Relative Accuracy Measures for Stroke Gestures

Radu-Daniel Vatavu
University Stefan cel Mare of Suceava
Suceava 720229, Romania
vatavu@eed.usv.ro

Lisa Anthony
UMBC Information Systems
1000 Hilltop Circle
Baltimore, MD 21250 USA
lanthony@umbc.edu

Jacob O. Wobbrock
Information School | DUB Group
University of Washington
Seattle, WA 98195-2840 USA
wobbrock@uw.edu

ABSTRACT

Current measures of stroke gesture articulation lack descriptive power because they only capture *absolute* characteristics about the gesture as a whole, not fine-grained features that reveal subtleties about the gesture articulation path. We present a set of twelve new *relative* accuracy measures for stroke gesture articulation that characterize the geometric, kinematic, and articulation accuracy of single and multi-stroke gestures. To compute the accuracy measures, we introduce the concept of a *gesture task axis*. We evaluate our measures on five public datasets comprising 38,245 samples from 107 participants, about which we make new discoveries; e.g., gestures articulated at fast speed are shorter in path length than slow or medium-speed gestures, but their path lengths vary the most, a finding that helps understand recognition performance. This work will enable a better understanding of users' stroke gesture articulation behavior, ultimately leading to better gesture set designs and more accurate recognizers.

Keywords

Gesture task axis; relative accuracy measures; gesture error; unistrokes; multi-stroke gesture; geometric accuracy; kinematic accuracy; articulation accuracy; toolkit.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

1. INTRODUCTION

As touch gesture interaction becomes more common, the need to understand how to design appropriate gesture interactions grows. One key aspect is understanding how users actually articulate pen and finger gestures. For example, variations in stroke gesture articulation have been studied in different ways in the literature, including examining the consistency between and within users [1], differences between user populations [9], and the impact of input devices [17]. However, all of these methods focus on characterizing *absolute* and *global* features of the gesture itself, such as total path length or articulation time. What has been missing are fine-grained analyses of gesture articulations to support an understanding of how gestures vary *relative* to each other and to recognizers' canonical template forms. Once we have this knowledge, better gesture set designs and more accurate template-based recognizers can be proposed.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICMI'13, December 9–13, 2013, Sydney, NSW, Australia
Copyright 2013 ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522875.

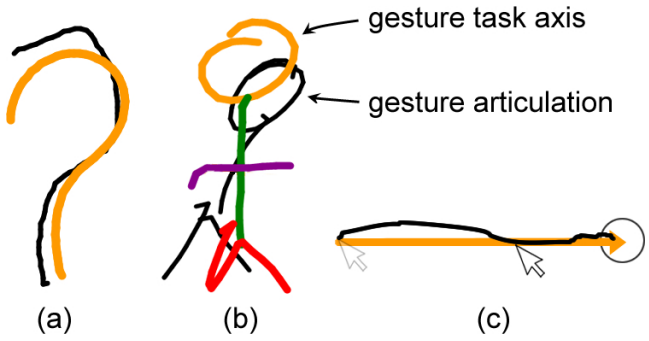

Figure 1: Gesture task axes for single and multi-stroke gestures (a, b), defined analogous to MacKenzie et al.'s [12] reference axis for pointing tasks (c). The gesture task axis, acting as representative articulation of a gesture type, is computed from a set of gesture samples.

We present a set of relative gesture accuracy measures that describe the way gestures unfold and what happens during their production in terms of their closeness to their ideal forms. Our measures are inspired by and analogous to MacKenzie et al.'s path-based pointing measures [12], which describe the accuracy of a pointing movement *while it unfolds*. As with that work, our work reveals the accuracy of gesture motion while it takes place, rather than just after-the-fact. MacKenzie et al. [12] defined the task axis as a straight line between the starting point of the user's mouse and the pointing target, to which the accuracy of users' pointing paths was compared (Figure 1c). We conceptualize the problem in the same way: a reference path or task axis is defined, and variations from this path are captured in the measures we present. We define the gesture task axis as a representative way to articulate a stroke gesture (Figure 1a,b), similar to the gesture templates stored by template-based gesture recognizers, such as $1 [24], $N [2,3], and $P [19]. Our accuracy measures then capture local deviations of users' candidate gestures relative to the gesture task axis, in terms of geometric, kinematic, and articulation accuracy. We evaluate our relative accuracy measures on five public gesture datasets comprising 38,245 samples from 107 participants and report new gesture findings, not revealed by absolute measures, that have implications for designing improved gesture sets and gesture recognizers.

The contributions of this work are: (1) a set of *relative stroke gesture accuracy measures* to characterize users' gesture articulation patterns; (2) an operational definition of the *gesture task axis*; (3) new findings regarding existing datasets enabled by our new accuracy measures; and (4) the Gesture RElative Accuracy Toolkit (GREAT) to compute the measures. These results can be used in new gesture studies to characterize in more detail how users make gestures. Ultimately, this work informs our understanding of designing gesture sets and recognizers to accommodate users' gesture articulation patterns.

2. RELATED WORK

In prior work, evaluating the performance of user stroke gesture articulation has been mainly confined to reporting conventional measurements, such as accuracy rates for recognizers [2,19,24], articulation time for user performance [6,24], and self-reported ratings for user preference [20,23]. Such high-level measures characterize gestural performance, but fail to capture more nuanced articulation behaviors. For example, pen and finger gestures have been found similar in terms of articulation time and shape distance, but different in aperture, corner distance, and intersecting points, which may impact future finger-gesture designs [17].

Beyond these general long-established measures in HCI, the articulation path of stroke gestures can be used to compute many geometric and kinematic features (e.g., see Blagojevic et al. [5] for a comprehensive set of 114 absolute gesture features). However, despite their potential to characterize articulation performance in fine detail, such features have been primarily used for gesture recognition [5,14,18,22], rather than to understand how users actually articulate gestures. In some cases, local features were used to inform gesture analysis approaches, such as the connection between curvature and tangential velocity employed by Cao and Zhai's model to predict gesture articulation time [6]. Only recently, a few studies have started to employ feature analysis to evaluate gesture articulation beyond error rates and task times. One example is Kane et al. [9], who compared gestures articulated by blind versus sighted people and reported differences picked up by newly-introduced features, such as gesture size variation and line steadiness. Tu et al. [17] defined new measures for stroke gestures, such as axial symmetry and intersecting points deviation. They were thus able to expose subtle differences between pen and finger gestures not revealed when employing only production time and proportional shape distance. Using absolute gesture features, Anthony et al. [1] found that increased geometric complexity of gesture shapes leads to people producing gestures less consistently.

In contrast to the work we present in this paper, gesture features investigated by all of these previous studies characterize *absolute* gesture articulation behavior, such as comparing path length or articulation time between different conditions. However, they cannot account for *relative* differences in individual gesture articulations. Yet such comparisons and analyses could inform the design of template-based recognizers, which operate based on comparisons between pairs of gestures (candidates and templates). To address this problem, we formulate new measures for quantifying and evaluating the accuracy of user gesture articulation behavior compared to some fixed example gesture (the gesture task axis), reflective of relative differences between individual executions.

3. RELATIVE ACCURACY MEASURES

We evaluate the precision of stroke gesture articulation relative to the gesture task axis in terms of (1) geometric, (2) kinematic, and (3) articulation accuracy. Geometric accuracy reflects how well users are able to reproduce a gesture, given its geometric shape alone. Kinematic accuracy captures differences in the time domain and, therefore, informs how fluent or smooth the gesture path is. Articulation accuracy measures how consistent users are in producing stroke gestures by looking at the difference in number of strokes and stroke ordering. Articulation accuracy may also be interpreted as recall accuracy, showing how well users can reproduce the exact execution details of a given gesture (e.g., always start the gesture from the same point and follow the same direction, a constraint imposed by some recognizers [24]). In order to evaluate the three types of accuracy, we employ the concepts of *error* and *variability* from MacKenzie et al. [12], who relied on them to evaluate the accuracy of pointing tasks[1]. Error is an indicator of the absolute difference between a measurement and a reference, and variability is the standard deviation of a set of differences.

In the following discussion, we represent a stroke gesture as a series of 2-D points, $p = \{p_i = (x_i, y_i, t_i) \mid i = 1..n\}$. The task axis is denoted by $\overline{p} = \{\overline{p}_i = (\overline{x}_i, \overline{y}_i, \overline{t}_i) \mid i = 1..n\}$, and will be defined in the next section. We employ standard local features at point p_i: (1) arc-length $s_i{=}\sum_{j=2}^{i} \|p_j - p_{j-1}\|$ for $i{=}2..n$ and $s_1{=}0$; (2) turning angle $\theta_i{=}\angle(\overline{p_{i-1}p_i}, \overline{p_ip_{i+1}})$; and (3) local speed $v_i{=}(s_{i+1} - s_{i-1}) \,/\, (t_{i+1} - t_{i-1})$. Local values are aggregated into global measures: (4) path length, $\mathcal{L}(p){=}s_n$; (5) total absolute turning angle, $\Theta(p){=}\sum_{i=2}^{n-1} \theta_i$; (6) area of gesture bounding box, $\mathcal{A}(p){=}(\max_{i=1,n} x_i - \min_{i=1,n} x_i) \times (\max_{i=1,n} y_i - \min_{i=1,n} y_i)$; and (7) average speed, $\mathcal{S}(p){=}(\sum_{i=2}^{n-1} v_i)/(n-2)$. We employ these measures later in the paper when we compare absolute versus relative measures.

All the measures we will introduce are relative to the gesture task axis and, therefore, the candidate gesture (for which the accuracy is to be determined) and the task axis must first be aligned. The alignment is a 1:1 matching defined as a permutation function $\sigma : \{1, 2, ..., n\} \rightarrow \{1, 2, ..., n\}$, meaning that point $\overline{p}_i$ on the task axis is aligned to point $p_{\sigma(i)}$ on the candidate gesture. If the gestures are always executed in the same direction (e.g., such as the unistrokes of [24]), this function is the identity permutation, $\sigma(i) = i$. Otherwise, the alignment is produced with the \$P matching technique [19].

3.1 Geometric accuracy

Geometric accuracy measures evaluate the deviation of the candidate gesture from the task axis in terms of shape distance, and capture tendencies of the users to stretch and bend strokes during articulation.

1. Shape Error (ShE) represents the average absolute deviation of the candidate gesture points from the task axis in terms of the Euclidean distance:

$$\text{ShE}(p) = \frac{1}{n} \sum_{i=1}^{n} \left\| p_{\sigma(i)} - \overline{p}_i \right\| \tag{1}$$

ShE relates to the cost function of the \$1 recognizer [24], the Proportional Shape Distance of SHARK^2 [10], and the \$P recognizer [19], depending on the point alignment procedure.

2. Shape Variability (ShV) computes the standard deviation of the distances between the points of the candidate and the task axis:

$$\text{ShV}(p) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(\left\| p_{\sigma(i)} - \overline{p}_i \right\| - \text{ShE}(p) \right)^2} \tag{2}$$

[1]MacKenzie et al. evaluated pointing tasks using, among other measures, movement error (ME), movement offset (MO), and movement variability (MV), computed relative to the pointing task axis [12].

Low ShV values show uniform shape distance errors across the entire gesture path, while large values indicate that errors are larger for some parts of the gesture and smaller for others.

3. Length Error (LE) measures users' tendencies to "stretch" gesture strokes with respect to the task axis:

$$\text{LE}(p) = \sum_{i=1}^{n} \left| s_{\sigma(i)} - s_i \right| \tag{3}$$

As we employ uniform resampling of gestures as a preprocessing step, similar to gesture recognizers [2,19,24], arc-lengths can be written as $s_i = (i-1) \cdot \frac{\mathcal{L}}{n-1}$, which leads to:

$$\text{LE}(p) = |\mathcal{L}(p) - \mathcal{L}(\overline{p})| \tag{4}$$

This compact form has the advantage of being generalizable for multi-stroke gestures, for which directly comparing arc-lengths of points aligned by the $P point-cloud algorithm is irrelevant. We can also measure the stretching behavior as the difference in gesture bounding box area, for which we derive the following definition analogous to LE:

4. Size Error (SzE) measures users' tendencies to "stretch" gesture strokes in terms of the gesture area size:

$$\text{SzE}(p) = |\mathcal{A}(p) - \mathcal{A}(\overline{p})| \tag{5}$$

5. Bending Error (BE) measures users' tendencies to "bend" the strokes of the articulated gesture with respect to the gesture task axis. It is defined as the absolute average of the differences between corresponding turning angles at the i^{th} point, measured on the gesture and the task axis:

$$\text{BE}(p) = \frac{1}{n} \sum_{i=1}^{n} \left| \theta_{\sigma(i)} - \overline{\theta}_i \right| \tag{6}$$

6. Bending Variability (BV) computes the standard deviation of the differences in turning angle:

$$\text{BV}(p) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(\left| \theta_{\sigma(i)} - \overline{\theta}_i \right| - \text{BE}(p) \right)^2} \tag{7}$$

3.2 Kinematic accuracy

Kinematic accuracy measures evaluate articulation differences in the time domain, and capture how fluent or smooth the articulated path is in terms of production time and speed.

7. Time Error (TE) measures the difference in articulation time (total duration) between the candidate and the task axis:

$$\text{TE}(p) = |\mathcal{T}(p) - \mathcal{T}(\overline{p})| \tag{8}$$

8. Time Variability (TV) represents the standard deviation of the differences between timestamps measured at each individual point on the gesture path:

$$\text{TV}(p) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(\left| t_i - \overline{t}_i \right| - \text{TE}(p) \right)^2} \tag{9}$$

9. Speed Error (VE)[2] measures the difference in the speed profiles of the candidate and the gesture task axis:

$$\text{VE} = \frac{1}{n} \sum_{i=1}^{n} |v_i - \overline{v}_i| \tag{10}$$

[2]Where letter "V" in VE comes from "Velocity," which is an innocent abuse of notation to prevent multiple, confounding abbreviations startingwith "S".

10. Speed Variability (VV) represents the standard deviation of the local differences between the speed profiles:

$$\text{VV} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(|v_i - \overline{v}_i| - \text{VE}(p) \right)^2} \tag{11}$$

3.3 Articulation accuracy

Articulation accuracy measures how consistent users are in producing the individual strokes of gestures, for which perfect accuracy is required by some recognizers (e.g., always start strokes from the same starting point and following the same direction) or use cases (e.g., Japanese script).

11. Stroke Count Error (SkE) reports the difference in the number of strokes between the candidate and the task axis.

12. Stroke Ordering Error (SkOE) is an indicator of stroke ordering accuracy, computed as the absolute difference between the $1 cost measure (defined as the sum of Euclidean distances between chronologically-aligned points) [24] and the $P cost measure (defined as the sum of Euclidean distances between point clouds) [19]:

$$\begin{aligned} \text{SkOE}(p) &= |\$1(p, \overline{p}) - \$P(p, \overline{p})| \\ &= \left| \sum_{i=1}^{n} \|p_i - \overline{p}_i\| - \sum_{i=1}^{n} \left\| p_{\sigma(i)} - \overline{p}_i \right\| \right| \end{aligned} \tag{12}$$

If the candidate gesture has been articulated in the same way as the gesture task axis, the ordering error will be low, as both $1 and $P will return approximately the same value. Should any difference exist in stroke ordering between the candidate and reference, SkOE will reflect it with a larger value.

4. THE GESTURE TASK AXIS

In the above measures, we use the term *gesture task axis* to mean a specific articulation of a gesture type that serves as a reference, against which the accuracy of other, candidate gestures is computed. The task axis can be a geometrically-perfect shape definition of the gesture (similar to MacKenzie et al.'s straight line for pointing tasks [12]), or it can be a representative gesture sample acquired from the user that captures the articulation specificity of both the user and the input device, as found in the training sets of template-based gesture recognizers [2,19,24]. Both alternatives represent reasonable design choices for computing gesture task axes to account for differences in gesture articulation behavior. The ideal shape can be specified as a series of geometric primitives, such as lines and arcs. Previous work has employed such representations for gestures, either for the purpose of recognition [13] or for gesture analysis [6]. However, despite the desirable objectivity of this perfectly-shaped reference, such a representation may not necessarily be reflective of actual user articulation patterns, which often reveal "chunking" and "corner-cutting" behaviors in an attempt to produce gestures faster [6] (p. 1503). Also, such a reference may not reflect the true accuracy of allographic handwriting [15] (e.g., individual differences in letters that personalize one's handwriting, such as $\mathcal{A}$, $\mathcal{A}$, A, or $\mathscr{A}$ for the letter "A"), for which it may artificially overemphasize differences where they do not exist. Therefore, an alternative would be a subjective, user-dependent reference, reflective of users' articulation patterns. Following these considerations, we arrived at three possible definitions for the *gesture task axis*:

1. The geometric gesture task axis (GEOMETRIC), defined by the gesture set designer by employing geometric primitives, such as lines and curves.
2. The average gesture task axis (AVERAGE), computed as the average shape of a set of user-captured gesture samples.
3. The canonical template form (TEMPLATE) supplied to a recognizer to which articulated gestures will be compared in a template-based matching approach.

The GEOMETRIC gesture task axis needs to be specified by the designer as a set of geometric primitives in a CAD-like manner. In order to help this process, we developed a simple application that takes the designer's stroke input and converts it into lines and curves. A stroke is replaced by a line if all the intermediate points are within a threshold distance from the line segment defined by the first and last points of the stroke. Otherwise, the stroke is modeled using a polynomial interpolating spline, for which the shape can be entirely customized by editing and manipulating its control points.

The AVERAGE gesture task axis is meant to be reflective of actual user-articulated gestures. Inspired by previous work on 2D shape averaging [7,16], we devised a simple technique to compute the "centroid gesture" of a set of samples. Let T be a set of samples for a given gesture symbol, $T = \{t_j | j = 1..|T|\}$, where each sample, t_j, is represented as a series of two-dimensional points. We first resample gestures to the same number of points and translate them such that their centroid point is at the origin (using for example the pseudocode made available with the $1, $N, or $P recognizers [2,19,24]). This approach allows us to work with point sets of equal cardinality, which simplifies the subsequent gesture alignment procedures, as well as assuring translation-invariance for the computed accuracy measures. One of the gesture samples is selected as a reference, to which all the other gestures are being aligned[3]. If the articulation patterns of the gestures in the set are alike (e.g., a rectangle that is always executed from the same starting point and in the same direction, such as in the $1 gesture set [24]), the alignment is a simple 1:1 matching that connects points with the same index on the two gestures. Otherwise, the point cloud alignment of the $P gesture recognizer [19] is used to best match the points of the two gestures for gestures that are executed with different stroke directions or stroke ordering, such as the gestures in [2]. Figure 2 illustrates the two point alignment types. Once the alignments are computed, point p_i of the reference gesture is associated to a set of $|T| - 1$ points from the remaining samples in the set. We then average the x and y coordinates of these points in all the gestures to compute the i^{th} point of the AVERAGE task axis.

We use the AVERAGE result to define the TEMPLATE gesture task axis as the gesture from the set that is "closest" to the average gesture (in terms of the nearest-neighbor classification procedure implemented by the $1 and $P recognizers, depending on how point alignments were computed). Using this approach, the TEMPLATE task axis is an actual user-articulated sample, representative of the ones stored by template-based gesture recognizers [2,19,24] in their training sets, and also representative of the "average" articulation performance of the user.

4.1 Pilot Study for the Gesture Task Axis

We note that the GEOMETRIC task axis is artificially created and, consequently, can only be used to assess geometric accuracy, as timestamps are unavailable. The AVERAGE task axis allows computation of the geometric and kinematic accuracy measures, but for single strokes only (in the case of multi-strokes, AVERAGE is a point-cloud [19], for which the execution details get lost during the alignment procedure). On the other hand, the TEMPLATE gesture task axis can compute all the geometric, kinematic, and articulation-related accuracy measures for both single and multi-strokes, as it represents an actual articulation of a stroke gesture. However, in order to better inform our choice for the task axis, we conducted a pilot evaluation, in which we computed the task axes for the single-stroke gestures of the $1 set [24] (p. 159) and the multi-stroke gestures of the MMG set [2] (p. 245), against which we computed the geometric accuracy measures (8,000 total gesture samples). We found the values significantly correlated (at p=.01), with Pearson r coefficients ranging between .138 and .903 for single-strokes (average r=.562) and between .117 and .909 for multi-strokes (average r=.659). These results confirm the hypothesis that correlations will exist between the values of relative measures computed against different task axes, simply because of their relative nature. Also, as anticipated, the objective GEOMETRIC task axis produced larger values (e.g., higher errors) for the accuracy measures than the subjective task axes (almost twice as large on average), because of its inability to capture allographic differences and "short-cutting" behaviors during articulation [6]. Also, given that results were similar for AVERAGE and TEMPLATE, but TEMPLATE represents an actual user articulation, we employ in this paper the TEMPLATE task axis as reference for computing relative accuracy measures. Figure 3 illustrates the TEMPLATE gesture task axes for single and multi-stroke gestures from actual users' data [2,20,22,24].

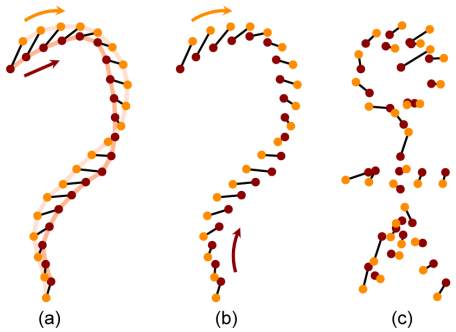

Figure 2: Gesture points are aligned in their chronological order of input if the two gestures conform to the same articulation pattern, such as the case of the "question mark" symbols (a) executed from the same starting point and in the same direction. Otherwise, points are aligned using the point-cloud matching procedure of $P [19], such as for the "question mark" symbols executed in opposite directions (b) or the "person" symbol (c) executed with different number of strokes and stroke ordering.

5. CASE STUDIES

We discuss the applicability of our relative accuracy measures to several existing stroke gesture experiments and datasets reported in previous work [2,3,8,20,22,24]. In order not to overload this paper with exhaustive numerical data supplied by the entire set of measures, we focus on one case study example per accuracy type: geometric, kinematic, and articulation-related. We discuss geometric accuracy measures on the $1 dataset [24] and compare our findings to the recognition results reported in that work. We then discuss kinematic accuracy measures for the gesture set of Vatavu et al. [20], who were interested in the time profile of stroke gestures for estimating execution difficulty. Finally, we employ three multi-

[3]Previous work [16] has showed that how the reference is selected has little influence on the final result.

Figure 3: Gesture task axes (orange lines) computed using the TEMPLATE scheme for single- and multi-stroke gestures from public datasets: the "question mark" symbol [24]; "flower" and "stairs" [20]; "pitchfork" and "asterisk" [2]; and the "person" symbol [22].

stroke datasets [2,8,22] to reveal new findings about users' articulation behavior in terms of stroke count and stroke ordering. Where possible, we compare differences in the values reported by absolute versus relative measures (such as path length versus length error).

5.1 Effect of Articulation Speed on Geometric Accuracy

We employ for this experiment the $1 gesture set [24], composed of 16 gesture types (p. 159) articulated by 10 participants with 10 repetitions each[4]. Participants entered gestures at low, medium, and fast speeds, corresponding to the instructions "as accurately as possible" (for low), "balance speed and accuracy" (medium), and "as fast as you can" (fast). The authors of the $1 work found a significant effect of articulation speed on recognition errors for all tested recognizers ($1, Rubine, and DTW). The smallest error occurred for medium speed, explained that "at medium speeds, subjects' gestures were neither overly tentative nor overly sloppy" (p. 166). In the following, we provide supporting data for this hypothesis by employing our relative measures, and reveal new findings on users' gesture articulation.

We computed task axes for each gesture type in the set under both user-dependent and user-independent training scenarios. For the first scenario, task axes were computed for each participant individually, while for the second, all participants' data contributed to the computation of the task axes. We then computed accuracy measures for candidate gestures, which were selected with a leave-one-out cross-validation testing procedure[5]: one gesture sample was selected as the candidate, while all the others were used to compute the task axis.

There was a significant effect of articulation speed on Shape Error ($\chi^2(2)$=70.355 for user-dependent and $\chi^2(2)$=77.889 for user-independent training, p<.001), with medium gestures showing the smallest difference in Shape Error and Shape Variability for the user-dependent case. This result confirms the original authors' finding on medium gestures exhibiting the highest recognition accuracy, as we can confirm such gestures are "closer" to their centroid, while the point-to-point distances exposed lower Shape Variability than encountered at other speeds (see Figure 4, next page). Slow and fast gestures, which delivered lower recognition accuracies, also presented larger Shape Error and Variability values in the user-dependent scenario, which is the scenario reported in [24]. The same negative correlation between recognition accuracy and Shape Error/Variability was found for the user-independent scenario[6]: an increase in Shape Error/Variability corresponds to lower recognition rates. These findings reveal new insights about the effect of articulation speed on users' stroke gesture executions, but also help explain why $1's recognition rates vary under increased articulation speed. Using our relative measures, we confirm the $1 authors' hypothesis about fast gestures being more "sloppy" and show that shape errors are equally introduced by over-focus on accuracy (i.e., slow gestures being more "tentative"). We note that this confirmation cannot be delivered by absolute performance measures (e.g., recognition rates reported in [24]) that only show differences between conditions, without providing explanations for these differences.

Besides these results, we can further employ our relative stretching and bending accuracy measures to report new findings about the gestures in the $1 set. For instance, articulation speed had a significant effect on absolute gesture length ($\chi^2(2)$=191.649, p<.001) and area ($\chi^2(2)$=276.001, p<.001), with both length and area decreasing with increased speed (Figure 5 left, next page). However, a closer look employing our relative measures shows people producing gestures with different variations in length and area depending on the articulation speed (Figure 5 right, next page), with medium-speed gestures being the most accurately articulated (p<.001).

Interestingly, participants exhibited a tendency to stretch their strokes more when producing fast gestures, as indicated by Length Error, but slow articulations led to larger variations in gesture area, as shown by Size Error ($p < .001$). We also found that participants exhibited a tendency to bend their strokes relative to the average behavior. The significant effect of speed on the absolute turning angle ($\chi^2(2) = 887.134$, $p < .001$) was also revealed in the Bending Error and Variability relative measures (Figure 6, next page). However, the relative measures characterize users' articulations at more subtle levels of detail. For example, we now know that the average shape "bending" expected from users across consecutive articulations of gestures from the $1 set lies between 0.22 and 0.26 radians (12.6−15.0°), and users are more accurate in producing the curvature of these shapes at fast speeds.

5.2 Effect of Gesture Practice on Kinematic Accuracy

We employ for this experiment the gesture set of Vatavu et al. [20], which contains 18 gestures (p. 94) executed by 14 participants with 20 repetitions each[7]. The set includes both familiar (practiced) and unfamiliar (new) gestures, verified by asking participants. Vatavu et al. found that articulation time correlated best with perceived difficulty; i.e., gestures that took longer to articulate were generally perceived as more difficult by users. However, no explanation is provided for the cause of this phenomenon. Next, we reveal new findings about users' articulation behavior in the time domain by employing our kinematic relative measures.

Absolute measures show that familiar gestures were articulated with lower overall duration (1323 vs. 2433 ms, Z=−40.975, p<.001) and faster (0.7 vs. 0.5 pixels/ms, Z=−37.660, p<.001)

[4] http://depts.washington.edu/aimgroup/proj/dollar/

[5] This has the desirable property of being almost unbiased [21] (p. 255).

[6] Because the $1 work [24] does not report user-independent recognition rates, we computed them by following the same training/testing procedure.

[7] http://www.eed.usv.ro/~vatavu/index.php?menuItem=downloads

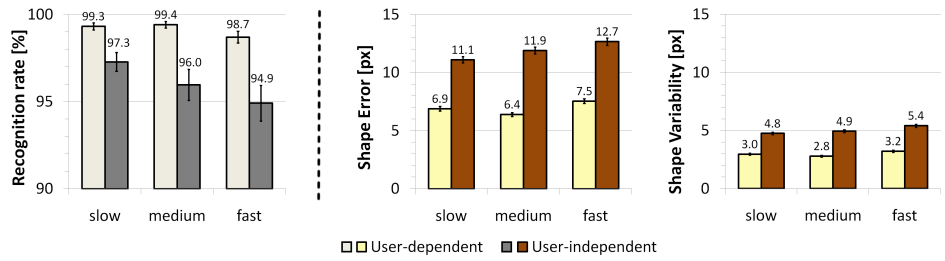

Figure 4: Recognition rates for the $1 recognizer (left) compared to Shape Error and Variability measures (right). Note how Shape Error and Variability reflect differences in recognition rate (e.g., medium gestures have the smallest errors and are the most accurately recognized in the user-dependent case).

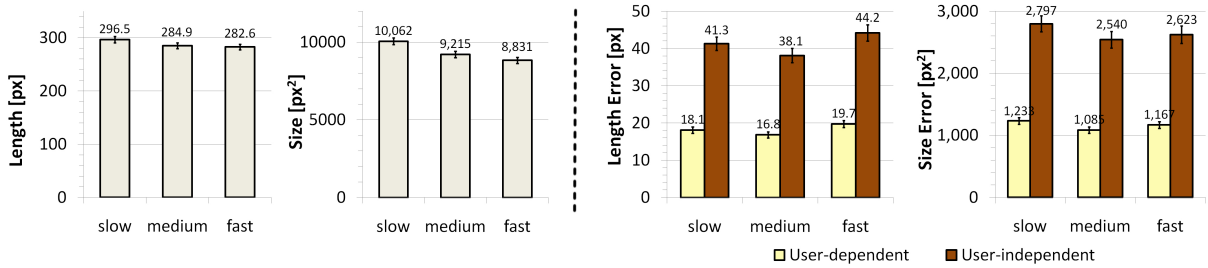

Figure 5: Absolute gesture length and size (left) compared to relative error measures (right). Note how the relative measures reveal articulation characteristics not captured by absolute measures (e.g., even though fast gestures are shorter in path length, their lengths varied the most).

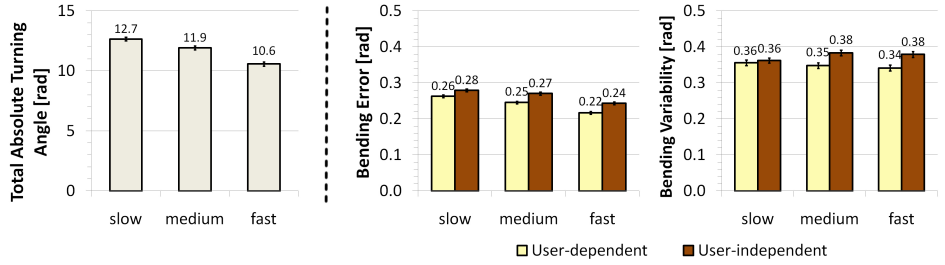

Figure 6: Absolute bending (left) compared to relative bending accuracy measures (right). Note how the relative measures replicate the findings of the absolute measures for the user-dependent case, but Bending Variability shows a different trend for the user-independent scenario.

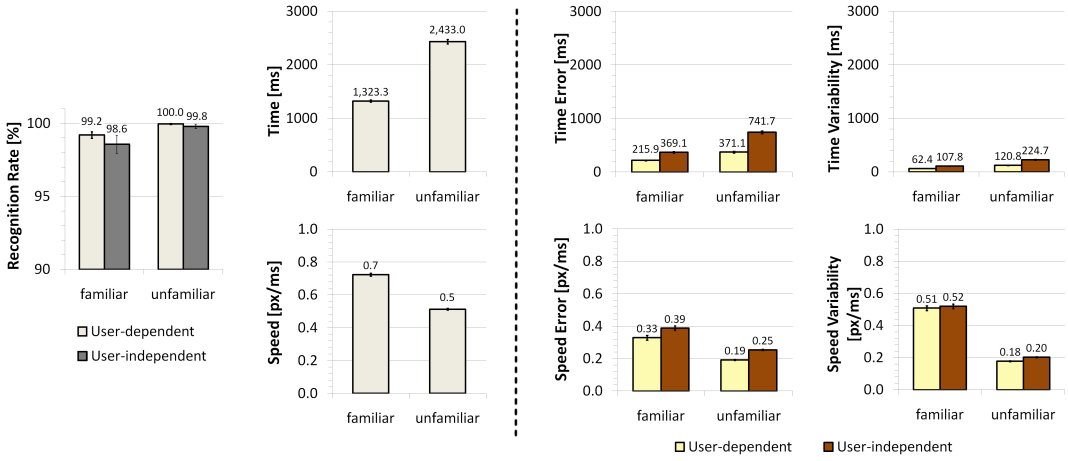

Figure 7: Recognition rates ($1) and absolute kinematic measures (left) vs. relative measures (right). Note how relative measures reveal articulation behavior not captured by absolute measures (e.g., familiar gestures are more variable in speed), which correlates with differences in recognition rates.

than unfamiliar ones. This result is enough in most cases to confirm the practice effect on gesture articulation, which separates novice from expert users of a gesture interface [4]. However, our relative accuracy measures show more insight on the effect of practice. For instance, participants articulated familiar gestures with smaller Time Error and Variability than measured for unfamiliar ones (as confirmed by Wilcoxon tests for both user-dependent and independent scenarios, $p<.001$). This result shows that practice leads not only to smaller articulation times, but also to smaller errors, as individual articulations are produced more consistently in the time domain (Figure 7, previous page).

The analysis of relative speed accuracy measures also shows surprising results, hard to infer otherwise: Speed Error and Variability were found to be twice as large for familiar than for unfamiliar gestures ($p<.001$). We see that, when entering familiar gestures, users are faster, decreasing their carefulness. These nuanced findings in the time domain are in accordance with the recognition rates[8] of familiar gestures, which are significantly smaller than those of unfamiliar gestures (Figure 7, left, previous page) for both user-dependent ($Z=-7.842$, $p<.001$, small Cohen effect $r=.25$) and independent training ($Z=-6.098$, $p<.001$, medium effect $r=.36$).

5.3 Articulation Accuracy of Multi-Stroke Gestures

In this section we employ our relative accuracy measures to report new findings on how users articulate multi-stroke gestures. We use the MMG gesture set [2,3] (3,200 samples, 16 gesture types, and 20 participants: 10 participants employed a stylus, the other 10 used a finger)[9]; the HHReco set [8] (6,986 samples, 13 gestures, and 19 participants)[10] and the Nic-Icon set [22] (13,819 samples, 14 gestures, and 34 participants)[11].

We found a significant effect of gesture instrument (stylus versus finger) on Stroke Ordering Error ($Z=-3.255$ for user-dependent and $Z=-4.228$ for user-independent training, $p<.001$), with finger gestures exhibiting more variation in stroke ordering than stylus gestures for the user-dependent scenario (Figure 8 right, next page). This finding can be explained by the years of experience with handwriting with a pen that make people fall back on those patterns more often when using the pen than the finger; only recently have people started making gestures on touchscreens by using their fingers. However, stylus stroke ordering patterns seem to be more consistent within- than between-users, as more variation was found for stylus gestures in the user-independent training. There was no significant effect of instrument on recognition accuracy[12] of \$P [19] in the user-dependent scenario ($U=201788.00$, $z=-0.637$, n.s.). The effect *was* significant, however, for user-independent training ($U=27585.50$, $z=-3.104$, $p<.01$), with finger gestures being more accurately recognized than stylus gestures (Figure 8 left, next page).

Analysis on the absolute number of strokes employed by participants[13] showed no significant differences between the average stroke count of these sets (Figure 8, left), as indicated by a Mann-Whitney test ($U=59.00$, $z=-1.55$, n.s.). However, the relative Stroke Count Error measure revealed higher accuracy within users and lower consistency (higher errors) between users. This finding is prominently reflected by the HHReco set ($Z=-53.09$, $p<.001$), for which the user-dependent SkE approaches zero (showing almost perfect within-user consistency), while the articulations of different users are different on average by one stroke (Figure 8, right). The findings enabled by our relative accuracy measures are in accordance with the recognition results obtained for the HHReco and NicIcon sets (Figure 8, left). In the user-dependent case, users are more consistent in terms of number of strokes in the HHReco set (SkE=0.02) than in the NicIcon set (SkE=0.18), which translates into higher recognition rates for HHReco (94.9% vs. 92.9%). In the user-independent case, the situation is reversed: there is more consistency between users for the NicIcon set (SkE=0.42) than for HHReco (SkE=1.10) and, therefore, the recognition rates we obtained for NicIcon are higher this time (75.8% vs. 72.1%). We note again how the relative accuracy measures we have introduced capture these subtle differences and provide more insight into how users articulate their stroke gestures.

6. SUMMARY, IMPLICATIONS, AND TOOLKIT SUPPORT

By employing our relative measures we were able to reveal new aspects of how users articulate stroke gestures, not captured by absolute measures. For instance, our findings show how articulation speed affects shape consistency relative to templates stored by gesture recognizers. Users are more accurate when balancing speed and accuracy than when aiming for either high accuracy or high speed. Based on these findings, we recommend training procedures that collect gesture samples articulated at different speeds. Such a recommendation is important for scenarios in which faster gesture input is more likely (e.g., on the go) and for recognition approaches that rely on geometric features to discriminate between gestures, such as the Rubine recognizer [14] and its derived forms [5].

Our findings also show that practice affects both articulation time and speed error, but in opposite ways: whereas familiar gestures are articulated more consistently in the time domain, they also exhibit larger speed errors. When entering familiar gestures, users are faster, decreasing their carefulness. Consequently, designers should employ recognition approaches that are not sensitive to how gestures are different in time and speed, or should capture a sufficient number of training samples to account for these variations. For example, Rubine's recognizer employs time features, among other gesture descriptors, and Rubine recommends 15+ training samples per gesture type [14] (p. 335).

We also found differences in stroke ordering for both finger and stylus gestures, which encourages the use of recognition approaches robust to the order in which users enter strokes [2] or gesture recognizers innately designed to handle such variations [19].

In the end, we release the Gesture RElative Accuracy Toolkit (GREAT) to the community for further gesture studies and for practitioners of gesture interfaces who need to tweak gesture sets and recognizers for various application contexts. We release this tool to enable new gesture discoveries between different conditions (e.g., different user populations [9], input devices [17], or gesture types [11]) and thus foster improved gesture interface designs. GREAT computes our set of twelve relative accuracy measures and can be downloaded for free online[14].

[8] Recognition rates were computed using the \$P recognizer as per the procedure of [19], by varying the number of training samples per gesture type $T = 1, 2, 4, 8$ and the number of training participants $P = 1, 2, 4, 8$.

[9] http://depts.washington.edu/aimgroup/proj/dollar/ndollar.html

[10] http://embedded.eecs.berkeley.edu/research/hhreco/

[11] http://unipen.nici.ru.nl/NicIcon/index.php

[12] Recognition rates were computed using the \$P recognizer as per the procedure of [19], by varying the number of training samples per gesture type $T = 1, 2, 4, 8$ and the number of training participants $P = 1, 2, 4, 8$.

[13] Only on the NicIcon and HHReco sets, as the MMG participants were instructed to produce gestures with predefined number of strokes, which makes Stroke Count Error analysis unavailable for this set.

[14] http://depts.washington.edu/aimgroup/proj/dollar/great.html

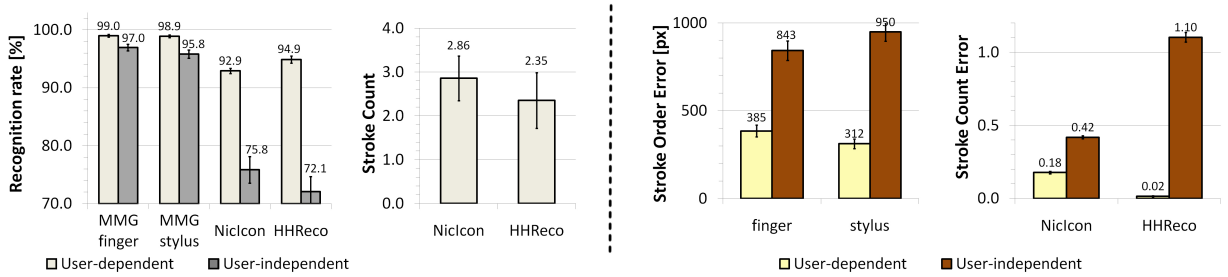

Figure 8: Recognition rates ($P) and absolute stroke count measure (left) versus relative Stroke Ordering and Stroke Count Error (right). Note how the Stroke Count Error reveals differences in user consistency for the HHReco and NicIcon sets, not captured by the absolute number of strokes.

7. CONCLUSION

We reported in this paper twelve new accuracy measures to examine relative differences between users' stroke gesture articulations, for which we have introduced the concept of a gesture task axis. We showed how our measures go beyond prior work by capturing relative gesture differences, and reveal new findings on users' stroke gesture articulation behaviors from five public datasets, especially finding relationships between users' articulation and recognition accuracy for different recognizers. We hope the contributions of this work will lead to a better understanding of users' gesture articulation behaviors for diverse contexts, and we are eager to see how the community might use the new relative measures and our toolkit to improve gesture interface designs.

8. REFERENCES

[1] Anthony, L., Vatavu, R.-D., and Wobbrock, J. O. Understanding the consistency of users' pen and finger stroke gesture articulation. Canadian Inf. Proc. Soc. (Toronto, Ont., Canada, 2013), 87–94.

[2] Anthony, L., and Wobbrock, J. O. A lightweight multistroke recognizer for user interface prototypes. GI '10, Canadian Inf. Proc. Soc. (Toronto, Ont., Canada, 2010), 245–252.

[3] Anthony, L., and Wobbrock, J. O. $N-Protractor: a fast and accurate multistroke recognizer. GI '12, Canadian Inf. Proc. Soc. (Toronto, Ont., Canada, 2012), 117–120.

[4] Bau, O., and Mackay, W. E. Octopocus: a dynamic guide for learning gesture-based command sets. UIST '08, ACM (New York, NY, USA, 2008), 37–46.

[5] Blagojevic, R., Chang, S. H.-H., and Plimmer, B. The power of automatic feature selection: Rubine on steroids. SBIM '10, Eurographics Association (Aire-la-Ville, Switzerland, 2010), 79–86.

[6] Cao, X., and Zhai, S. Modeling human performance of pen stroke gestures. CHI '07, ACM (New York, NY, USA, 2007), 1495–1504.

[7] Chen, S. E., and Parent, R. E. Shape averaging and it's applications to industrial design. *IEEE Comput. Graph. Appl. 9*, 1 (Jan. 1989), 47–54.

[8] Hse, H., and Newton, A. Recognition and beautification of multi-stroke symbols in digital ink. *Computers & Graphics 29*, 4 (2005), 533–546.

[9] Kane, S. K., Wobbrock, J. O., and Ladner, R. E. Usable gestures for blind people: understanding preference and performance. CHI '11, ACM (New York, NY, USA, 2011), 413–422.

[10] Kristensson, P.-O., and Zhai, S. SHARK^2: a large vocabulary shorthand writing system for pen-based computers. UIST '04, ACM (New York, NY, USA, 2004), 43–52.

[11] Long, Jr., A. C., Landay, J. A., and Rowe, L. A. Implications for a gesture design tool. CHI '99, ACM (New York, NY, USA, 1999), 40–47.

[12] MacKenzie, I. S., Kauppinen, T., and Silfverberg, M. Accuracy measures for evaluating computer pointing devices. CHI '01, ACM (New York, NY, USA, 2001), 9–16.

[13] Paulson, B., and Hammond, T. Paleosketch: accurate primitive sketch recognition and beautification. IUI '08, ACM (New York, NY, USA, 2008), 1–10.

[14] Rubine, D. Specifying gestures by example. *SIGGRAPH Comput. Graph. 25*, 4 (July 1991), 329–337.

[15] Schomaker, L. From handwriting analysis to pen-computer applications. *IEEE Electron. Commun. Eng. J. 10*, 3 (1998), 93–102.

[16] Sebastian, T. B., Klein, P. N., Kimia, B. B., and Crisco, J. J. Constructing 2D curve atlases. The IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA '00), IEEE Computer Society (Washington, DC, USA, 2000), 70–77.

[17] Tu, H., Ren, X., and Zhai, S. A comparative evaluation of finger and pen stroke gestures. CHI '12, ACM (New York, NY, USA, 2012), 1287–1296.

[18] Vatavu, R.-D. 1F: One accessory feature design for gesture recognizers. IUI '12, ACM (New York, NY, USA, 2012), 297–300.

[19] Vatavu, R.-D., Anthony, L., and Wobbrock, J. O. Gestures as point clouds: a $P recognizer for user interface prototypes. ICMI '12, ACM (New York, NY, USA, 2012), 273–280.

[20] Vatavu, R.-D., Vogel, D., Casiez, G., and Grisoni, L. Estimating the perceived difficulty of pen gestures. INTERACT'11, Springer-Verlag (Berlin, Heidelberg, 2011), 89–106.

[21] Webb, A. *Statistical Pattern Recognition, 2nd Edition*. John Wiley & Sons Ltd., West Sussex, England, 2003.

[22] Willems, D., Niels, R., van Gerven, M., and Vuurpijl, L. Iconic and multi-stroke gesture recognition. *Patt. Rec. 42*, 12 (2009), 3303–3312.

[23] Wobbrock, J. O., Morris, M. R., and Wilson, A. D. User-defined gestures for surface computing. CHI '09, ACM (New York, NY, USA, 2009), 1083–1092.

[24] Wobbrock, J. O., Wilson, A. D., and Li, Y. Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. UIST '07, ACM (New York, NY, USA, 2007), 159–168.