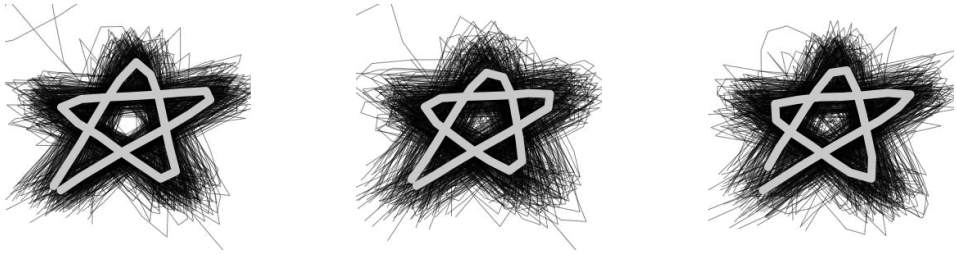
Large-Scale User Perception of Synthetic Stroke Gestures

Luis A. Leiva
Sciling | Valencia, Spain
name@sciling.com



(a) (b) (c)

Figure 1. Examples of human and synthetic gestures. Can you guess which are which? See the answer at the end of this paper.

ABSTRACT

Researchers are increasingly being concerned with the resemblance of synthetic gestures; i.e., how human-like they are, as perceived by end users. However, evaluations in this regard have been scarce and/or inconclusive. In this paper, we compared stroke gestures produced by two modern synthesizing techniques (GPSR and G3) against the same gestures produced by humans. We conducted an online study involving 623 participants, who provided binary assessments for near 6K gesture images. We found that it is difficult to tell human and synthetic gestures apart, but also that gestures synthesized with G3 are perceived as if they were human-generated more often than those synthesized with GPSR. Our results enable a deeper understanding of synthetic gestures' production, which can inform the design of gesture interaction.

Author Keywords

Gesture Synthesis; Bootstrapping; Sigma-Lognormal Model; Gesture Path Stochastic Resampling; Rapid Prototyping

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces; I.5.2 Pattern Recognition: Design Methodology

INTRODUCTION

Gesture interaction is a longstanding research area in HCI, evidenced by early works from 1960 like the Sketchpad project [28] and the RAND tablet [7]. Today, stroke gestures (also known as pen gestures, touch gestures, or hand markings) are becoming more and more relevant to mainstream products such as touchscreen-capable devices like smartphones and tablets. Compared to traditional interactions, stroke gestures have the potential to lower cognitive load and the need for visual attention [5, 33]. Stroke gestures also may improve the usability of UIs, by replacing standard shortcuts by more accessible triggers [15, 18].

Training a high-quality gesture recognizer requires providing a large number of examples to enable good performance on unseen, future data. However, recruiting participants, data collection and labeling, etc. necessary for achieving this goal are usually time-consuming and expensive [1, 14]. Previous works have proposed to address this problem by generating synthetic samples [2, 8, 12, 27]. Modern synthesizing techniques like G3 [16] and GPSR [29] have demonstrated that training gesture recognizers with synthetic data generated from real users can significantly improve recognition accuracy. However, researchers have seldom evaluated the quality of gesture synthesis from the perspective of how "human-like" they really are. It is crucial that synthetic gestures have a realistic appearance, not only for display purposes, but because severely deformed synthetic samples may lead to poor recognizer performance [29].

In this paper, we investigate the user perception toward unistroke and multistroke gestures synthesized with GPSR and G3 vs. actual human gestures. We conducted an online study "in the wild", involving 623 participants who provided binary assessments for near 6K gesture images. We found that it is difficult to tell human and synthetic gestures apart, but also that gestures synthesized with G3 are perceived as if they were human-generated more often than those synthesized with GPSR. Researchers and practitioners can finally be confident that synthetic stroke gestures are actually reflective of how users perceive real gestures. Taken together, our results enable a deeper understanding of synthetic gestures' production, which can inform the design of gesture interaction by (1) automatically augmenting current gesture sets with more human-like samples and, in consequence, (2) building more accurate gesture recognizers.

DIS 2017, June 10–14, 2017, Edinburgh, United Kingdom

ISBN 978-1-4503-4922-2/17/06…$15.00

DOI: http://dx.doi.org/10.1145/3064663.3064724

RELATED RESEARCH

Approaches to gesture synthesis mainly fit into two categories [29]: those that replicate stroke features from an existing dataset and those that apply perturbations to a given gesture sample. Approaches in the former category are unsuitable for rapid prototyping as they require a reasonably large amount of data to begin with as well as advanced knowledge in machine learning. The latter category aims for lightweight, easy to understand and ready-to-use approaches that address common UI demands. There is a third option involving an interactive approach, such as in Gesture Script [20], where developers describe the gesture structure and its parts. However, having to provide too detailed information for each gesture can be time-consuming. This is why we focus on perturbation-based approaches to gesture synthesis.

Most relevant to this work, G3 [16, 21] produces synthetic stroke gestures by means of the Kinematic Theory [24] and its associated $\Sigma\Lambda$ model [25]. Concretely, G3 creates a kinematic model of a user-provided gesture sample and introduces local and global perturbations to the model parameters. Then, the perturbed models are used to generate synthetic gestures that in turn look realistic, as shown in a study like the one we conducted in this paper but at a smaller scale [17].

Also relevant to this work, GPSR [29] produces synthetic stroke gestures by lengthening and shortening gesture subpaths within a given sample to produce realistic variations via stochastic (nonuniform) resampling. A preliminary study revealed significances between GPSR and a custom implementation of G3, but not between GPSR and real samples. It is unclear thus if these results will hold the same at a large scale, like the study we conducted in this paper.

So far, there is little work that has advanced our knowledge of how users perceive synthetic stroke gestures. Galbally et al. [11] examined the human likeness of synthetic handwritten signatures (25 participants), Leiva et al. [17] replicated that study with stroke gestures (236 participants), and Taranta et al. [29] conducted an informal evaluation (unknown number of participants, presumably small). All these previous works found a high degree of similarity between synthesized and human samples, but it is unclear which synthesizing technique actually performs better. Therefore, an evaluation like the one conducted in this paper has been missing.

EVALUATION

We conducted an online study to assess the subjective perception that non-expert human observers have of unistroke and multistroke gestures.

Datasets

We used two well-known public datasets in HCI that are also available in synthesized forms [17, 29].

GDS: Comprises 16 *unistroke* gesture classes, 5,280 samples in total [32]. Ten users provided 10 samples per class at 3 articulation speeds (slow, medium, fast) using an iPAQ Pocket PC (stylus as input device).

MMG: Comprises 16 *multistroke* gesture classes, 9,600 samples in total [4]. Twenty users provided 10 samples per class at 3 articulation speeds (same as in GDS) using either finger (half of the users) or stylus as input device on a Tablet PC.

The set of MMG + GDS gestures, produced by humans and by the two synthesizing methods, was made available online at https://g3.prhlt.upv.es/guessit/.

Participants

The study was advertised in social networks, ensuring that users had no expert knowledge on gesture recognition; e.g., we did not survey special interest groups on gesture interaction or machine learning. In order to increase the chance of participation, we did not collect explicit demographic information and participants were incentivized with a comparison of their results to others, via social sharing buttons at the end of the study.

We report on data collected between November 2016 and December 2016, which attracted 6,035 pageviews. During this time, 623 volunteers from 30 different countries took part in the online study. Most of the participants came from Europe (92%) and the Americas (5.1%), and used either a desktop PC (48%), a mobile device (45%), or a tablet (7%). The study was completed between 30 seconds and a minute on average.

Procedure

This study aimed for recreating as much as possible the settings used by previous works [11, 17]. Each user was presented with 10 gesture images (1 image at a time) drawn at random among all gesture samples available. The user had to click on a button to indicate whether the gesture shown was human or artificial; see Figure 2. The maximum time permitted to assess each gesture was 4 seconds at most. This was so because the overall objective of this study was not making a detailed and profound analysis of each gesture, but estimating the general visual appearance of gesture samples after a short inspection. A "Skip this guess" button allowed the user to discard the current gesture shown and load a different one. Users could take the study more than once, if desired (only 22 users did it).

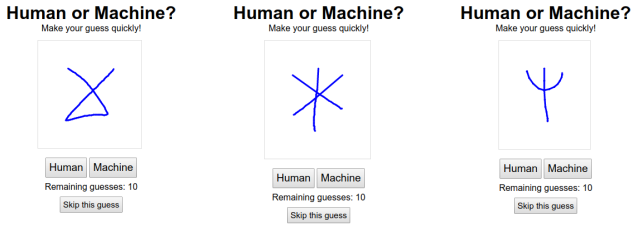


Figure 2. Screenshots of the study application.

Design

We considered 4 independent variables (factors) for analysis:

1. *Producer*, with 3 levels: Human, GPSR, G3;
2. *Dataset*, with 2 levels: GDS, MMG;
3. *Speed*, with 3 levels: slow, medium, fast;
4. *Device*, with 2 levels: stylus, finger.

The dependent variable considered is *Guessing Accuracy*: the user's capability to distinguish between human and synthetic gestures, i.e., the success classification rate.

The study is a repeated measures within-subjects design, since the same user makes a number of randomized guesses while being exposed to potentially any treatment level of every factor combination. The data were analyzed using a generalized linear mixed model for binomial proportions as omnibus test, which combines the advantages of ordinary logit models (akin the ANOVA test) with the ability to account for random subject and item effects [6].

RESULTS

The omnibus test to assess the main effects and interactions present in our data revealed a significant main effect regarding *Producer* $\left[\chi^2_{(2,N=4457)} = 26.11,\ p < .001\right]$. In addition, a significant interaction was found for *Producer*Dataset* $\left[\chi^2_{(2,N=4457)} = 8.80,\ p = .012\right]$ and *Producer*Device* $\left[\chi^2_{(2,N=4457)} = 21.81,\ p < .001\right]$. No other main effects or high-order interactions were significant. We therefore split the data by producer type and performed pairwise comparisons using the Wilcoxon rank sum test as post-hoc test of significance, with appropriately adjusted significance levels to guard against the risk of over-testing the data. All post-hoc tests used the Bonferroni correction.

Table 1 and Figure 3 summarize the overall guessing accuracy for each producer type. Notice that the distributions are non-Gaussian, as they represent the proportion of average successes in a series of independent binary assessments.

Human (N=1,937)		**GPSR** (N=2,032)		**G3** (N=1,989)	
54.53	[52.1, 56.9]	50.02	[47.7, 52.4]	44.92	[42.6, 47.3]

Table 1. Overall guessing accuracy (in %, mean and 95% confidence intervals). N = 5,958 gestures in total.

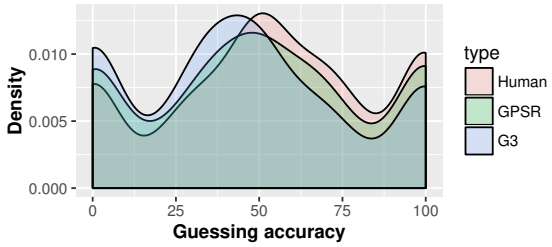


Figure 3. Distribution of guessing accuracy by producer type, averaged across all subjects.

Derived from the data above we can see that (i) real gestures were marked as synthetic $100 - 54.53 = 45.4\%$ of the time; (ii) GPSR gestures were mistaken with a real sample 49.9% of the time; and (iii) G3 gestures were mistaken with a real sample 55.1% of the time. Simply put, it is not easy to distinguish one type of gestures over the other, suggesting thus that synthetic samples have actual human-like appearance.

To better understand the differences between the three producers, we ran the pairwise comparisons and observed statistically significant differences in all cases: human vs. GPSR ($p < .01$), human vs. G3 ($p < .001$), and GPSR vs. G3 ($p < .01$). These results reveal that users performed differently when they were presented with a human or a synthetic gesture sample; i.e., users tended to guess correctly human samples more often, but also tended to classify synthetic samples as human. This is particularly true for G3, which had the lower guessing accuracy rate, and is further supported by the median guessing accuracies, which were *Mdn*=50% for both human and GPSR samples, and *Mdn*=33% for G3 samples. In sum, users committed more errors when assessing artificial samples, and particularly more when assessing gestures generated with G3.

Dataset analysis

We also investigated the impact of dataset (unistroke vs. multistroke gestures), given the interaction effect observed in the omnibus test. Table 2 summarizes the results. The post-hoc test on the GDS dataset (unistrokes) revealed statistically significant differences between synthetic and human samples ($p < .01$) but no differences were found between GPSR and G3 ($p = .32, n.s.$). Next, the post-hoc test on the MMG dataset (multistrokes) revealed statistically significant differences between G3 and the other approaches ($p < .01$).

Dataset	**Human** (N=985)		**GPSR** (N=1,008)		**G3** (N=985)	
GDS	55.77	[52.3, 59.2]	45.59	[42.2, 49.0]	43.15	[39.8, 46.5]
MMG	53.47	[50.1, 56.8]	53.88	[50.6, 57.1]	46.39	[43.1, 49.6]

Table 2. Guessing accuracy (in %, mean and 95% CIs) on the GDS dataset (N = 2,978 unistrokes) and MMG (N = 2,980 multistrokes).

This analysis suggests that both synthesizing techniques succeed at producing realistic unistroke gestures, however G3 seems more appropriate for rendering multistrokes, as it had much lower guessing accuracy. In general, multistroke gestures are more complex in nature. There are spatial constraints between consecutive strokes that sometimes are hard to represent accurately; see e.g. the "asterisk" or "six point star" samples in Figure 4.

Input device analysis

We also investigated the impact of input device (stylus vs. finger gestures) over the MMG dataset,[1] given the interaction effect observed in the omnibus test. Table 3 summarizes the results. The post-hoc test on the stylus gestures revealed statistically significant differences between human samples and both synthesizing techniques ($p < .001$), and no differences were found between GPSR and G3 ($p = .126, n.s.$). Next, the post-hoc test on the finger gestures revealed statistically significant differences between GPSR and the other approaches ($p < .01$).

Device	**Human** (N=952)		**GPSR** (N=1,024)		**G3** (N=1,004)	
Stylus	59.76	[55.2, 64.3]	49.74	[45.1, 54.4]	44.75	[40.2, 49.3]
Finger	46.68	[41.8, 51.5]	57.80	[53.3, 62.3]	48.17	[43.5, 52.9]

Table 3. Guessing accuracy (in %, mean and 95% CIs) of gestures drawn with stylus (N = 1,509) and finger (N = 1,471).

This analysis suggests that both synthesizing techniques succeed at producing gestures drawn with a stylus. However, for finger gestures G3 was on par with human samples, and both of which outperformed GPSR.

[1] GDS gestures were drawn with a stylus, so they were excluded.

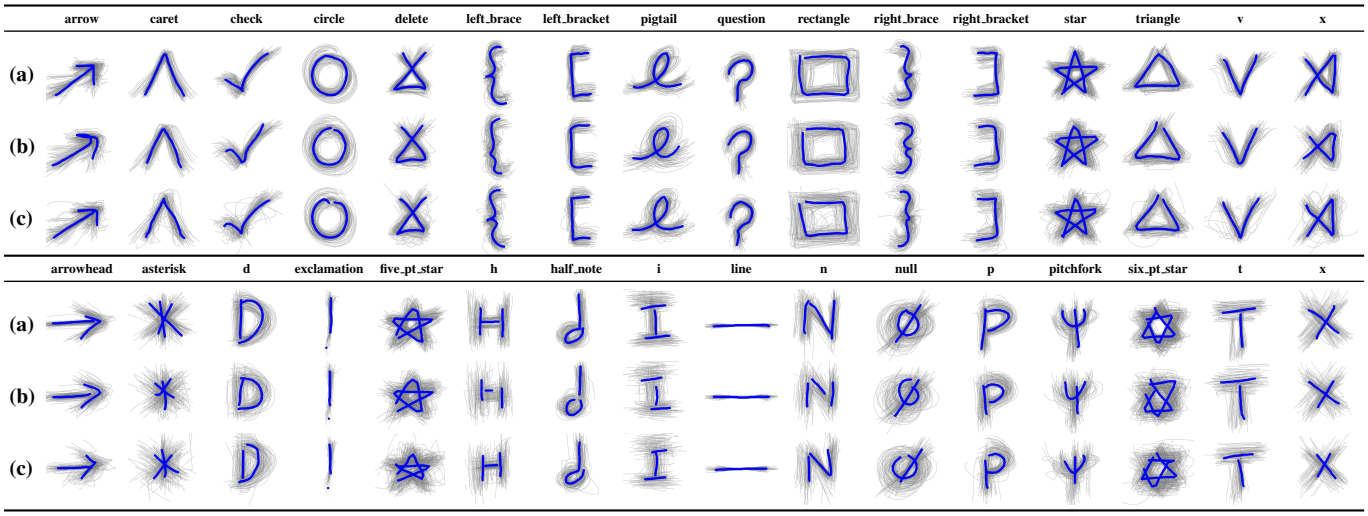


Figure 4. Gestures samples from GDS (top subtable) and MMG (bottom subtable). Each gesture group (a,b,c) is labeled the same way as in Figure 1.

Discussion

Overall, participants struggled to tell human and synthetic gestures apart. We conclude therefore that the visual appearance of the synthetic samples is very similar and close to that of human gestures, evidencing thus that they "look and feel" the same to human observers. This is an important result because recent literature on gesture analysis has shown how users produce different gesture articulations in various conditions [3, 13, 31]. Further, being able to produce synthetic gestures with a realistic appearance not only fulfills display purposes, but also improves recognition performance [29].

Previous user studies on stroke gestures perception [11, 17, 29] revealed that the number of mistaken real and synthetic samples was very similar. This paper advances our knowledge further, revealing that, at a large scale, both GPSR and G3 produce compelling results in terms of unistroke gestures, and that G3 delivers better results than GPSR in terms of multistrokes. We also found an effect of input device on guessing accuracy, suggesting that both synthesizing techniques perform similarly for stylus gestures and that G3 outperforms GPSR for finger gestures.

The synthesizing approaches studied in this paper have been proven to be strong contenders in the research literature. GPSR introduces reasonable perturbations on a stroke gesture sample directly, operating over the sequence of 2D coordinates. This makes this approach computationally efficient, with minimal coding overhead. On the contrary, G3 extracts a number of complex model parameters from a given sample and perturb their parameters. This approach becomes more resilient to gesture type and input device, however this apparent superiority comes at the expense of performance: G3 may take several seconds to process a gesture sample, including the delay times incurred by network latency (G3 is available as a web service, GPSR is standalone software).

On a side note, the concepts concerning internal models of human movements have been well supported by behavioral studies in the field of sensory motor control. Overall, it is assumed that users are "ideal" motion planners who choose movement trajectories to minimize an expected loss [30, 26]. Currently, we can find two compelling theories to describe those movements: the Minimization Theory [10] and the Kinematic Theory [24]. Actually, it has been shown that their concepts are linked and describe, with different arguments, a model of velocity profiles [9, 17].

Users tend to be reluctant to invest time and effort upfront to train or adjust software before using it [5]. Further, users are unwilling to provide more than a small set of samples for training [19]. Consequently, synthesizing techniques like GPSR and G3 are of high value, as they help to lower time and costs associated to recruiting users and subsequent data labeling. Furthermore, researchers can focus exclusively on UI design rather than fret over machine learning concepts or toolkits that may not be available for their platform. Eventually, both GPSR and G3 can be used for rapid prototyping, allowing developers to define new gestures on demand. However, if realtime response is mandatory and only 2D information is needed, we would recommend to use GPSR. For more advanced applications such as modeling mouse movements [22] or handwriting behavior [23] we would recommend G3.

CONCLUSION AND FUTURE WORK

This work provides evidence against the implied alternate hypothesis of a difference between human and synthesized stroke gestures. Our findings have demonstrated the importance of our study: until now the "human likeness" of synthesized gestures has been measured indirectly, intermediated by classification accuracy. Researchers and practitioners can be finally confident that synthesized gestures (either via GPSR or G3) are actually reflective of how users perceive gestures, and that G3 is more appropriate for rendering realistic multistroke and touch-based gestures.

For future work, the relationship between what *looks* realistic and what *is* realistic should be studied further. As of today, there is no gold standard to assess how realistic a gesture sample really is. The method undertaken by this and previous work assumes that humans a reasonable judges of realism, but it might not be the case.

ACKNOWLEDGMENTS

I thank Daniel Martín-Albo and Réjean Plamondon for fruitful discussions during our previous (and current) research collaborations on gesture synthesis. I also thank the anonymous DIS reviewers for their constructive and helpful feedback.

APPENDIX

Figure 4 depicts all the gestures in both datasets drawn with a stylus at medium speed, which represents the tradeoff between drawing accuracy and execution pace.

Solution to Figure 1

Group (a) was produced by humans, group (b) was synthesized using GPSR, and group (c) was synthesized using G3.

REFERENCES

1. Shahriyar Amini and Yang Li. 2013. CrowdLearner: Rapidly creating mobile recognizers using crowdsourcing. In *Proc. UIST '13*.
2. Eric Anquetil, Laurent Miclet, Sabri Bayoudh, and others. 2007. Synthetic on-line handwriting generation by distortions and analogy. In *Proc. IGS '07*.
3. Lisa Anthony, Radu-Daniel Vatavu, and Jacob O. Wobbrock. 2013. Understanding the consistency of users' pen and finger stroke gesture articulation. In *Proc. GI '13*.
4. Lisa Anthony and Jacob O. Wobbrock. 2012. $N-protractor: a fast and accurate multistroke recognizer. In *Proc. GI '12*.
5. Caroline Appert and Shumin Zhai. 2009. Using strokes as command shortcuts: Cognitive benefits and toolkit support. In *Proc. CHI '09*.
6. Norman E. Breslow and David G. Clayton. 1993. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **88**(421).
7. Malcolm R. Davis and Tom O. Ellis. 1964. The RAND tablet: A man-machine graphical communication device. In *Proc. AFIPS '64*.
8. Moussa Djioua and Réjean Plamondon. 2008. *The Kinematic Theory and Minimum Principles in motor control: A conceptual comparison*. Technical Report EPM-RT-2008-03. École Polytechnique de Montréal.
9. Moussa Djioua and Réjean Plamondon. 2010. The limit profile of a rapid movement velocity. *Hum. Mov. Sci.* **29**(1).
10. Tamar Flash and Neville Hogan. 1985. The coordination of arm movements: an experimentally confirmed mathematical model. *J. Neurosci.* **5**(7).
11. Javier Galbally, Réjean Plamondon, Julián Fierrez, and Javier Ortega-García. 2012. Synthetic on-line signature generation. Part II: Experimental validation. *Pattern Recogn.* **45**(7).
12. John M. Hollerbach. 1981. An oscillation theory of handwriting. *Biol. Cybern.* **39**(2).
13. Shaun K. Kane, Jacob O. Wobbrock, and Richard E. Ladner. 2011. Usable gestures for blind people: Understanding preference and performance. In *Proc. CHI '11*.
14. Daniel Kohlsdorf and Thad E. Starner. 2013. MAGIC Summoning: Towards automatic suggesting and testing of gestures with low probability of false positives during use. *J. Mach. Learn. Res.* **14**(1).
15. Luis A. Leiva, Vicent Alabau, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal. 2015. Context-aware gestures for mixed-initiative text editing UIs. *Interact. Comput.* **27**(1).
16. Luis A. Leiva, Daniel Martín-Albo, and Réjean Plamondon. 2016. Gestures à Go Go: Authoring synthetic human-like stroke gestures using the kinematic theory of rapid movements. *ACM T. Intel. Syst. Tec.* **7**(2).
17. Luis A. Leiva, Daniel Martín-Albo, and Réjean Plamondon. 2017a. The kinematic theory produces human-like stroke gestures. *Interact. Comput*. Advance online publication. DOI: 10.1093/iwc/iww039.
18. Luis A. Leiva, Daniel Martín-Albo, and Radu-Daniel Vatavu. 2017b. Synthesizing stroke gestures across user populations: A case for users with visual impairments. In *Proc. CHI '17*.
19. Yang Li. 2010. Protractor: a fast and accurate gesture recognizer. In *Proc. CHI '10*.
20. Hao Lü, James A. Fogarty, and Yang Li. 2014. Gesture Script: Recognizing gestures and their structure using rendering scripts and interactively trained parts. In *Proc. CHI '14*.
21. Daniel Martín-Albo and Luis A. Leiva. 2016. G3: bootstrapping stroke gestures design with synthetic samples and built-in recognizers. In *Proc. MobileHCI '16*.
22. Daniel Martín-Albo, Luis A. Leiva, Jeff Huang, and Réjean Plamondon. 2016b. Strokes of insight: User intent detection and kinematic compression of mouse cursor trails. *Inform. Process. Manag.* **56**(6).
23. Daniel Martín-Albo, Luis A. Leiva, and Réjean Plamondon. 2016a. On the design of personal digital bodyguards: Impact of hardware resolution on handwriting analysis. In *Proc. ICFHR '16*.
24. Réjean Plamondon. 1995. A kinematic theory of rapid human movements. Part I: Movement representation and control. *Biol. Cybern.* **72**(4).
25. Réjean Plamondon and Moussa Djioua. 2006. A multi-level representation paradigm for handwriting stroke generation. *Hum. Mov. Sci.* **25**(4–5).
26. Philip Quinn and Shumin Zhai. 2016. Modeling gesture-typing movements. *Human-Computer Interaction*. Advance online publication. DOI: 10.1080/07370024.2016.1215922.
27. Patrice Simard and Yann LeCun. 1991. Reverse TDNN: an architecture for trajectory generation. In *NIPS '91*.

28. Ivan E. Sutherland. 1963. *Sketchpad: A man-machine graphical communication system*. Technical Report 296. Lincoln Laboratory, MIT.

29. Eugene M. Taranta, II, Mehran Maghoumi, Corey R. Pittman, and Joseph J. LaViola, Jr. 2016. A rapid prototyping approach to synthetic data generation for improved 2D gesture recognition. In *Proc. UIST '16*.

30. Julia Trommershäuser, Laurence T. Maloney, and Michael S. Landy. 2003. Statistical decision theory and trade-offs in the control of motor response. *Spat. Vis.* **16**(3–4).

31. Huawei Tu, Xiangshi Ren, and Shumin Zhai. 2012. A comparative evaluation of finger and pen stroke gestures. In *Proc. CHI '12*.

32. Jacob O. Wobbrock, Andrew D. Wilson, and Yang Li. 2007. Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. In *Proc. UIST '07*.

33. Shumi Zhai, Per Ola Kristensson, Caroline Appert, Tue H. Anderson, and Xiang Cao. 2012. Foundational issues in touch-surface stroke gesture design — an integrative review. In *Foundations and Trends in Human-Computer Interaction*. Vol. 5.