar", "features", or "contradictory" according to the category of the new drawing used in the trial, and ordering trials were considered to be "similar" or "not" depending on whether or not one of the new drawings in the trial was classified as "similar". Since only two drawings were classified as "different", the results from trials involving these drawings were included only in the "all trials" results and not in the results from any other subgroup of trials.

The multiple columns for each measure in Figures 4(a) and 4(b) show the average correctness for each category of trials. Figure 4(b) breaks down the trials beyond the two "similar" and "not" groupings, but the discussion is only in terms of the two groups. There were no ordering trials in which both drawings were similar, so that combination is not shown.

The categorization used is based on human judgment and the assumption that features play a key role in similarity. The latter assumption is supported

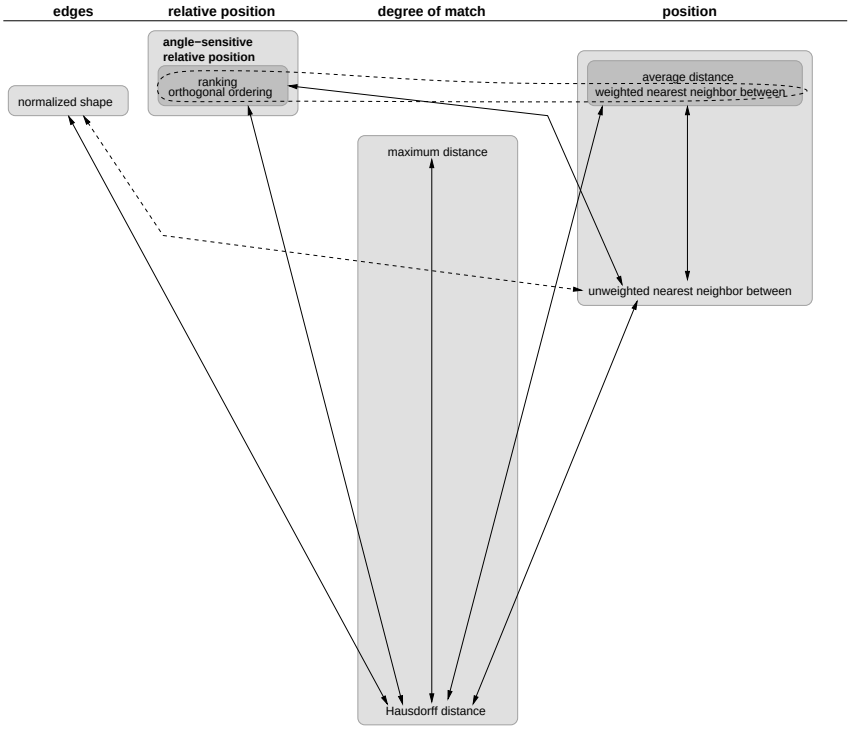


Figure 5: Similarities and differences in measure performance for the rotation task. The vertical position of the measure name indicates its average correctness over all trials, with the lowest at the bottom. Gray-shaded areas group related measures. Measures with similar performance are grouped by ovals; arrows indicate differences between measures or groups of measures. The line style indicates the significance level: solid for $p = .01/p = .99$, dashed for $p = .05/p = .95$.

by the feedback portion of the study (section 5.2). It is intended that the ordering "similar", "features", and "contradictory" represent groups of new drawings that are increasingly different from the base drawing to which they are compared. This goal seems to have been met: in the ordering task, users picked the drawing from the most similar group in 80% of the trials involving drawings from different groups, and the most common user response was the drawing from the most similar group in 90% of the trials. Also, using Student's t-test to compare the distances between base and new drawings in each group shows that, for all of the similarity measures under consideration except Hausdorff distance, the means of the groups are significantly different and follow the expected pattern ($p = .01$). Since most of the measures being tested seem to capture at least some of the users' ideas of similarity, this lends support to the categorization.

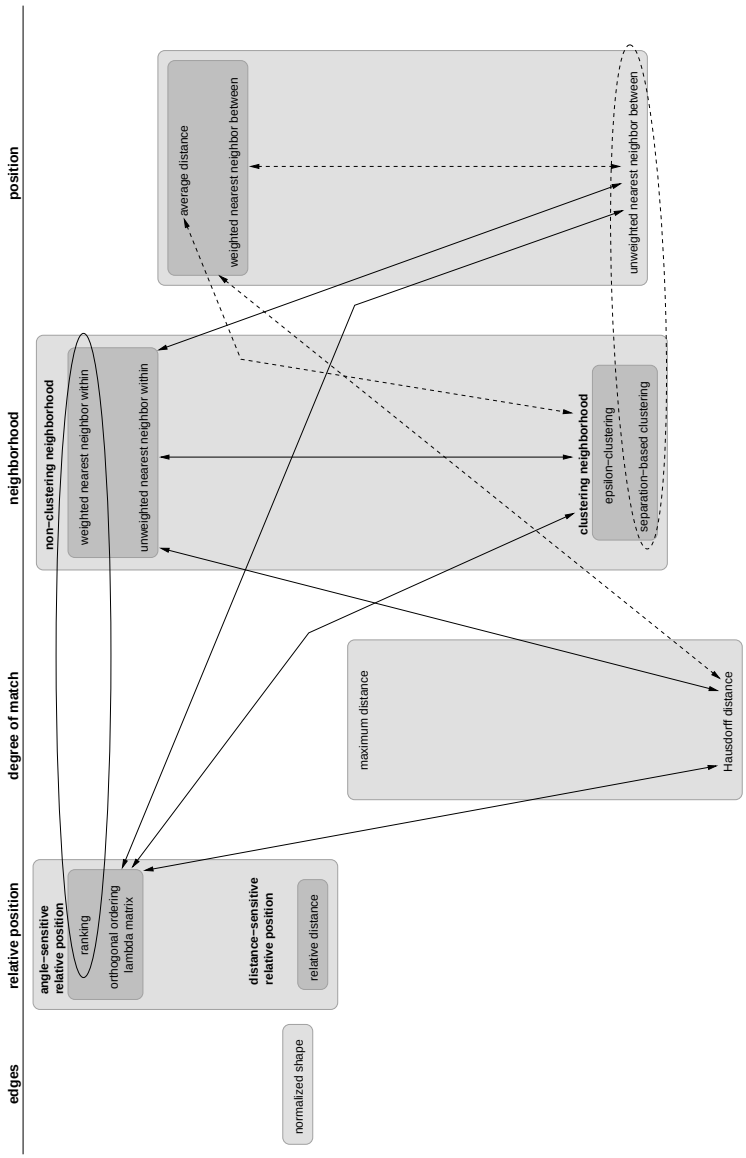


Figure 6: Similarities and differences in measure performance for the ordering task.

Observations Within measures, the expected drop in performance between "similar" and "features" trials and between "features" and "contradictory" trials for rotation, and between "similar" and "not" trials for ordering is significant for nearly every measure ($p = .01$). The only exceptions are for Hausdorff distance in the rotation task (differences are not significant, and average correctness is

higher for "contradictory" than for "features") and ϵ-clustering for the ordering task (drop in performance is significant only at $p = .05$).

4.3 Mode Correctness

The "correct" answer for any given trial is based on a user's opinion, and because approximately 34 students gave responses for each set of drawings, there was the potential for getting different "correct" user responses for the same set of drawings. Since each measure always chooses the same response for the same set of drawings, the best correctness score a measure could receive is if it always makes the choice that was most common among the users.

Let $f(T, r)$ be the frequency with which the users completing trial T picked response r from the choices D_i. Also define the most common response T_{mode} as the response for which $f(T, r)$ is maximized, and the correctness of the most common response for trial T and user k as

$$C(\text{mode}, T, k) = \begin{cases} 1 & \text{if } T_{\text{mode}} = T_k \text{ or } T_k = \text{tie} \\ 0 & \text{otherwise} \end{cases}$$

Then the "best possible score" for a measure for trial T is:

$$C_{\text{best},T} = \frac{\sum_{k \in \mathcal{K}} C(\text{mode}, T, k)}{|\mathcal{K}|}$$

where $\mathcal{K}$ is the set of users completing trial T. The "mode" columns of Figures 4(a) and 4(b) show the average value of C_{best} over the various groups of trials.

The *relative correctness* for a given measure and trial is the ratio of the measure's correctness to the mode correctness for that trial. The numbers listed by the measure columns in Figure 4(a) give the average relative correctness for each group of trials for which the average correctness is given; the numbers shown in Figure 4(b) give the average relative correctness for the "similar" and "not" trial groups.

Observations The first observation is that, as one might expect, the average mode correctness drops for the less similar trials. The differences in average mode correctness between "similar", "features", and "contradictory" trials in the rotation task and between "similar" and "not" trials in the ordering task are significant ($p = .01$). The drop in correctness between the more similar and more different trials is due to an increase in user disagreement over the correct answer for each trial. Table 1 summarizes the degree to which users preferred some drawings to others for different groups of trials.

The mode correctness is also affected by the number of users choosing "can't decide" for each trial — a large number of "can't decide" answers will raise the value of the mode correctness. For rotation, the percentage of "can't decide" answers grew from 1.1% for "similar" drawings to 3.9% for "features" drawings

rotation	all trials	similar	features	contradictory
preference	92%	100%	94%	83%
no preference	-	-	-	-

ordering	all trials	similar	not
preference	70%	94%	61%
no preference	1.9%	-	2.6%

Table 1: Percentage of trials for which users preferred one or more choices over the others, and for which there was no preference. Preference was determined using the χ^2 test to compare the distribution of frequencies for each possible choice in a trial with a uniform distribution (the expectation if there was no preference among the choices in the trial). "Preference" indicates a significant difference between the distribution of user responses and the uniform distribution ($p = .05$); "no preference" indicates no difference between the distribution of user responses and the uniform distribution ($p = .95$).

and 6.7% for "contradictory" drawings. For ordering, the percentage grew from 2.7% for trials involving "similar" drawings to 6.1% for trials not involving "similar" drawings. The increase in "can't decide" answers partially offsets the drop in mode correctness due to increased user disagreement, but did not cancel out the effect.

A second observation concerns the relative correctness. While the average correctness is low for all measures, the relative correctness results are better. For both rotation and ordering, the average correctness is at least 90% of the average mode correctness for most measures. (Exceptions are Hausdorff distance, unweighted nearest neighbor between, and the clustering neighborhood measures (ϵ-clustering and separation-based clustering).) This indicates that low correctness values for these measures, especially in the rotation task, are primarily due to disagreements between users about the correct answer rather than major failings on the part of the measures.

An interesting note is that the average relative correctness of most measures drops with the more difficult trials — the lower performance for these trials is not due only to increased user disagreement about the correct answer. Table 2 lists the measures for which the drop is significant. The drop in relative correctness, combined with the presence of some user preference for many of even the most different trials, suggests that there are more subtle similarities that users are picking up on that the measures do not capture. The good relative performance for "similar" trials is the result of the drawings in those trials being similar in many ways — including those ways captured by the measures — while the poorer relative performance for more different trials is the result of the measures' missing something that the users are seeing.

rotation	ordering
maximum distance	maximum distance
unweighted nearest neighbor between	weighted nearest neighbor between
ranking	average distance
normalized shape	constant-weighted orthogonal ordering
	linear-weighted orthogonal ordering
	ranking
	λ-matrix
	relative distance
	unweighted nearest neighbor within
	weighted nearest neighbor within

Table 2: Measures for which the average relative correctness is significantly lower between "similar" or "features" trials and "contradictory" trials (rotation) or between "similar" and "not" trials (ordering); $p = .05$.

4.4 Border vs. Full Point Sets

Based on user feedback (see section 5.2), a second point set using only those points near the borders of the drawing was tested for point-based measures. There was no significant difference in performance between the two point sets for any measure ($p = .95$), and so results for the borders-only point set are not included in Figures 4(a) and 4(b).

4.5 Per-User Correctness

The low correctness for even the mode (65% for the most similar drawings) in the rotation task indicates that not all users agree on what the "correct" answer is for a given trial. This suggests that users have different ideas about what factors make drawings look more similar, or different ideas about the relative importance of different aspects of similarity.

Let $\mathcal{T}_k$ be the set of trials for which T_k is defined, i.e., the set of trials user k completed. Then the correctness of measure M for user k is

$$C_{M,k} = \frac{\sum_{T \in \mathcal{T}_k} C(M, T, k)}{|\mathcal{T}_k|}$$

Figure 7 shows $C_{M,k}$ for each measure/user combination in the rotation task.

Observations The most striking observations from Figure 7 are that the measures that do badly overall (Hausdorff distance and unweighted nearest neighbor between) perform badly for every user, and the measures that perform well overall tend to all perform well or all perform badly for an individual user. The results are similar, but less dramatic, for the ordering task.

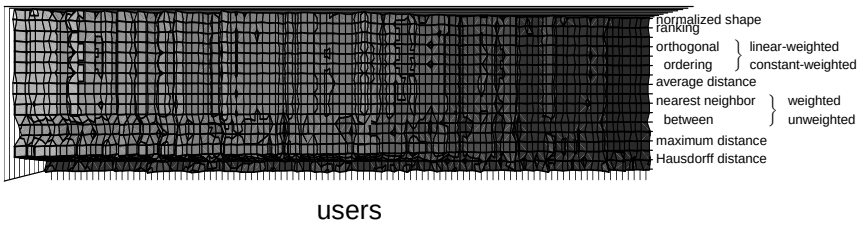


Figure 7: Average correctness per user for the rotation task. Colors indicate the average correctness, from light gray (1) to black (0). Each measure label (except normalized shape) spans two rows, where the top row shows the borders-only point set and the bottom row the full point set.

rotation	all trials	similar	features	contradictory
borders-only better	5.8%	-	-	7.1%
full better	2.0%	-	-	-
no difference	23%	100%	54%	39%

ordering	all trials	similar	not
borders-only better	6.9%	4.9%	6.9%
full better	8.8%	8.8%	9.8%
no difference	1.0%	40%	2.9%

Table 3: Percentage of users which are significantly better predicted by one of the point sets used ($p = .05$), and the percentage for which there is no difference ($p = .95$).

Of interest, though, is that there is a small group of users who are significantly better- or worse-predicted by the borders-only point set than the full point set. The $C_{M,k}$ values for a user k for selected measures were compared for the full and borders-only point set using Student's t-test. The measures used were for the better measures: all of the point-based measures except Hausdorff distance, unweighted nearest neighbor between, and the clustering neighborhood measures (ϵ-clustering and separation-based clustering). Table 3 summarizes the results.

The cases for which the borders-only point set performs better meshes well with the feedback from the study (section 5.2), though many more users commented on the importance of the borders than had a significant improvement for that point set. One possibility for this is that the change in the border may be representative of the change in the drawing as a whole, so that full point set performs well for users even if they are focusing primarily on the borders of the drawing. For the rotation task, another possibility is that many of the measures

are already more sensitive to changes in the border regions than in the centers of the drawings because rotation moves points near the borders farther than those near the center. This means that the full point set versions will tend to perform similarly to the border-only sets for rotation.

Another interesting note is that a larger number of users were better-predicted by the border point set for more different trials than for the similar trials. This suggests that the borders become important when the obvious similarities disappear — when the drawings are similar, users can easily select their responses based on the overall look of the drawing, but when the drawings become more different, they begin focusing on the borders to look for distinguishing characteristics.

Focusing on cases where the borders-only point set outperforms the full point set ignores those instances where the reverse is true. In fact, more users are predicted better by the full point set than by the borders-only point set for the ordering task — a pattern which is not seen in the results from the rotation task. This may be the result of users looking for different things in each task — the overall look may be more important in the ordering task, while users gravitate towards the borders when looking for hints as to the proper orientation in the rotation task. The border-sensitivity of many measures under rotation may also explain this difference — if both full and border-only point sets are sensitive to the same things, they will tend to perform in the same way and thus the full point set will not perform better.

4.6 Other Rotation Angles

The rotation task focused on differences of $\pi/2$ in the rotation angle — a very large difference, though it is the only meaningful difference for orthogonal drawings since a user can easily tell that a rotation by some other angle is not the best match. Ideally, there would be no change in a measure's correctness if additional rotation angles are considered, since the user's "correct" answer would not change.

Figure 8 shows the average correctness over all trials for three sets of orientations: the $\pi/2$ multiples (with and without an initial flip around the x-axis) discussed so far (labelled "$\pi/2$"), the $\pi/2$ multiples augmented by the four additional multiples of $\pi/4$ (labelled "$\pi/4$"), and the $\pi/2$ multiples augmented by the four additional rotations $\pi/36$, $\pi/2 + \pi/36$, $\pi + \pi/36$, and $3\pi/2 + \pi/36$ (labelled "$\pi/36$ offsets"). (The shape measure is not included in the figure because it is defined only in terms of orthogonal drawings.) The $\pi/4$ set adds several additional orientations, though the angle between successive rotations is still relatively large; the $\pi/36$ offsets add rotation angles which are very close to angles a user would consider. The additional rotations were not presented to the users, as it was assumed that no one would pick one of them.

Observations Clearly the ideal outcome did not happen — over all trials, there is a drop in the average correctness with the addition of the extra orienta-

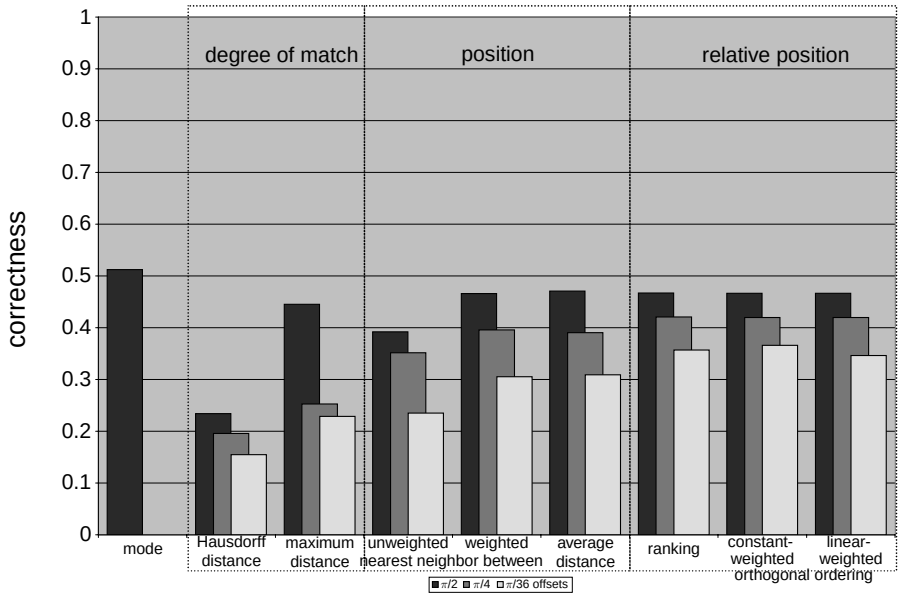


Figure 8: Measure correctness for rotation with other sets of orientations.

tions and the drop is greater for the $\pi/36$ offsets group than for the $\pi/4$ group. Table 4 summarizes the significant results.

If the correctness criterion is relaxed so that measures are only expected to choose a rotation near the correct one, the drop in performance is eliminated. The expanded definition of "correct" is as follows: $C_{\text{near}}(M, T, k) = 1$ if T_k is "tie" or if the rotation angle of T_M is within $\pm\pi/4$ of the rotation angle of T_k for "$\pi/4$" and within $\pm\pi/36$ for "$\pi/36$ offsets." The improvement in the results indicates that when the wrong rotation is chosen, it tends to be near the right one. This suggests that while the measures would likely perform less satisfactorily when asked to pick the correct rotation in non-orthogonal applications, they would still perform reasonably well if the goal is only to obtain approximately the right orientation.

5 Other Results

5.1 Difference Task

The goal in the difference part was to be able to use the user's response times as an indicator of similarity, the idea being that a user can locate the new vertex faster if the drawings are more similar. As a test of the validity of this, the times on the difference part were used to order the pairs of drawings used in the ordering task. The results were very unsatisfactory, achieving only 45% correctness on average (compare to Figure 4(b), where even the worst of the

$\pi/2$ **vs.** $\pi/4$	all trials	similar	features	contradictory
Hausdorff distance				
maximum distance	>>		>>	>>
nearest neighbor between				
unweighted		=		
weighted	>	==		>>
average distance	>	==		>>
orthogonal ordering				
constant-weighted		==		>
linear-weighted		==		>
ranking		==		>

$\pi/2$ **vs.** $\pi/36$ **offsets**	all trials	similar	features	contradictory
Hausdorff distance	>>		>	
maximum distance	>>	>>	>>	>>
nearest neighbor between				
unweighted	>>	>	>>	>
weighted	>>	>	>>	>>
average distance	>>		>>	>>
orthogonal ordering				
constant-weighted	>>		>>	>
linear-weighted	>>		>>	>
ranking	>>		>>	>>

Table 4: Comparing the average performance for each group of trials for different sets of orientations. ">" and ">>" indicate a drop in performance with additional drawings; "=" and "==" indicate that the performance did not change. The number of symbols indicates the significance level: ">" and "=" for $p = .05/p = .95$, ">>" and "==" for $p = .01/p = .99$.

measures under consideration reached 62% correctness). As a result, the times on the difference task are not a good indicator of similarity and are not suitable for evaluating measures with respect to the magnitude criterion.

5.2 User Feedback

The students' responses to the final questionnaire yielded several interesting notes. As might be expected, the responses as to what makes two drawings look similar in the rotation and ordering parts included a sizable percentage (35%) who said preserving the position, size, number of large vertices was important and another large percentage (44%) who said they looked for distinctive clusters and patterns of vertices, such as chains, zigzags, and degree 1 vertices. More surprising was that 44% of the students said that borders and corners of the drawing are more important than the interior when looking for similarity. This is supported by research in cognitive science indicating that people often treat

filled and outline shapes as equivalent, focusing primarily on the external contour (Wickelgren [18]). A number of these students mentioned the importance of "twiddly bits around the edges" — distinctive clusters and arrangements of vertices, made more obvious by being on the border. Related comments were also that the orientation and aspect ratio of the bounding box should remain the same, and that the outline of the drawing should not change. Another sizable group (34%) commented that the "general shape" of the drawing is important.

For the question about the difference part, several users expressed frustration at the difficulty of the task. The usefulness of the "big picture" view — looking at the overall shape of the drawing — was contested, with nearly equal numbers reporting that the overall look was useful in the task, and that it was confusing and misleading. About 16% of the users mentioned limited use of the overall look, using it on a region-by-region basis to quickly eliminate blocks that remained the same and falling back on simply scanning the drawing or matching corresponding vertices and tracing edges when the regions were too different. Another 24% reported using vertex-by-vertex matching from the beginning. A similar-sized group (20%) figured out shortcuts, such as that the edges added along with the new vertex frequently caused one of the neighboring vertices to have a degree larger than 4 and thus be drawn with a larger box, so they scanned for the neighbors of the large boxes to find the new vertex. Overall, just over a quarter of the users (28%) reported searching for vertices with extra edges rather than searching for new vertex directly.

For the final question, about what a graph drawing algorithm should take into account if the look of the drawing should be preserved, the most common answers echoed those from the rotation/ordering question: maintaining vertex size and shape, the relative positions of vertices, the outline of the drawing, and clusters.

6 Conclusions and Future Work

Table 5 summarizes the best and worst of the measures evaluated.

As groups, the angle-sensitive relative position measures (orthogonal ordering, ranking, and λ-matrix) and the non-clustering neighborhood measures (weighted and unweighted nearest neighbor within) performed significantly better than the degree-of-match measures (Hausdorff distance and maximum distance) and the clustering neighborhood measures (ϵ-clustering and separation-based clustering). As a result, the orthogonal ordering, ranking, λ-matrix, and nearest neighbor within measures are given the highest ranking.

Two of the three position measures (weighted nearest neighbor between and average distance) also perform significantly better than the degree-of-match and clustering neighborhood measures, and so are also ranked well. They are given the second-level ranking because there was a slightly greater dropoff in their performance when additional rotation angles were introduced, suggesting that they might not perform quite as well for non-orthogonal drawings.

Shape and relative distance are also given an above-average ranking be-

best measures	
angle-sensitive relative position	ranking
	orthogonal ordering (constant- and linear-weighted)
	λ-matrix
non-clustering neighborhood	nearest neighbor within (weighted and unweighted)
position	average distance
	weighted nearest neighbor between
edges	shape
distance-sensitive relative position	relative distance
middle-of-the-road measures	
degree of match	maximum distance
worst measures	
clustering neighborhood	ϵ-clustering
	separation-based clustering
position	unweighted nearest neighbor between
degree of match	Hausdorff distance

Table 5: Overall ranking of the measures evaluated. Measures are listed in groups from best to worst; groups are separated by horizontal lines. There is little difference between measures in a single group.

cause of their performance on the ordering task — while not significantly worse than the top-ranked measures, they were also not significantly better than the bottom-ranked measures. It should be noted that shape was among the best measures for the rotation task.

Maximum distance received a middle-of-the-road ranking because of its middle-of-the-road performance in both tasks — while it performed as well the several top-ranked measures for the "similar" rotation trials, its average correctness was between the top- and bottom-ranked measures for both rotation and ordering. Also, when the degree-of-match measures were considered as a group, their performance was significantly worse than the top-ranked groups.

The clustering neighborhood measures and the unweighted nearest neighbor between measure all performed significantly worse than the top-ranked groups of measures, and so are given a low ranking. Hausdorff distance is given the bottom ranking because it performs significantly worse than most other measures in most groups of trials.

Several other conclusions can be drawn from this study:

- The difference in relative correctness between the more similar and more different trials for both rotation and ordering task suggests that there are

more subtle notions of similarity which are being missed by the measures tested, and which, if incorporated, would improve their performance.

- The per-user analysis suggests that while it is meaningful to talk about "good" measures and "bad" measures in overall terms, to get the maximum performance it may be necessary to tailor the specific similarity measure used to a particular user.

- The difficulty of the difference part suggests that the amount of difference between the drawings that is considered reasonable varies greatly with the task — when the user simply needs to recognize the graph as familiar, the perimeter of the drawing and the position and shape of few key features are the most important. On the other hand, when trying to find a specific small change, the drawings need to look very much alike or else the user needs some other cues (change in color, more distinctive vertex names, etc.) in order to highlight the change.

The students' responses on the questionnaire suggest several possible directions for future investigation.

- The number of students who mentioned focusing on drawing borders was surprising, and additional study is needed to further investigate the importance of borders.

- Large vertices are identified as being especially important, which could lead to a scheme in which changes in the position and size of large vertices are weighted more heavily than other vertices.

- Another major focus was clusters of vertices — both the presence of clusters in general, and the presence of specific shapes such as chains and zigzags. The relatively poor showing of the clustering measures indicates that they are not making use of clusters in the right way. The fact that the students reported looking for specific shapes suggests an approach related to the drawing algorithms of Dengler, Friedell, and Marks [6] and Ryall, Marks, and Shieber [16]. These algorithms try to produce drawings which employ effective perceptual organization by identifying Visual Organization Features (VOFs) used by human graphic designers. VOFs include horizontal and vertical alignment of vertices, particular shapes such as "T" shapes, and symmetrically placed groups of vertices. VOFs can also be used not to guide the creation of drawings from scratch, but to identify features in an existing drawing that may be important because they adhere to a particular design principle. This is related to the work of Dengler and Cowan [5] on semantic attributes that humans attach to drawings based on the layout — for example, symmetrically placed nodes are interpreted as having common properties. A similarity measure could then measure how well those structures are preserved, and an interactive graph drawing algorithm could focus on preserving the structures.

References

[1] T. Biedl, J. Marks, K. Ryall, and S. Whitesides. Graph multidrawing: Finding nice drawings without defining nice. In S. Whitesides, editor, *Graph Drawing (Proc. GD '98)*, volume 1547 of *Lecture Notes Comput. Sci.*, pages 347–355. Springer-Verlag, 1998.

[2] T. C. Biedl and M. Kaufmann. Area-efficient static and incremental graph drawings. In R. Burkard and G. Woeginger, editors, *Algorithms (Proc. ESA '97)*, volume 1284 of *Lecture Notes Comput. Sci.*, pages 37–52. Springer-Verlag, 1997.

[3] S. Bridgeman and R. Tamassia. Difference metrics for interactive orthogonal graph drawing algorithms. *Journal of Graph Algorithms and Applications*, 4(3):47–74, 2000.

[4] S. S. Bridgeman, J. Fanto, A. Garg, R. Tamassia, and L. Vismara. InteractiveGiotto: An algorithm for interactive orthogonal graph drawing. In G. Di Battista, editor, *Graph Drawing (Proc. GD '97)*, volume 1353 of *Lecture Notes Comput. Sci.*, pages 303–308. Springer-Verlag, 1997.

[5] E. Dengler and W. Cowan. Human perception of laid-out graphs. In S. H. Whitesides, editor, *Graph Drawing (Proc. GD '98)*, volume 1547 of *Lecture Notes Comput. Sci.*, pages 441–443. Springer-Verlag, 1998.

[6] E. Dengler, M. Friedell, and J. Marks. Constraint-driven diagram layout. In *Proc. IEEE Sympos. on Visual Languages*, pages 330–335, 1993.

[7] G. Di Battista, A. Garg, G. Liotta, R. Tamassia, E. Tassinari, and F. Vargiu. An experimental comparison of four graph drawing algorithms. *Comput. Geom. Theory Appl.*, 7:303–325, 1997.

[8] U. Fößmeier. Interactive orthogonal graph drawing: Algorithms and bounds. In G. Di Battista, editor, *Graph Drawing (Proc. GD '97)*, volume 1353 of *Lecture Notes Comput. Sci.*, pages 111–123. Springer-Verlag, 1997.

[9] J. E. Goodman and R. Pollack. Multidimensional sorting. *SIAM J. Comput.*, 12(3):484–507, Aug. 1983.

[10] K. A. Lyons, H. Meijer, and D. Rappaport. Algorithms for cluster busting in anchored graph drawing. *J. Graph Algorithms Appl.*, 2(1):1–24, 1998.

[11] A. Marzal and E. Vidal. Computation of normalized edit distance and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):926–932, Sept. 1993.

[12] K. Misue, P. Eades, W. Lai, and K. Sugiyama. Layout adjustment and the mental map. *J. Visual Lang. Comput.*, 6(2):183–210, 1995.

[13] S. North. Incremental layout in DynaDAG. In *Graph Drawing (Proc. GD '95)*, volume 1027 of *Lecture Notes Comput. Sci.*, pages 409–418. Springer-Verlag, 1996.

[14] A. Papakostas, J. M. Six, and I. G. Tollis. Experimental and theoretical results in interactive graph drawing. In S. North, editor, *Graph Drawing (Proc. GD '96)*, volume 1190 of *Lecture Notes Comput. Sci.*, pages 371–386. Springer-Verlag, 1997.

[15] A. Papakostas and I. G. Tollis. Interactive orthogonal graph drawing. *IEEE Trans. Comput.*, C-47(11):1297–1309, 1998.

[16] K. Ryall, J. Marks, and S. Shieber. An interactive system for drawing graphs. In S. North, editor, *Graph Drawing (Proc. GD '96)*, volume 1190 of *Lecture Notes Comput. Sci.*, pages 387–393. Springer-Verlag, 1997.

[17] R. Tamassia, G. Di Battista, and C. Batini. Automatic graph drawing and readability of diagrams. *IEEE Trans. Syst. Man Cybern.*, SMC-18(1):61–79, 1988.

[18] W. A. Wickelgren. *Cognitive Psychology.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1979.