Understanding the Consistency of Users' Pen and Finger Stroke Gesture Articulation

Lisa Anthony[*], Radu-Daniel Vatavu[†], Jacob O. Wobbrock[‡]
[*]UMBC Information Systems, 1000 Hilltop Circle, Baltimore, MD 21250 USA
[†]University Stefan cel Mare of Suceava, Suceava 720229, Romania
[‡]Information School — DUB Group, University of Washington, Seattle, WA 98195-2840 USA

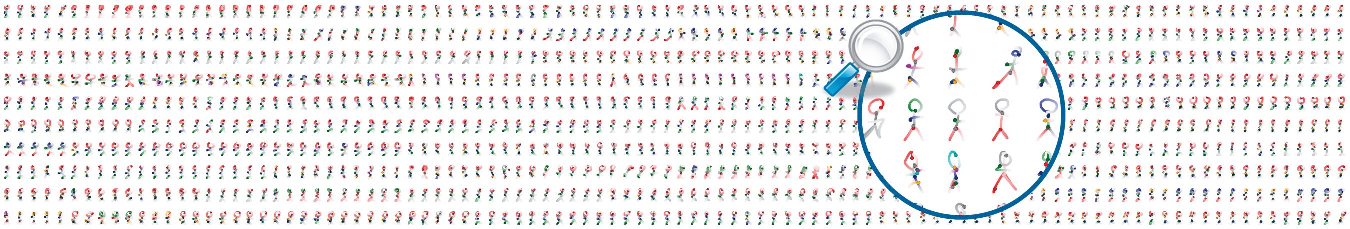


Figure 1: Nearly one thousand executions for the "person" symbol [46], performed by 34 users, with 52 different production patterns identified. These samples represent just 2.5% of the total number of 40,305 gestures that we analyzed for this study. NOTE: colors show stroke ordering and were automatically produced by GECKo, a gesture clustering toolkit that we release as a companion software application to this study.

Abstract

Little work has been done on understanding the articulation patterns of users' touch and surface gestures, despite the importance of such knowledge to inform the design of gesture recognizers and gesture sets for different applications. We report a methodology to analyze user consistency in gesture production, both between-users and within-user, by employing articulation features such as stroke type, stroke direction, and stroke ordering, and by measuring variations in execution with geometric and kinematic gesture descriptors. We report results on four gesture datasets (40,305 samples of 63 gesture types by 113 users). We find a high degree of consistency within-users (.91), lower consistency between-users (.55), higher consistency for certain gestures (e.g., less geometrically complex shapes are more consistent than complex ones), and a loglinear relationship between number of strokes and consistency. We highlight implications of our results to help designers create better surface gesture interfaces informed by user behavior.

Index Terms: H.5.2 [Information interfaces and presentation (e.g., HCI)]: User interfaces-input devices and strategies.

1 Introduction

Touch and surface gesture interaction is becoming a dominant form of everyday interaction as smartphones and tablet computers come to be more widespread. In addition to standard swipe, flick, and pinch gestures, sketch- or handwriting-based gestures are being used for a variety of applications [5, 23, 28]. Supporting gesture interaction requires recognizers to be integrated into the system and trained to the specific gestures the system will support. However, most recognizers have inherent limitations in the types of gestures they can discriminate (cf., [1, 25]). In the past, recognition algorithms have been tailored to specific applications, and much trial-and-error is employed while tweaking recognition parameters and thresholds in order to improve recognition rates on a specific gesture set [2]. Long et al. [25] observed that individual gestures within a gesture set affect recognition of each other, leading to post-hoc removal of specific gestures to tweak performance [8].

In this context, little work has been done to understand the range of users' gesture articulation patterns (which may impact recognition), despite the fact that currently popular gesture recognizers like \$1 [50], \$N [2, 3], and Protractor [24] require an explicitly-defined template for each gesture articulation to be recognized. Other recognizers, such as \$P [43], were specifically designed to ignore articulation differences altogether, which led indeed to improved recognition accuracy, but with the side effect of losing the capability to discriminate between directional strokes [43] (p. 278). For an example of the degree of variability possible with multistroke gestures, Figure 1 illustrates nearly 1,000 executions for the "person" symbol [46], among which we identified 52 distinct production patterns. Even a simple "asterisk" can be articulated in up to $2^3 \cdot 3! = 48$ different ways in terms of stroke direction and ordering (Figure 2 shows 14 of them). If we knew more about how users actually make gestures (e.g., which articulations are most common), we could design recognizers that capitalize on critical consistencies and differences in within- and between-user articulation in order to improve accuracy. We could also design gesture sets that run less risk of conflicts due to how users make gestures [25].

We analyze in this paper *user consistency* in touch and pen gesture production, focusing on (a) articulation features, such as number of strokes, stroke ordering, and stroke direction, and (b) execution variation, captured by geometric and kinematic descriptors. We report consistency in gesture articulation patterns for four previously published datasets including 40,305 samples of 63 gesture types produced by 113 users. Specifically, we find a high degree of consistency within users (.91), lower consistency between

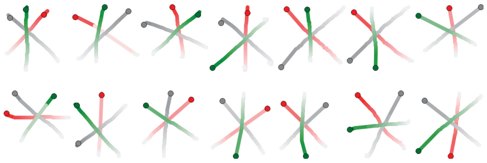


Figure 2: Fourteen different ways to articulate an "asterisk" identified among 200 executions from 20 users [3].

[*]lanthony@umbc.edu
[†]vatavu@eed.usv.ro
[‡]wobbrock@uw.edu

users (.55), higher consistency for certain gesture types (e.g., less geometrically complex shapes are more consistent than complex shapes), and a loglinear relationship between number of strokes and consistency. The contributions of this work are as follows: (1) a methodology and supporting tool for clustering gesture sets and analyzing user consistency in stroke gesture articulation (reported as agreement rates [48]); (2) a set of operationalized gesture features that can be used to characterize user consistency in articulating stroke gestures; (3) empirical findings on within- and between-user consistency on real gestures; and (4) practical implications for gesture interface prototypers to improve the performance of their designs. We are the first to examine how consistent humans are at producing stroke gestures, and our results are based on the largest experiment ever conducted on gesture input behavior (40,305 gesture samples) with high replicability (gesture data is from public datasets, and our gesture clustering tool is public released).

2 Related Work

2.1 User Consistency and User-Defined Gestures

Little work has attempted to understand the full range of users' gesture articulation patterns. One example is Hammond and Paulson [15], who examined the number of strokes users drew when sketching primitive shapes (e.g., lines, squares, circles, and curves), in order to inform the design of a multistroke sketch recognizer. Sezgin and Davis [37] used observed consistencies in stroke ordering to improve sketch recognition in domains such as course-of-action diagrams and circuit diagrams. Kane et al. [18] investigated gesture differences between blind and sighted users, Mauney et al. [30] explored the impact of different cultures on gesture articulation, and Tu et al. [40] studied pen versus finger gestures. However, no one has conducted as thorough an examination of user consistency (e.g., many different users, domains, gesture types, and features) as we present.

Agreement between user gestures has been examined in the context of user-defined gestures, as a replacement for expert designs that may be too tailored to technical constraints [49]. In fact, Morris et al. [31] noted that expert designers tend to propose gesture sets that are too complex compared to user-elicited gestures. Wobbrock et al. [49] defined a methodology for eliciting user-defined gesture sets by asking users, given the effect of a gesture, to demonstrate the cause that would invoke it. Many studies have since been conducted following this approach (for a survey, see Vatavu [42]), with all results showing user consistency in proposing gestures for similar tasks, even across domains [22]. A somewhat different methodology to explore the joint user-sensor motion space was introduced by Williamson and Murray-Smith [47]. Their work employs positive reinforcement to reward the originality of users while exploring the space of motions they are able to perform and sensors are able to capture. However, none of these studies have examined low-level features of user-defined gestures in order to understand how to build recognizers that can accommodate them.

These user-defined gesture elicitation studies can be large and expensive. Predictive models based on users' perceptions of gesture similarity represent a suitable alternative to help designers choose the best set of gestures for an application [26, 44]. For example, Vatavu et al. [44] investigated gesture execution difficulty, while Long et al. [26] focused on visual similarity to group gestures. We use similar features as these approaches but go beyond their simple demonstrations of user consistency in the *perception* of gesture shapes to report consistency in actual gesture *articulation*.

2.2 User Consistency and Gesture Recognition

Gesture recognition approaches vary but many use features similar to the ones we examine in this paper [15, 26, 35, 46]. We discuss the original work from which we borrow features as they are introduced in the analysis. By studying user consistency for features commonly used to recognize gestures, we can estimate how discriminative the features are and their impact on accuracy. Furthermore, it is well accepted among handwriting recognition research that recognition rates are higher for user- and task-dependent cases [11, 29]. Recognition accuracy is usually higher for domain-specific (e.g., smaller) applications [27], or for writer-dependent systems in which the recognizer has been trained on the writing of a given user [38]. This accuracy boost partially comes from users' internally consistent handwriting [9]. Writer identification research has also found that handwriting styles can be highly individual [39], but it is possible to cluster writing styles between users to improve accuracy of recognition algorithms [9]. Preliminary work in multi-touch gesture interaction has found similar heterogeneity between users [36]. We extend such findings on within- and between-user consistency from handwriting recognition to general gesture recognition, which includes more symbol types to be drawn.

2.3 Kinesthetics and Motor Control

Research in motor control theory has sought to understand the kinematic processes that occur during handwriting and, especially, what affects handwriting variability [10]. Two proposed models of fine human movement production are the Sigma- and Delta-lognormal models of the Kinematic Theory of Rapid Human Movements [34]. They state that generation of a complex movement requires the central nervous system to generate an action plan in the form of a series of virtual targets, reached via rapid strokes of the neuromuscular system. Kinematic Theory has been used to investigate the variability of handwriting patterns by considering local fluctuations of individual strokes and global fluctuations in how these strokes are sequenced [10]. Modeling of handwriting distortion has generated synthetic signature and gesture specimens exhibiting the same lognormal characteristics as genuine human movements [13] while recognizers trained on such data deliver improved accuracy [12].

Such studies offer mathematically-precise modeling of fine human movements with proven impact on the design of pattern recognition systems [12, 13]. However, note that motor control theory's definition of a "stroke" is different from what HCI researchers usually define as a pen trajectory between two consecutive pen-down and pen-up events [17]. In motor control, a stroke is a subcomponent of the pen movement, and exhibits a stereotypical bell-shaped velocity profile [21]. Unistrokes [14] would therefore be composed of multiple such strokes. In contrast to motor control theory, which looks at low-level analysis of human movement production, we are interested in a high-level understanding of stroke gestures and user production patterns. We therefore analyze user consistency by adopting the high-level HCI definition of a stroke. We focus on understanding differences in the number of strokes, stroke direction, and stroke ordering as they are naturally produced by users during single stroke and multistroke gesture articulation. We also aim for an understanding of user drawing behavior by employing today's HCI research tools to assess user consensus [42, 49].

3 Study Method

To develop an understanding of general patterns of user consistency both within- and between-users in gesture articulation: (1) we semi-automatically clustered gesture samples together based on articulation similarities (i.e., number of strokes, stroke orders, and stroke directions) in order to compute the degree of agreement per gesture type and dataset; and (2) we computed a set of features for each of the 40,305 gestures in four previously published datasets in order to understand how variation in feature values may be related to gesture articulation consistency according to our clustering.

3.1 Gesture Clustering

Recent work has shown the benefits of clustering gestures for re-organizing the structure of training sets and improving recognition accuracy [19, 33]. For example, Ouyang and Li [33] employed clustering to merge similar gesture patterns supplied by many contributors in an attempt to construct a large, and continuously evolving, gesture dictionary for touchscreen mobile devices. Keskin et al. [19] used clustering as a preprocessing step to reconfigure the number and structure of gesture classes according to the actual similarity present between samples in the training set.

Inspired by the success of these recent approaches to leveraging clustering algorithms for improving recognition performance, we also decided to employ gesture clustering techniques, but this time for the purpose of understanding user consistency in gesture articulation patterns. We cluster large gesture data sets in order to identify and group together similar production patterns that people naturally employ while articulating single stroke and multistroke gestures. To this end, we implemented the agglomerative hierarchical approach [45] (p. 363) with the complete-link method (p. 367), similar to [33]. The clustering technique starts with simple clusters consisting of one gesture sample only, and then iteratively merges clusters that are close together with respect to some similarity function. The process stops when the similarity between clusters falls below a given threshold and clusters cannot be merged any longer. During pretests, we experimented with different similarity measures inspired by gesture metrics [24, 32, 50] and gesture features [46]. We finally adopted a simple definition for gesture similarity by relying on the normalized Euclidean distance [20, 50], as we found it to deliver the best results. Therefore, for clustering purposes, we define the similarity between gestures a and b as follows:

$$similarity(a,b) = \begin{cases} 1 - \frac{\|a-b\|}{n} & \text{if } S(a) \ = S(b) \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

where $S(a)$ represents the number of strokes of gesture a; $\frac{\|a-b\|}{n}$ the normalized Euclidean distance between gestures a and b [50]; and n the number of sampling points for each gesture ($n = 64$) [50]. After normalization[1], the values of $similarity(a,b)$ fall in $[0..1]$, with 0 denoting no similarity at all and 1 denoting a perfect match.

Although accurate, the clustering results obtained using this automatic procedure were not perfect (as expected, since clustering relied on the existing body of knowledge on gesture recognizers, which inherently exhibit classification errors [2, 3, 24, 35, 50]). However, reporting precise measurements of user gesture articulation consistency requires perfect clustering. Therefore, we adopted a two-step hybrid clustering approach, in which a human operator verified the output of the automated clustering process and performed corrections where necessary by splitting and merging computer-generated clusters. As manual editing proved tedious, we developed several gesture visualization techniques to assist the process which are now part of the publicly released tool GECKo (GEsture Clustering ToolKit). (We detail these techniques and GECKo later in the paper.) The hybrid two-step clustering methodology led to perfect (though subjective) gesture clusters based on the following criteria: (a) gesture type, (b) number of strokes, (c) stroke order, (d) stroke direction, and (e) starting angle (e.g., orientation).

Then, to measure user consistency in producing gestures, we calculated agreement rates based on Wobbrock et al.'s method [48], previously successfully adopted for evaluating gesture sets [31, 41, 42, 48, 49]. Specifically, if gesture a has been produced in m different ways, for which we know the clustering partition $P = \{P_1, P_2, ..., P_m\}$, then the agreement rate of a is defined as:

$$AR_a = \sum_{i=1,m} \left(\frac{|P_i|}{|P|}\right)^2 \qquad (2)$$

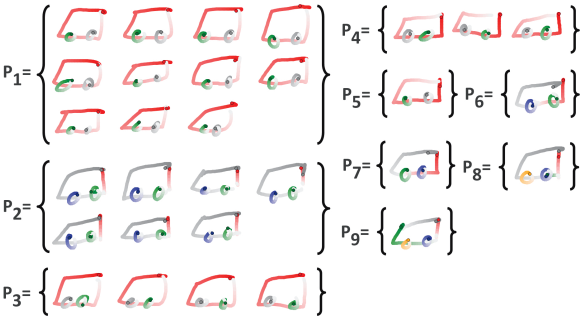


Figure 3: Nine different ways in which a user drew a car [46], consisting in different numbers of strokes, stroke direction, and stroke ordering. Resultant agreement rate is .22.

where $|P_i|$ represents the number of samples in cluster P_i. For example, Figure 3 illustrates different ways of drawing a "car" symbol, produced by a single user [46], showing articulations that differ in terms of number of strokes, stroke direction, and stroke ordering. The agreement rate for this gesture is therefore:

$$AR_{car} = \left(\frac{11}{30}\right)^2 + \left(\frac{7}{30}\right)^2 + \left(\frac{4}{30}\right)^2 + \left(\frac{3}{30}\right)^2 + \left(\frac{1}{30}\right)^2 + \left(\frac{1}{30}\right)^2 + \left(\frac{1}{30}\right)^2 + \left(\frac{1}{30}\right)^2 + \left(\frac{1}{30}\right)^2 = 0.22$$

We calculate agreement rates per gesture type in two conditions: within-user (using data from one user at a time), and between-users (by clustering the within-user clusters across all users). This agreement rate analysis is discussed in section 4.

3.2 Gesture Consistency Features

While gesture clusters capture information about preferred gesture articulation patterns in terms of numbers of strokes, stroke direction, and stroke ordering, we are also interested in execution variation of the articulated strokes, which we compute with geometric and kinematic gesture descriptors. In order to do so, we examined gesture features from the existing literature on gesture recognition [26, 35, 46]. We started our feature collection by considering all the features from previous studies that (1) we filtered during a first theoretical analysis based on their potential to correlate with articulation consistency, and (2) we filtered again based on actual measurements and correlation results. This preliminary analysis led to a final set of twelve representative features (Table 1). The set contains features that describe gesture path length and size [46], gesture structure (i.e., number of strokes), orientation (start and ending angles), shape (e.g., sharpness and curviness [26]), and kinematics (i.e., production time and speed [35]).

3.3 Gesture Datasets

We employ several existing gesture datasets in this work: (1) the Mixed Multistroke Gesture (MMG) corpus [3]; (2) the Algebra Learner mathematics input corpus [4]; (3) the HHReco geometric shape dataset [16]; and (4) the NicIcon crisis management dataset [46]. Key characteristics of these datasets relevant to this work are given in Table 2. Three of the datasets (MMG, HHReco, and NicIcon) were collected from adult users entering individual gesture samples one at a time in a gesture collection tool. The fourth dataset (Algebra Learner) was collected from middle and high school users ($11-17$ years old) solving algebraic equations (later hand-segmented and labeled). In all, we employ four datasets containing gestures of 63 different types executed by 113 unique users, for a total of 40,305 executions.

[1] Normalization and resampling represent common preprocessing techniques employed by gesture recognizers in order to reduce gesture variability and increase classification accuracy [2, 3, 50].

Gesture Feature	Units	Computation
1. Geometric features (selected from Rubine [35], Long et al. [26], Willems et al. [46])		
Number of strokes	count	number of paired pen-down and pen-up events
Path length	pixels	cumulative sum of the Euclidean distance between adjacent points
Area of the bounding box	pixels^2	height $(y_{max} - y_{min})$ multiplied by width $(x_{max} - x_{min})$ of the bounding box
Cosine of starting angle	-	Rubine f_1 feature
Cosine of ending angle	-	similar to Rubine's f_1 but for the end of the gesture
Line similarity	-	distance between starting and ending points divided by path length
Global orientation	degrees	angle of the diagonal of the gesture bounding box (Rubine f_4)
Total turning angle	degrees	sum of the absolute value of the angles at each point (Rubine f_{10})
Sharpness	degrees	sum of the squared angles at each gesture point (Rubine f_{11})
Curviness	degrees / pixel	total turning angle divided by path length (Long et al. [26], feature 13)
2. Kinematic features		
Production time	ms	difference between t_{max} and t_{min}
Average speed	pixels / ms	path length divided by production time

Table 1: Gesture features employed during analysis.

Dataset	Users	Gestures	Multi strokes	Single strokes	Total samples	Max strokes	Gesture types
MMG [3]	20	16	87%	13%	3,200	5	arrowhead, asterisk, D, exclamation point, five-pointed star, H, half-note, I, line, N, null symbol, P, pitchfork, six-pointed star, T, X
Algebra [4]	40	20	30%	70%	15,309	2	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, x, y, a, b, c, +, -, =, (,)
HHReco [16]	19	13	60%	40%	7,791	9	arch, callout, crescent, cube, cylinder, ellipse, heart, hexagon, parallelogram, pentagon, square, trapezoid, triangle
NicIcon [46]	34	14	89%	11%	14,005	4	accident, bomb, car, casualty, electricity, fire, fire brigade, flood, gas, injury, paramedics, person, police, roadblock
Total	**113**	**63**			**40,305**		

Table 2: Properties of the four previously published datasets used in this work.

4 Gesture Consistency Findings

We found a high degree of within-user agreement (.91, SD $= .18$), but a lesser degree of consistency between users (.55, SD $= .31$). A Wilcoxon signed-rank test confirmed this difference is significant ($Z = -6.59$, $p < .001$, $N = 63$). This finding supports prior work in handwriting recognition [9, 39] and multitouch gestures [36] indicating that users are highly individual and internally consistent, but that there are also some stylistic "classes" across users that can be reliably consistent. To understand how agreement in gesture articulation manifests in gesture execution features, we correlated between- and within-user agreement rates per gesture type with the average values for each gesture feature per gesture type (Table 3). We present the remainder of the findings by examining how these gesture articulation features are relevant to the agreement rates.

4.1 Relationship of Agreement to Number of Samples

One might expect between-user consistency to depend on how many and which users the gestures come from. For example, a dataset consisting entirely of a small set of users of the same age, handedness, cultural background, etc., might yield 100% agreement for all gesture types. A dataset of many users of diverse cultures, languages, ages, etc., might have very low agreement rates between users. In our case with 113 users, we tested the relationship between number of users who drew a gesture type and the average between-user agreement rates and found a moderately strong positive correlation ($r = .322$, $p < .01$): the more people whose samples we have for a given gesture type, the more agreement we find. We do not have detailed demographic information available for all datasets, so we cannot confidently remark on diversity, but future work could examine differences among various cultures (e.g., [30]).

Another measure of the expected agreement one might achieve given a certain amount of gesture samples is the number of executions per person per gesture type. How many samples per person are needed to reach good coverage? In our data, there is no significant relationship between number of executions and within-user agreement, but there is a strong negative correlation with between-user agreement ($r = -.495$, $p < .01$). This finding indicates that, as the number of samples per user increases, the agreement decreases. We theorize that adding more executions per person may simply expose extra styles of drawing a gesture that are idiosyncratic to that user, lowering overall agreement with other users.

Feature	AR within-user	AR between-user
Number of strokes	$-.687^{**}$	$-.614^{**}$
Speed	$.530^{**}$	$.311^{*}$
Sharpness	$-.395^{**}$	$-.439^{**}$
Total turning angle	$-.375^{**}$	$-.436^{**}$
Line similarity	$.313^{*}$	$.627^{**}$
Path length	n.s.	$-.536^{**}$
Area of bounding box	n.s.	$-.470^{**}$
Production time	n.s.	$-.418^{**}$
Global orientation	n.s.	$.270^{*}$
Curviness	n.s.	$.301^{*}$
Start angle	n.s.	n.s.
End angle	n.s.	n.s.

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

NOTE: N=43 for production time and speed (the Algebra set does not include timestamps, N=20), N=63 for all other features.

Table 3: Pearson correlation coefficients between gesture features and agreement rate values. NOTE: features are listed in decreasing order of the within-users agreement correlation coefficient.

4.2 Relationship to Gesture Category and Familiarity

Space prevents us from showing a figure of the agreement rates for all 63 gesture types that appeared in the four datasets we examined. Instead, we show a frequency distribution of the percent of gesture types that had similar agreement values (Figure 4a). We find that 87% of gestures are above .75 within-user agreement, whereas only

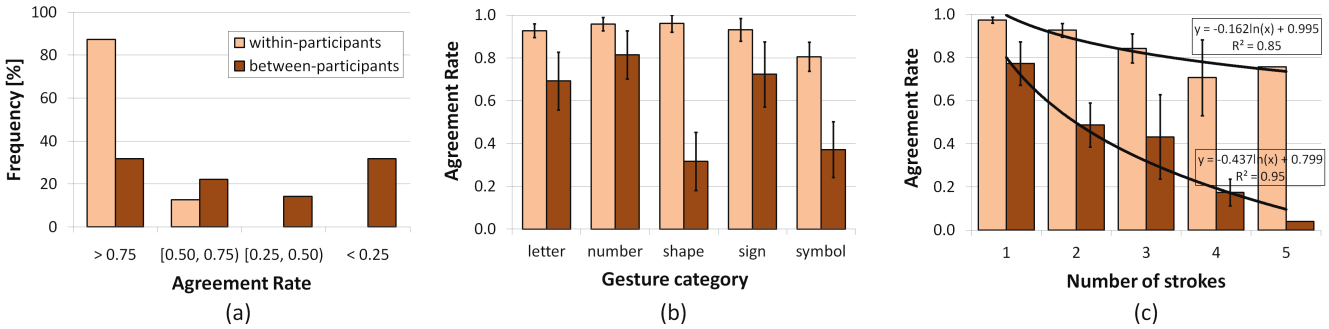


Figure 4: User consistency summary: (a) frequency distribution of proportion of gesture types exhibiting similar agreement rates; (b) agreement rate by gesture category; (c) loglinear relationship between number of strokes and agreement rates.

Gesture Category	
Letter	a, b, c, x, y, D, H, I, N, P, T, X
Number	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Shape	arch, cube, cylinder, ellipse, heart, hexagon, moon, parallelogram, pentagon, square, trapezoid, triangle, 5pt star, 6pt star, line
Symbol	callout, pitchfork, accident, bomb, car, casualty, electricity, fire, fire brigade, flood, gas, injury, paramedics, person, police, roadblock
Sign	equal, left-parenthesis, minus, plus, right-parenthesis, arrowhead, asterisk, exclamation, half note, null
Gesture Familiarity	
Familiar	a, b, c, x, y, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, equal, left-parenthesis, minus, plus, right-parenthesis, arrowhead, asterisk, exclamation, half note, line, null, D, H, I, N, P, T, X, arch, cube, cylinder, ellipse, heart, hexagon, moon, parallelogram, pentagon, square, trapezoid, triangle
Nonfamiliar	5pt star, 6pt star, pitchfork, callout, accident, bomb, car, casualty, electricity, fire, fire brigade, flood, gas, injury, paramedics, person, police, roadblock

Table 4: Gesture category and familiarity groupings.

32% of gestures are above .75 between-user agreement. There are no gesture types below .50 agreement within users, indicating that personal outliers are rare. Several gesture types had perfect (1.00) average within-user agreement (i.e., no variability in how they were executed for all individual users) including “2”, “arch”, “D”, “ellipse”, “heart”, “line”, “moon”, “P”, and “pentagon”. Gesture types “1”, “left-parenthesis”, “right-parenthesis”, “6”, “3”, and “exclamation point” were also above .96 within-user agreement. Only a few gesture types had perfect between-user agreement: “D”, “line”, “P”, “2”, “c”. All of these gestures are common gestures that users probably have written or drawn thousands of times in their lives, plausibly leading to a practice effect that increases agreement.

To examine this possibility, we considered both the category of each gesture type and its potential *familiarity* to the users who drew it. When we consider the types of gestures that were included in the datasets we investigated, several categories of gestures emerged: (a) *letters*, e.g., “a”, “b”, “c”; (b) *numbers*, e.g., “1”, “2”, “3”; (c) *shapes*, e.g., “square”, “ellipse”, “triangle”; (d) *symbols*, e.g., “callout”, “pitchfork”, “car”; and (e) *signs*, e.g., “plus”, “minus”, “equal” (Table 4). We found that agreement rates for letters and numbers were higher than for other gesture types, such as shapes and symbols (Figure 4b). Specifically, letters and numbers had .75 (SD = .22) between-user agreement; other gestures had only .44 (SD = .31), and this difference was confirmed significant by a Mann-Whitney U test ($U = 195.50$, $Z = -3.68$, $p < .0005$).

This relationship could be the result of explicit training to write letters and numbers in a certain way as part of penmanship practice in school (i.e., in which the manner of making the letters and numbers are part of teaching their form), or it could simply be the result of heavy practice of these types of gestures as compared to the other types. To explore this issue in more depth, we next grouped gesture types into ones we felt likely to have been practiced (at least in Western cultures) vs. ones that were not (Table 4). Indeed, we find that familiar gestures show higher agreement rates both within users (.95 practiced, .81 not) and between users (.62 practiced, .37 not). A Mann-Whitney U test confirmed a significant difference for between-users ($U = 212.00$, $Z = -2.94$, $p < .005$). This finding suggests that increased degree of comfort with a gesture decreases the variation in articulating that gesture.

4.3 Relationship to Gesture Entry over Time

As an approximation of the effect of practice, we attempted to measure the degree to which consistency might change over time within the course of one gesture collection session. We hypothesized that a learning effect could influence user consistency in gesture production patterns, as users articulate more samples of a gesture. As the number of executions increases for a gesture type, it is likely for the articulations of that gesture to converge to some “preferred” production patterns, which would make users seem less consistent at the beginning, but more consistent as they progress. To test this hypothesis, we sorted all samples for each user and gesture type in chronological order, split them half way, and computed agreement rates for each half. However, a Mann-Whitney U test did not reveal any significant differences for either within- (.88 vs. .90, $U = 515000.500$, $Z = -1.894$, n.s.), nor between-user agreement rates (.44 vs. .44, $U = 919.500$, $Z = -0.043$, n.s.). This finding shows that the users from our sets exhibited the same level of consistency in their articulation patterns from their initial execution to the last one, making them equally consistent over time. This finding seems to hold for blocks of up to 15 gesture executions (as the maximum number of samples for a gesture was 30 for the HHReco and NicIcon sets), but may change with more practice over the course of a lifetime (i.e., thousands of executions).

4.4 Relationship to Geometric Complexity

We consider a number of our features as indicators of *geometric complexity*: number of strokes, total turning angle, line similarity, and sharpness. If we examine how much agreement there is among users with respect to number of strokes they generate while drawing gestures, regression analysis shows a logarithmic model as the best fit, with the following loglinear relationships (Figure 4c):

(a) within-users: $y = -0.162 \cdot ln(x) + 0.995$ ($R^2 = .85$)

(b) between-users: $y = -0.437 \cdot ln(x) + 0.799$ ($R^2 = .95$)

The relationship has a negative coefficient, meaning that, as number of strokes increases, agreement rates decrease. Prior work in human visual perception [7] has found that perceived similarity of objects is typically correlated with the logarithm of quantitative measurements of those objects. It is therefore interesting that the reverse relationship also seems to hold (at least for number of strokes): humans tend to draw objects that are visually similar in a logarithmic relationship. The high degree of fit for these functions is encouraging; we may be able to use them as predictors of expected agreement on a candidate gesture given the expected number of strokes, which would help designers choose good gestures.

Another potentially interesting measure of complexity is the total turning angle of the gesture. If a gesture passes through many curves and wiggles (such as a "g") during its path, is it likely to have higher or lower agreement rates among users executing that gesture type? Such gesture types tend to give gesture recognizers more trouble, especially template matchers such as $N [2, 3], and the reason could be decreased user consistency. We can explore this relationship by first computing the average total turning angle per gesture type of all the gestures in the four datasets we examined, which ranges from $min = 148.8°$ (for "line" gestures) to $max = 2861.7°$ ("6pt star"), SD $= 462.3°$. Indeed, there was a moderate negative correlation between the total turning angle and within-user agreement rate ($r = -.375$, $p < .01$) and a strong negative correlation with between-user agreement rate ($r = -.436$, $p < .01$). Both correlations are negative, meaning that agreement rates decrease as the total turning angle increases. This result is consistent with the independent measure of complexity mentioned above, number of strokes; in both cases, increased geometric complexity leads to lower user consistency, even within a single user.

The other two features we identify as relevant to geometric complexity, line similarity and sharpness, exhibit similar relationships to agreement. The moderate and strong *positive* correlations to agreement that line similarity shows (Table 3), and the moderate and strong *negative* correlations that sharpness shows, both indicate that lower complexity is related to higher agreement.

4.5 Relationship to Kinematics

When we examined the distribution of agreement rates earlier in the paper, we noted that there are no gesture types with below .50 within-user agreement, indicating fairly consistent within-user behavior. We wondered if this consistency could still be expected if users are rushed or otherwise distracted, and we examined this possibility through the production time and average speed features in our data. Unexpectedly, we found a strong *positive* relationship between speed and within-user agreement ($r = .530$, $p < .01$). This result indicates that faster gesture entry does not co-occur with atypical gesture articulation; in fact, the opposite is true. Supporting this finding is prior work on the $1 gesture recognizer [50], which found that faster gestures were better recognized, indicating higher consistency when rushed. (No significant correlation was found between production time and within-user agreement, likely because quicker gestures could also be caused by less complex gestures, already shown to be correlated to agreement.)

In addition, we found a strong negative relationship between production time and between-user agreement ($r = -.418$, $p < .01$) and a moderate positive relationship between speed and between-user agreement ($r = .311$, $p < .05$). These results continue to indicate that, as users enter their gestures faster (and with shorter durations), agreement actually increases, even between users. This result could be a factor of confidence: users possibly draw gestures faster when they feel more comfortable with them. Comfort level could come from repeated practice of the same gestures (e.g., letters and numbers are commonly written), and so is related to our earlier findings regarding gesture category and familiarity as well.

4.6 Relationship to Gesture Size

Gesture length and area, as indicators of articulation size, are negatively correlated with between-user agreement ($r = -.536$, and $r = -.470$, $p < .01$) but not significantly correlated with within-user data. Smaller gestures have higher agreement, most likely because less variation is possible kinematically in smaller motions.

5 IMPLICATIONS

The implications of these findings apply to the design of application gesture sets and inform the structure of recognizer training sets. Also, new classification rules working on top of existing gesture recognizers can be designed based on the results of this study, for example, to prune training sets and to improve performance for specific gesture sets. We list potential improvements that can be implemented based on our findings of the present study as a set of guidelines (a-g) for practitioners, and we accompany each specific guideline with practical examples:

❶ **Application gesture set design**.

(a) Where possible, prefer unistroke gestures (e.g., commands examined by [50]) as their execution is more consistent for both individual users (.97) and between users (.77) than the execution of multistroke gestures.

(b) Respect emerging standards and/or prefer gestures already likely to be familiar to users (e.g., letters, shapes, numbers in general; "pigtail" to select and "cross" to delete, in specific).

(c) In spite of a need to create gestures that are fairly distinctive [25], avoid introducing gestures with too high a degree of geometric complexity (e.g., large total turning angle).

✍ **Example**: Applying criteria (a-c) for the MMG set of gestures [3], we would suggest pruning the "6pt star" gesture, which had the lowest agreement rate from the set, .73 for within-user and .21 for between-users, and also the second lowest recognition rate of 16 gestures, just 93% [3] (p. 120). Applying these principles for the $1 gesture set [50], we would suggest pruning the curly braces, which also correlates with users rating them poorly: 2 on a 5-point Likert scale [50] (p. 166), the lowest rating out of 16 gestures. The curly braces also had the highest recognition error (1.67%) out of all gestures of the $1 set [50] (Table 1, p. 166).

❷ **Training set design**.

(d) Collect more training samples for gestures that appear to be less consistent (e.g., more strokes) in order to cover more of their variability within the training set.

(e) Prune the large training sets needed by some recognizers [2, 3], removing unlikely articulations of multistroke gestures.

✍ **Example**: The most important implication of guideline (d) is that the number of training samples per gesture type does not need to be the same for all gestures in the set. This is a simple consequence of our study but no one has actually examined this option before. Instead, the existing practice of testing recognition performance of gesture recognizers has only considered equal sampling for all gesture types [2, 3, 24, 50].

✍ **Example**: Recognizers such as $N [2, 3], which represents all possible permutations of a given gesture to keep user training costs down, could use guideline (e) to prune the available set of permutations (or mark some as less likely) once a particular user enters a few samples. New samples from the same user are not likely to deviate much from this core pruned set, and we have already noted that writer-dependent recognition is more accurate [38]. We also noted that the effect of familiarity with certain gesture types and categories is strong; previously practiced letters, numbers, and signs (e.g., handwriting gestures) show much higher between-user agreement rates. These gesture types are then candidates for much more aggressive pruning of the possible gesture articulation space when designing recognizers for them.

❸ **Design of supporting classification rules**.

(f) Use simple rejection rules to assist recognizers in discriminating between confusable gesture types with close confidence scores [2, 3, 50].

(g) Exploit differences in gesture articulation to allow multiple commands to use the same gesture with different articulations.

✍ **Example**: The lower between-user agreement rates (.55) and the negative correlation between number of executions and between-user agreement rates indicate it may be more difficult to prune the gesture space for a multiuser system without cutting gestures that matter. However, we can use the analysis of the gesture features and how they relate to agreement in order to prune more precisely. Simple rules based on guideline (f) can be devised to improve recognition performance on gesture classes with high degrees of confusion, such as: (1) if the candidate gesture has 3+ strokes then it can't be of type X, Y, or Z because these gesture types are never produced with more than two strokes; (2) if the turning angle of the candidate is larger than a given threshold, then it can't be of type X or Y because these gesture types consist of simple lines only.

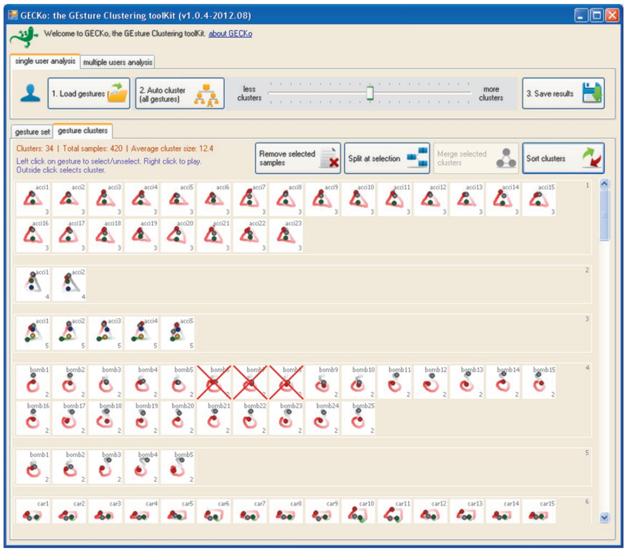
✍ **Example**: One challenge of gesture set design is that users often desire to use the same (or similar) gesture for multiple commands [25], but these are difficult for recognizers to distinguish. Therefore, gesture sets must be designed considering the impact of each individual gesture type on the others with respect to possible recognition confusions [25]. Expert designers can consider gesture articulation differences to support guideline (f), for instance, allowing users to draw a clockwise circle to select items and a counterclockwise circle to delete items, assuming the recognizer being used can distinguish between them.

5.1 Impact on Research and Practice

Besides the above, many other applications of our findings can be imagined for improving the performance of today's gesture recognition techniques and the design practice of gesture sets for applications. Furthermore, probably the most important implication of this work is to draw the community's attention towards the amount of (not-before-measured) variation in articulating stroke gestures. We believe that reporting such findings to the community, while backed up by the largest experiment ever conducted on gesture input behavior (40,305 gestures), represents a solid starting point for extended investigation into how users articulate gestures and how that can be exploited for the design of future surface gesture interfaces. We look forward to seeing how our findings will be exploited by practitioners and how other researchers will make use of this extensive dataset on gesture articulation patterns and agreement rates that we have generated. To this end, we release the logs of our manually clustered gesture sets[2].

5.2 The GECKo Tool

We mentioned the GEsture Clustering toolKit (GECKo) that we implemented to assist a human operator in editing and correcting the cluster partition initially generated by an automated clustering procedure. We devised several visualization strategies for gesture articulations to make such editing easier. We briefly summarize them here, as we believe they could prove useful to implement in other applications that need to display stroke gestures as well. The goal was to increase the visual similarity of gestures exhibiting the same number of strokes, stroke directions, and stroke orderings, and to visually highlight dissimilarity in any of these factors. We found the following visualization techniques effective: (1) display strokes in different colors following a fixed color scheme (e.g., first stroke is always displayed in red, second stroke gray, and so on); (2) highlight the starting point of each stroke with a small disc; (3) fade stroke color in the direction of articulation; (4) display the total number of strokes next to the gesture image. For cases in which similarity could not be assessed visually, the human operator could play an animation of the gesture execution. GECKo reports within- and between-user agreement rates after clustering. In the spirit of providing open toolkits [5, 6, 25], we provide GECKo as a free tool, available for public download (see above).


GECKo: the GEsture Clustering toolKit (v1.0.4-2012.08)

Welcome to GECKo, the GEsture Clustering toolKit. about GECKo

Clusters: 34 | Total samples: 420 | Average cluster size: 12.4

Left click on gesture to select/unselect. Right click to play.
Outside click selects cluster.

Figure 5: GECKo: the gesture clustering toolkit.

6 Conclusion

We report in this paper a methodology to analyze user consistency in touch and surface gesture execution, focusing on (a) gesture articulation described in terms of stroke number, ordering, and direction, and (b) execution variation measured by geometric and kinematic gesture features. We report the results of applying this methodology to four previously published datasets from different domains (40,305 samples of 63 gesture types by 113 users). We found a high degree of consistency within users (.91), lower consistency rates between users (.55), higher rates of consistency for certain gesture types, and a loglinear relationship between number of strokes and consistency. We use our findings to propose a set of guidelines for helping designers of gesture interfaces to improve their gesture sets and recognizers.

We note that 40,305 gestures is the largest experiment on gesture input behavior ever conducted. We generated a large quantity of data and observations which are easily replicable (by virtue of using public data sets), together with delivering the actual techniques and tools to obtain them (i.e., agreement rate analysis of gesture clusters obtainable via the GECKo toolkit). By doing so, we not only confirm, formally and for the first time, expected user behavior in producing gesture shapes (e.g., positive correlation between user consistency and gesture complexity, or users being highly individual and internally consistent), but we also highlight new findings (e.g., a loglinear relationship between user consistency and number of gesture strokes). We have also developed a method to visualize gestures and gesture clusters in the GECKo toolkit, which will be useful for gesture designers to explore variability in their gesture sets, given some initial gesture data. We plan to add new features to GECKo to assist in designing gesture sets through prediction of gesture articulation patterns informed by this work. In the end, we believe that this work lays the foundation for further investigation into how users articulate gestures and how these findings can be exploited in the design of future surface gesture interfaces. Contributions of this work will lead to advanced gesture recognizers and adaptable gesture set designs that capitalize upon observed user behavior and preferred gesture articulation patterns.

[2] http://depts.washington.edu/aimgroup/proj/dollar/gecko.html

REFERENCES

[1] L. Anthony, Q. Brown, J. Nias, B. Tate, and S. Mohan. Interaction and recognition challenges in interpreting children's touch and gesture input on mobile devices. In *Proc. of ITS '12*, pages 225–234, New York, NY, USA, 2012. ACM.

[2] L. Anthony and J. O. Wobbrock. A lightweight multistroke recognizer for user interface prototypes. In *Proc. of GI '10*, pages 245–252, Toronto, Ont., Canada, 2010. Canadian Information Processing Soc.

[3] L. Anthony and J. O. Wobbrock. $N-Protractor: a fast and accurate multistroke recognizer. In *Proc. of GI '12*, pages 117–120, Toronto, Ont., Canada, 2012. Canadian Information Processing Soc.

[4] L. Anthony, J. Yang, and K. Koedinger. A paradigm for handwriting-based intelligent tutors. *International Journal of Human-Computer Studies*, 70(11):866–887, 2012.

[5] C. Appert and S. Zhai. Using strokes as command shortcuts: cognitive benefits and toolkit support. In *Proc. of CHI '09*, pages 2289–2298, New York, NY, USA, 2009. ACM.

[6] D. Ashbrook and T. Starner. MAGIC: a motion gesture design tool. In *Proc. of CHI '10*, pages 2159–2168, New York, NY, USA, 2010.

[7] F. Attneave. Dimensions of similarity. *American Journal of Psychology*, 63(4):516–556, 1950.

[8] S. Chatty and P. Lecoanet. Pen computing for air traffic control. In *Proc. of CHI '96*, pages 87–94, New York, NY, USA, 1996. ACM.

[9] J.-P. Crettez. A set of handwriting families: style recognition. In *Proc. of ICDAR '95*, pages 489–494, Washington, DC, USA, 1995. IEEE Computer Society.

[10] M. Djioua and R. Plamondon. Studying the variability of handwriting patterns using the kinematic theory. *Human Movement Science*, 28(5):588–601, 2009.

[11] C. Frankish, R. Hull, and P. Morgan. Recognition accuracy and user acceptance of pen interfaces. In *Proc. of CHI '95*, pages 503–510, New York, NY, USA, 1995. ACM.

[12] J. Galbally, J. Fierrez, J. Ortega-Garcia, and R. Plamondon. Synthetic on-line signature generation. Part II: Experimental validation. *Pattern Recognition*, 45(7):2622–2632, 2012.

[13] J. Galbally, R. Plamondon, J. Fierrez, and J. Ortega-Garcia. Synthetic on-line signature generation. Part I: Methodology and algorithms. *Pattern Recognition*, 45(7):2610–2621, 2012.

[14] D. Goldberg and C. Richardson. Touch-typing with a stylus. In *Proc. of CHI '93*, pages 80–87, New York, NY, USA, 1993. ACM.

[15] T. Hammond and B. Paulson. Recognizing sketched multistroke primitives. *ACM Trans. Interact. Intell. Syst.*, 1(1):4:1–4:34, Oct. 2011.

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

[17] P. Isokoski. Model for unistroke writing time. In *Proc. of CHI '01*, pages 357–364, New York, NY, USA, 2001. ACM.

[18] S. K. Kane, J. O. Wobbrock, and R. E. Ladner. Usable gestures for blind people: understanding preference and performance. In *Proc. of CHI '11*, pages 413–422, New York, NY, USA, 2011. ACM.

[19] C. Keskin, A. T. Cemgil, and L. Akarun. DTW based clustering to improve hand gesture recognition. In *Proc. of HBU '11*, pages 72–81, Berlin, Heidelberg, 2011. Springer-Verlag.

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

[21] F. Lacquaniti, C. Terzuolo, and P. Viviani. The law relating the kinematic and figural aspects of drawing movements. *Acta Psychologica*, 54(1-3):115–130, 1983.

[22] M. J. LaLomia and K. C. Cohen. Gesture consistency for text, spreadsheet, graphic and form fill editing. *SIGCHI Bull.*, 23(4):40–41, 1991.

[23] Y. Li. Gesture search: a tool for fast mobile data access. In *Proc. of UIST '10*, pages 87–96, New York, NY, USA, 2010. ACM.

[24] Y. Li. Protractor: a fast and accurate gesture recognizer. In *Proc. of CHI '10*, pages 2169–2172, New York, NY, USA, 2010. ACM.

[25] A. C. Long, Jr., J. A. Landay, and L. A. Rowe. Implications for a gesture design tool. In *Proc. of CHI '99*, pages 40–47, New York, NY, USA, 1999. ACM.

[26] A. C. Long, Jr., J. A. Landay, L. A. Rowe, and J. Michiels. Visual similarity of pen gestures. In *Proc. of CHI '00*, pages 360–367, New York, NY, USA, 2000. ACM.

[27] G. Lorette. Handwriting recognition or reading? What is the situation at the dawn of the 3rd millenium? *International Journal on Document Analysis and Recognition*, 2(1):2–12, 1999.

[28] H. Lü and Y. Li. Gesture avatar: a technique for operating mobile user interfaces using gestures. In *Proc. of CHI '11*, pages 207–216, New York, NY, USA, 2011. ACM.

[29] V. Märgner and H. E. Abed. ICFHR 2010 - arabic handwriting recognition competition. In *Proc. of ICFHR '10*, pages 709–714, Washington, DC, USA, 2010. IEEE Computer Society.

[30] D. Mauney, J. Howarth, A. Wirtanen, and M. Capra. Cultural similarities and differences in user-defined gestures for touchscreen user interfaces. In *CHI '10 Extended Abstracts*, pages 4015–4020, New York, NY, USA, 2010. ACM.

[31] M. R. Morris, J. O. Wobbrock, and A. D. Wilson. Understanding users' preferences for surface gestures. In *Proc. of GI '10*, pages 261–268, Toronto, Canada, 2010. Canadian Information Processing Soc.

[32] C. Myers and L. Rabiner. A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60(7):1389–1409, 1981.

[33] T. Ouyang and Y. Li. Bootstrapping personal gesture shortcuts with the wisdom of the crowd and handwriting recognition. In *Proc. of CHI '12*, pages 2895–2904, New York, NY, USA, 2012. ACM.

[34] R. Plamondon, C. O'Reilly, J. Galbally, A. Almaksour, and E. Anquetil. Recent developments in the study of rapid human movements with the kinematic theory: Applications to handwriting and signature synthesis. *Pattern Recognition Letters*, 2012 (to appear).

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

[36] M. Schmidt and G. Weber. Enhancing single touch gesture classifiers to multitouch support. In *Proc. of ICCHP '10*, pages 490–497, Berlin, Heidelberg, 2010. Springer-Verlag.

[37] T. Sezgin and R. Davis. Sketch interpretation using multiscale models of temporal patterns. *IEEE Computer Graphics and Applications*, 27(1):28–37, 2007.

[38] N. K. Smithies, S. and J. Arvo. Equation entry and editing via handwriting and gesture recognition. *Behaviour & Information Technology*, 20(1):53–67, 2001.

[39] S. Srihari, S.-H. Cha, H. Arora, and S. Lee. Individuality of handwriting: A validation study. In *Proc. of ICDAR '01*, pages 106–109, Washington, DC, USA, 2001. IEEE Computer Society.

[40] H. Tu, X. Ren, and S. Zhai. A comparative evaluation of finger and pen stroke gestures. In *Proc. of CHI '12*, pages 1287–1296, New York, NY, USA, 2012. ACM.

[41] R.-D. Vatavu. User-defined gestures for free-hand TV control. In *Proc. of EuroiTV '12*, pages 45–48, New York, NY, USA, 2012. ACM.

[42] R.-D. Vatavu. A comparative study of user-defined handheld vs. freehand gestures for home entertainment environments. *Journal of Ambient Intelligence and Smart Environments*, 5(2):187–211, 2013.

[43] R.-D. Vatavu, L. Anthony, and J. O. Wobbrock. Gestures as point clouds: a $p recognizer for user interface prototypes. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 273–280, New York, NY, USA, 2012. ACM.

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

[45] A. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, West Sussex, England, 2003.

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

[47] J. Williamson and R. Murray-Smith. Rewarding the original: explorations in joint user-sensor motion spaces. In *Proc. of CHI '12*, pages 1717–1726, New York, NY, USA, 2012. ACM.

[48] J. O. Wobbrock, H. H. Aung, B. Rothrock, and B. A. Myers. Maximizing the guessability of symbolic input. In *CHI '05 Extended Abstracts*, pages 1869–1872, New York, NY, USA, 2005. ACM.

[49] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *Proc. of CHI '09*, pages 1083–1092, New York, NY, USA, 2009. ACM.

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