Estimating the Perceived Difficulty of Pen Gestures

Radu-Daniel Vatavu[1], Daniel Vogel[2,3], Géry Casiez[2], Laurent Grisoni[2]

[1]University Stefan cel Mare of Suceava, Romania,
[2]LIFL, INRIA Lille & University of Lille, France, [3]Mount Allison University, Canada
vatavu@eed.usv.ro, dvogel@mta.ca, gery.casiez@lifl.fr, laurent.grisoni@lifl.fr

Abstract. Our empirical results show that users perceive the execution difficulty of single stroke gestures consistently, and execution difficulty is highly correlated with gesture production time. We use these results to design two simple rules for estimating execution difficulty: establishing the relative ranking of difficulty among multiple gestures; and classifying a single gesture into five levels of difficulty. We confirm that the CLC model does not provide an accurate prediction of production time magnitude, and instead show that a reasonably accurate estimate can be calculated using only a few gesture execution samples from a few people. Using this estimated production time, our rules, on average, rank gesture difficulty with 90% accuracy and rate gesture difficulty with 75% accuracy. Designers can use our results to choose application gestures, and researchers can build on our analysis in other gesture domains and for modeling gesture performance.

Keywords: gesture-based interfaces, pen input, gesture descriptors.

1 Introduction

There are three primary factors which contribute to a successful gesture-based interface: the acquisition technology, the recognizer, and the design of the gesture set. Technologies to acquire gestures [7,9,16,25], and gesture recognition algorithms [10,12,26], are now quite robust and widely available. However, developing techniques and criteria to help designers create an intuitive and easy-to-perform gesture set remain an active area of research. The challenge is that in order to successfully integrate into an application, a gesture has to satisfy multiple criteria: it must be unambiguously recognized [2,13,14]; fit well with its associated function [16,17,25]; be easy to learn and recall [17]; and be efficient to perform [16,17,25].

Researchers have offered two different strategies to assist designers. The first is to use predictive models to analytically evaluate candidate gestures. These have been successful for evaluating recognition ambiguity [2,14] and have made progress towards predicting actual performance time [4,8]. Unfortunately, creating accurate predictive models for non-recognition criteria such as performance time is difficult due to the complexity of gestural motion and criteria interdependencies — factors which are also influenced by an individual user's cognitive ability, physical skill, and cultural context. For these reasons, researchers have proposed a second strategy using

formal user studies for participatory design and gesture set evaluation [2,16,17,19,25]. Involving users in any design process is a good idea, but the effort to plan, run, and analyze these kinds of studies is large compared to using a predictive model.

We offer a practical solution in-between a model and a user study. Based on an estimate of actual production time, we found that designers can reasonably estimate user's perceived gesture *execution difficulty*. The notion of difficulty encompasses multiple criteria including the ease with which a gesture may be learned, remembered, and performed. This notion of difficulty has been mentioned in previous work [16,17,25], but there has been no previous attempt to examine it in detail or estimate it. In an experiment using single stroke pen gestures, we elicited a difficulty classification rating and a relative difficulty ranking from participants. Based on data from a second validation experiment, our results show that the difficulty ranking can be predicted with greater than 93% accuracy using measured production time and 87% using the Isokoski first-order predictive production time model [8]. Using a Bayes classification rule and measured production time, we can also classify the difficulty rating with 83% accuracy. Since the times predicted by the CLC predictive model [4] reduced the accuracy of our classification rule to 25%, we analyzed an alternative approach. We found that production time can be reasonably estimated by gathering a few samples of actual production time – a set of data which may already exist for the purpose of training a gesture recognizer. With three people supplying three gesture samples, our classification rule achieved 75% accuracy on average and increased the average accuracy of the estimated difficulty ranking to 90%.

Our findings that gesture difficulty can be predicted from production time, together with our results regarding the reasonable estimation of production time based on a very small set of data, provide designers with a general measurement encompassing multiple criteria to assess gesture sets without a full formal user study.

2 Previous Work

Creating a successful gesture-based interface is challenging. Once a vocabulary of gestures moves beyond a small set of directional strokes, it becomes more difficult to learn, remember, and use [11]. Techniques exist which assist with recall and help to transition users from novice to expert: examples include crib-sheet diagrams [11] and dynamic path guides [3]. While these techniques are effective, they assume that a good gesture set has already been created.

2.1 Gesture Design Tools

One way to make the designer's job easier is to use a gesture design tool. An example is Appert and Zhai's Stroke Shortcuts Toolkit [1] which includes a simple tool with a predefined dictionary of stroke primitives. The hope is that a designer's creativity is stimulated with a "structured design space that can be systematically explored". Long et al.'s Quill gesture design tool [13,14] goes further by providing metrics to help designers evaluate potential gesture sets. The metrics relate to recognition rate, and conveyed through values such as classification distance or visualized as confusion

matrices. Ashbrook and Starner's MAGIC tool [2] introduces gesture *goodness* as a metric. In an evaluation, this seemingly abstract metric was useful as a quantitative guideline compared to a specific breakdown of individual measures (such as inter-class variability graphs). However, goodness is also closely related to recognition rate. Although participants were also asked to design gestures that would be easy to remember, perform, and be socially acceptable, MAGIC, like Quill, does not provide any quantitative feedback for these criteria.

2.2 Models

Producing quantitative measurements to represent other criteria requires predictive models. For example, Long et al. [15] developed a model for predicting the perceived visual similarity of two gestures. Their model was generated by selecting a subset of geometric and dynamic features of gesture trajectories, and looking for a correlation with experimentally determined user rated visual similarity. The final model could predict visual similarity of two gestures reasonably well (correlated R^2=.56 with ground truth). One application is increasing recognition rate by avoiding ambiguous gestures, but the authors also argue that a visual similarity metric may be used to improve a gesture's fit with its function. For example, designers could assign visually similar gestures to similar operations (such as scroll up and scroll down), and dissimilar gestures to more abstract tasks such as cut and paste.

Isokoski [8] introduced a model to predict the relative ordering of gesture production times based on geometric complexity. The model sums the minimal number of straight segments needed to maintain a human recognizable shape in the gesture. This sum is interpreted as a complexity number and can be used as a first-order ranking of gesture production time: the model ranked production times of *Unistroke* characters with R^2=.85. Although there is ambiguity in the definition and calculation method, Isokoski's model has the advantage of being conceptually simple.

Cao and Zhai's [4] Curves, Lines and Corners (CLC) model goes beyond Isokoski by attempting to predict the actual production time of a single stroke gesture. After decomposing a gesture into curved and straight segments, the model calculates individual production times for curves based on Viviani's 2/3 power law of curvature [22] and a simple power term based on the length of straight lines (no time is calculated for corners, they are only used to segment lines and curves). The authors found that CLC works very well as a first order predictor (correlations with test data had R^2>.90), but over- or under-predicted arbitrary gestures times by 30% and over-predicted *Unistroke* and *Graffiti* gestures by more than 40%. Castellucci and MacKenzie also noted this type of performance for CLC [5]. Cao and Zhai attribute this behaviour to the model's inability to compensate for unfamiliar and little practiced gestures, or familiar and well-practiced gestures.

2.3 User Studies

Rather than rely on predictive models, researchers have suggested that user studies should be used to assist in the design and evaluation of gesture sets. For example,

Nielsen et al. [17] provide a user-centered procedure to design whole-hand gestures. The procedure requires two user studies, an initial study to gather user input to inform design and a subsequent study to evaluate. In a case study application, they report they were able to obtain a good gesture set, but the procedure was very time consuming. Also, key stages such as the generation of scenarios must be carefully prepared or else results may be substandard.

Wobbrock et al. [25] take a participatory design approach by eliciting a gesture set from users. Using wizard-of-oz techniques, they asked users to mimic the best multi-touch gesture to match a demonstrated action such as scale, rotate, move, etc. The study, as well as a follow-up [16], also gathered rankings for each candidate gesture's intuitiveness and ease-of-execution. Perhaps surprising, but the authors report that gestures which experienced designers propose are not always preferred by users [16].

2.4 Summary

Ideally, the best way to design an intuitive and easy-to-perform gesture set is to involve users like Nielsen et al. [17] and Wobbrock et al. [25] since even experienced designers cannot predict user preference [16]. But, faced with the large amount of effort required to plan, run, and analyze these studies [17], perhaps there is a way for designers to evaluate candidate designs using predictive models and/or minimal user data. Long et al.'s [15] visual similarity predictive model is interesting since it can guide designers with a gesture's fit with a function. Isokoski [8], and Cao and Zhai [4], have made progress towards estimating actual gesture production time, a measure which should directly relate to how efficient a gesture is to perform. However, Cao and Zhai [4] and Castellucci and MacKenzie [5] note that production time is a partial function of many factors and therefore an accurate predictive model remains elusive. Inspired by Long et al., as well as Ashbrook and Starner's success with a seemingly abstract post-hoc measure of goodness [2], we focus on a measure of *execution difficulty*.

3 Experiment 1: Measuring Execution Difficulty

The notion of execution difficulty (or the converse, ease-of-execution) is frequently mentioned [2,14,15,17] and has been measured for multi-touch gestures with post-experiment surveys [16,25], but there has been no attempt to estimate it a priori. Morris et al. associate difficulty with "carrying out the gesture's physical action" [16]. Carrying out an action refers directly to efficiency of performance, but also involves a cognitive process which relates to how easy a gesture is to learn and recall [4]. Thus, execution difficulty is a general quantitative measure which combines multiple design criteria: learn-ability, recall, and performance. More abstract measures, such as goodness [2] and general preference [26] may include additional criteria (such as social acceptability [19]), but the more general the measure, the more abstract it is due to more complex relationships of the underlying criteria. The challenge is how to estimate execution difficulty given a candidate gesture or gesture set, with the knowledge that it encompasses criteria which are known to be difficult to predict.

In the first experiment, we measure perceived execution difficulty for a set of single stroke gestures. If there is significant agreement across participants, then it is likely to be an intuitive measure suitable for a priori estimation. Using the participants' movement logs and the geometric gesture shapes, we compute quantitative measures ("descriptors") and test these for correlations with the participants' responses. If well correlated descriptors exist, and they can be estimated or computed directly, then designers have a way to estimate perceived difficulty of candidate gesture designs.

Participants

14 right-handed people (3 females) participated in the experiment (mean age 21 years, SD 1). 11 out of 14 participants had no pen-based interface experience.

Apparatus

Gestures were entered using a 17 inch (431 mm) Wacom DTU-710 Interactive Pen Display running at a resolution of 1280 x 1024 px (pixel pitch 0.264 x 0.264 mm) and capable of capturing pen input at 133Hz. The display was positioned horizontally to approximate a physical pen and paper context. A 2.4GHz computer ran a C# full screen application. The participant entered gestures in a 420 x 420 px (110 x 110 mm) square box centered in the display (Fig 1).

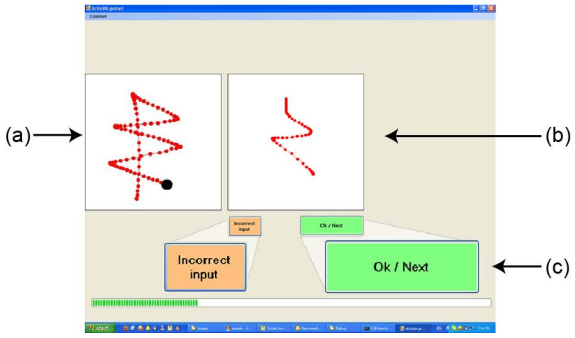


Fig. 1. Experiment Application: (a) current gesture to perform; (b) gesture input area; (c) post-entry choice buttons.

Task

Each trial began with the path of the current gesture to be entered shown on the left side of the display (Fig 1a). Participants were instructed to enter a continuous stroke for the gesture and to balance speed and accuracy. After performing the gesture, two buttons were enabled representing a choice between flagging their input as incorrect or continuing to the next gesture. Participants were instructed to flag a stroke as incorrect if the shape they entered was different from the target gesture, or if some accidental input occurred such as the pen slipping or moving unevenly. This was logged as an input error and the participant was asked to re-execute the gesture. Like Wobbrock et al. [25], we wanted our participants to decide whether a gesture was

similar to the template, avoiding any confounding effects due to the behavior of a recognizer. As an extra precaution, all participant executions were visually inspected by the authors and confirmed that they were correctly entered.

Gesture Set

There were 18 different single stroke gestures (Fig 2). The set contains 9 gestures designed to be *familiar* (i.e. letters and shapes used in everyday writing) and 9 gestures designed to be *unfamiliar* (e.g. the *twirl-omega* and *flower* shapes may appear familiar, but are unlikely to be practiced as a pen stroke, while *steep-hill* and *triangles-chain* are completely new shapes). As discussed earlier, Cao and Zhai [4] argue that familiarity affects actual performance time due to practice. The idea is that a more practiced gesture will result in a lower performance time in spite of high objective geometric complexity. For example, although the letter *g* is a rather complex series of twists and 180-degrees turns, it would be difficult to reproduce initially; but, with practice it can be executed very quickly. Since practice also relates to how easy a gesture is to learn and recall, *familiarity* is likely to relate to execution difficulty. We expected that more familiar gestures will be rated as easier to perform, even if they have high objective complexity.

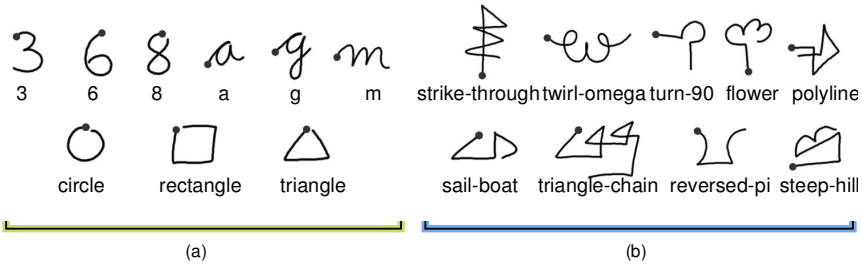


Fig. 2. The 18 gestures used in the experiment: (a) left 9 designed to be familiar; (b) right 9 designed to be unfamiliar.

Design

Each participant executed each gesture 20 times, with the 18x20 = 360 gestures presented in random order. The number of repetitions (20) was chosen larger than the current practice when eliciting gestures from users, be it for training gesture recognizers [26] or even for deriving performance models [4]. We purposely did this to ensure motor learning for all gestures so that participants would reach execution automaticity. Participants were allowed to take as many breaks as they wished. The experiment took approximately 40 minutes.

Post-Experiment Questionnaire

After the experiment, participants answered a short questionnaire regarding their perceived execution difficulty when performing the gestures. We gathered this information in two different ways: an individual execution difficulty *Rating* for each gesture using a 5-point Likert scale; and an ordered *Ranking* of all gestures according

to relative execution difficulty. The 5-point Likert scale rating question (Table 1) was presented as a 5 column table: participants entered ratings for the 18 gestures in any order they chose. Participants were asked to enter the rating by drawing the gesture in the column corresponding to the desired Likert rating. We hoped this would allow them to re-enact the gesture performance and make visual inspection easy. They could modify previous ratings at any time until they were confident of their final choices.

Table 1. Likert questions used to elicit execution difficulty *Rating*.

Likert rating	Associated explanation
1. very easy to execute	I executed these gestures immediately and effortlessly with absolutely no need to pay attention
2. easy to execute	I executed these easily, almost without paying attention
3. moderate difficulty	I occasionally paid special attention during execution
4. difficult to execute	I paid special attention with each execution
5. very difficult to execute	I had to concentrate for each execution. There were times when I did not get the right shape from the first attempt

The ordered ranking of all gestures according to ascending execution difficulty was completed after the Likert rating. This enabled participants to use the rating classes to assist with this otherwise difficult task. As before, we asked them to draw the gestures in order to revisit relative differences in difficulty as they completed the ranking.

We also asked participants to explain their perception of gesture difficulty: what they found difficult or easy for each gesture execution. Finally, we asked them to identify which shapes they found *Familiar* (they had seen and practiced before) in order to test our choice for familiar and unfamiliar gestures.

4 Results

We found a high degree of agreement between participant *Rating* of execution difficulty (Kendall's W=.78[1], $\chi^2(17)=185.60$, $p<.001$). The agreement was even stronger for *Ranking* which participants commented as being a difficult task (W=.82, $\chi^2(17)=195.17$, $p<.001$). Both coefficients are well above 0.5 indicating our sample size was appropriate with a large Cohen effect. Since *Rating* was designed to be used as a first approximation for *Ranking*, there was a significant correlation between their median ratings ($\rho_{(N=18)}=.97$, $p=.01$).

Fig 3 illustrates the median *Rating* and *Ranking* ratings for each gesture. A repeated-measures Friedman's ANOVA was used in order to test the influence of gesture type (nominal with 18 cases) over *Rating* and *Ranking*. The results showed a significant effect of *Gesture* over both *Rating* ($\chi^2(17)=185.60$) and *Ranking* ($\chi^2(17)=195.17$, at $p<.001$).

[1] Kendall's coefficient of concordance W in [0..1] where 0 denotes no agreement at all and 1 represents absolute agreement.

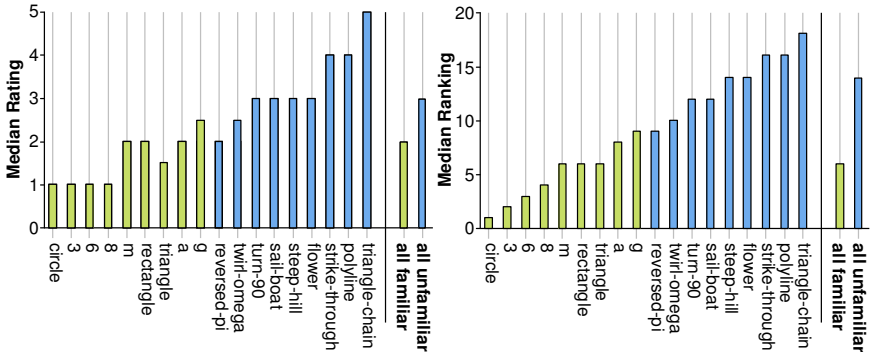


Fig. 3. Left: median gesture *Rating* (higher *Rating* values were perceived to be more difficult to execute). Right: median gesture *Ranking* (higher numerical *Ranking* for gestures perceived to be more difficult to execute). In both graphs, gestures are ordered by ascending *Ranking*.

Across all 14 participants there were 17 deviations (6.7% of the total responses) from our gestures set's assumed *Familiarity*. 14 deviations were assumed unfamiliar gestures: 7 participants found the *twirl-omega* gesture familiar, 4 *reversed-pi*, 2 *flower*, and one participant said the *sail-boat* and *steep-hill* were also familiar. The latter also noted that the assumed-to-be-familiar gestures *a* and *g* were unfamiliar because the starting point was not in the same location where they usually start those letters. As part of their comments regarding their perception of gesture difficulty, three participants noted the same issue of starting position with *a* and *g* and one participant with *8*, but they did not feel this made them unfamiliar. This relates to the problem of allographic variation in handwriting where individual differences in the formation of character shapes pose problems for handwriting recognizers [21]. Aside from *twirl-omega* where *Familiarity* deviations occurred with half of our participants, our assumed gesture familiarity was reasonable. We could treat these deviations as outliers since they represent less than 4% out of the total responses, but when possible *Familiarity* related analysis is based on actual participant responses.

The median *Ranking* and *Rating* across *all familiar* and *all unfamiliar* gestures (Fig 3) are significantly different according to a Wilcoxon signed-rank test ($z_{(N=14)}$=-3.402, p=.001 for *Rating* and $z_{(N=14)}$=-3.400, p=.001 for *Ranking*, both with large effects, r=-.64). These 9 assumed familiar gestures are among the 11 gestures assigned to the easiest *Rating* levels**,** and are among the lowest 10 gestures in ascending difficulty *Ranking* (Fig 3). The *twirl-omega* and *reversed-pi* (two out of three contentiously unfamiliar gestures) also share the two easiest median *Rating* levels, and *reversed-pi* has the same median ranking as the familiar gesture *g*.

5 Towards Estimating Execution Difficulty

Given the high agreement of perceived execution difficulty *Rating* and *Ranking* in experiment 1, we can search for a way to estimate difficulty in the absence of a formal experiment. Essentially, if a correlation exists with one or more characteristic gesture

descriptors, then those descriptors can be used to estimate execution difficulty. We examined many potential descriptors (Table 2): all of Rubine's static geometric descriptors and measured quantities [20], the additional geometric descriptors used by Long et al. [15], Hu invariant spatial curve moments commonly used in image processing for contours and shapes [18](p. 606), Isokoski's complexity measure [8], and the production time predicted by Cao and Zhai's CLC model [4].

Table 2. Descriptors (bold indicates significant correlation with *Rating* or *Ranking*).

Rubine's set [20]: Geometric	
1. Cosine of initial angle (cosine1)	7. Sine of angle between first and last points (sine2)
2. Sine of initial angle (sine1)	**8. Total length**
3. Size of bounding box (bbox size)	9. Total turning angle
4. Angle of bounding box (bbox angle)	**10. Total absolute turning angle (turn angle)**
5. Distance between first and last points	**11. Sharpness or (energy)**
6. Cosine of angle between first and last points (cosine2)	
Rubine's set [20]: Measured	
12. Production Time (time)	
13. Speed	
Long et al.'s visual similarity set [15]: Geometric	
14. Aspect	**18. Size of bounding box (density2)**
15. Total angle traversed / total length	19. Openness
16. Total angle / total absolute angle	**20. Area of bounding box (bbox area)**
17. Distance between first and last points (density1)	
Hu invariant spatial moments [18, p.606]: Geometric	
21 – 27. Hu1, **Hu2**, Hu3, Hu4, Hu5, Hu6, Hu7	
Model predictions	
28. CLC Predicted Production Time [4]	
29. Isokoski's complexity measure [8]	

The calculation of the Rubine, Long et al., and Hu descriptors are straightforward to apply to the geometric shape of the gesture, given the descriptions and equations in the cited works. We computed these measurements using two representations of geometric gesture shapes. To approximate a design scenario where the gestures have been drawn, but not performed, we used the target gesture shapes displayed in the left panel of our experimental application (i.e. the vector drawings in Fig 1). We will refer to these as geometric descriptors using *Drawn* representations. We also computed mean descriptors using the actual gesture geometries as performed by the participants in our experiment. Theoretically, this is a best case scenario for geometric descriptor performance, but with the potential issue of overfitting. We will refer to these geometric descriptors using *Performed* representations. Both *Drawn* and *Performed* representations were preprocessed similar to previous work [4,10,26] by normalizing without deformation, centering on the origin, and re-sampling uniformly into n=32 points. To calculate the CLC predicted production time, we used the PlayCLC program[2]. As noted earlier, the definition and calculation of Isokoski's complexity measure is ambiguous. By studying examples [8](p. 360) we developed quantitative

[2] from http://www.cs.toronto.edu/~caox/PlayCLC/PlayCLC.htm

guidelines to perform the necessary reduction of arcs into line segments: if the angle α inscribed by an arc was greater than 270° use 3 segments; if α<120° use 1 segment; otherwise use 2 segments. We could verify these guidelines with our *3* and *circle* shapes, also included in Isokoski's examples.

Note that all descriptors based on geometry are static and will not change with practice. For example, a geometrically complex, but familiar gesture such as *g* may have a lower *Rating* compared to a geometrically simple, but unfamiliar gesture such as *sail-boat*. Rubine's Production Time and Speed descriptors are measured, i.e. they are computed from data gathered during actual gesture performance, so they include effects for practice. Of course, using this type of post-hoc measure for a-priori prediction seems paradoxical. Our initial rationalization is that some future model may be able to accurately predict these measures (such as an improved CLC model for Production Time), and we show later that the relevant measure of Production Time can be approximated with a very small set of informally gathered user data.

All of the potential descriptors in Table 2 were tested for correlations with execution difficulty *Rating* and *Ranking*. This was done overall, as well as separately with familiar and unfamiliar gesture groups. Descriptors with at least one significant Spearman correlation coefficient are listed in Table 3 (for geometric descriptors using *Drawn* representations in Fig 1) and Table 4 (for geometric descriptors using participant *Performed* representations).

Table 3. Correlations of geometric descriptors using *Drawn* representations. Spearman correlation of descriptor with median *Rating* and *Ranking* in descending order of overall *Rating* coefficients; coefficients are reported at p = .01 (**) and p = .05 (*) significance levels; N = 18 for all, N = 9 for familiar and N = 8 for unfamiliar gestures (*twirl-omega* was excluded). The largest coefficient in each column is shown in bold text (two bold coefficients in the same column are not significantly different).

	all		**familiar**		**unfamiliar**	
	Rating	*Ranking*	*Rating*	*Ranking*	*Rating*	*Ranking*
bbox size	**.78****	**.75****	n.s.	n.s.	**.93****	**.98****
bbox area	.72**	.70**	n.s.	n.s.	**.93****	**.98****
length	.60**	.60**	n.s.	**.67***	**.79***	**.86****
cosine2	n.s.	n.s.	**.73***	**.81****	-.86*	-.88**
density1	n.s.	n.s.	n.s.	n.s.	**.73***	.81*
Hu2	n.s.	-.50*	n.s.	n.s.	n.s.	n.s.

Production time has the highest correlations with *Rating* and *Ranking* overall; and, in all but one case, it is among the highest correlations when tested separately with familiar and unfamiliar gesture groups. Speed had the second highest (negative) correlation when all gestures were considered together, but not significant when tested separately with familiar and unfamiliar.

Note that there is evidence that production time should be a scale invariant.Viviani and Terzuolo [23] found that execution times for single strokes in handwriting are scale invariant. If we accept that a single stroke gesture is similar, then scale invariance should not be problematic. Isokoski [8] also provides additional evidence with his observation that average velocity increases with longer strokes.

Table 4. Correlations of geometric descriptors using *Performed* representations. Correlations reported as in Table 3.

	all		**familiar**		**unfamiliar**	
	Rating	*Ranking*	*Rating*	*Ranking*	*Rating*	*Ranking*
time	**.95****	**.96****	**.94****	**.84****	.79*	**.91****
-speed	**.87****	**.85****	n.s.	n.s.	n.s.	n.s.
length	.80**	.82**	**.94****	**.90****	.72*	**.81***
bbox size	.77**	.82**	n.s.	n.s.	n.s.	n.s.
Isokoski	.74**	.71**	**.70***	n.s.	n.s.	**.79***
bbox area	.70**	.75**	n.s.	n.s.	n.s.	n.s.
density2	.56*	.52*	**.90****	**.85****	.72*	**.76***
turn angle	.53*	.51*	n.s.	n.s.	.72*	**.83***
Hu2	-.48*	-.47*	n.s.	n.s.	n.s.	n.s.
CLC	.47*	n.s.	n.s.	n.s.	n.s.	**.79***
aspect	-.47*	n.s.	n.s.	n.s.	n.s.	n.s.
cosine2	n.s.	n.s.	**.86****	**.71***	**-.86****	-.88**
density1	n.s.	n.s.	n.s.	n.s.	.79*	**.86****
energy	n.s.	n.s.	n.s.	n.s.	n.s.	**.79***

In many cases, descriptors based on geometry had significantly lower correlation coefficients compared to measured values. An exception is length, which has all significant coefficients in Table 4 and all but one in Table 3. In the case of familiar gestures in Table 4, coefficients for length, along with density2, and cosine2 are not significantly different from actual production time. Although not significantly highest, Isokoski's complexity and two bounding box descriptors in Table 4 correlate reasonably well when all gestures were considered together, but are not even significant when tested separately with familiar and unfamiliar. In Table 3, some geometric descriptors such as the two bounding box descriptors and cosine2 correlate very well with unfamiliar gestures. Intuitively, the larger gestures may be more complex, and thus be more difficult to execute, but the high correlation of cosine2 is surprising. With low N values (8 for unfamiliar and 9 for familiar), there will be fewer significant differences between descriptors. The tendency for geometric descriptors to exhibit higher coefficients in either familiar or unfamiliar gesture groups is most likely because they cannot adapt to the effect of practice. This is similar to reasons given for the under- or over-estimation behavior of the CLC model [4,5].

Visual inspection of the most promising descriptors provides some intuition for their relative performance in predicting difficulty (Fig 4). Gestures are listed by ascending median *Ranking* on the horizontal axis, so a monotonic trend would suggest it is a good candidate for estimating *Ranking*. Actual production time ascends almost monotonically with *Ranking* demonstrating that gestures rated as being more difficult to execute have a greater production time. The static geometric descriptors for the most part increase with difficulty overall, but irregularities are much more pronounced suggesting a weaker fit. For example, letters *a*, *g* and *m* have long lengths, yet they are rated as easy to execute. This again speaks to familiarity: despite objective complexity, practiced gestures are rated with lower execution difficulty.

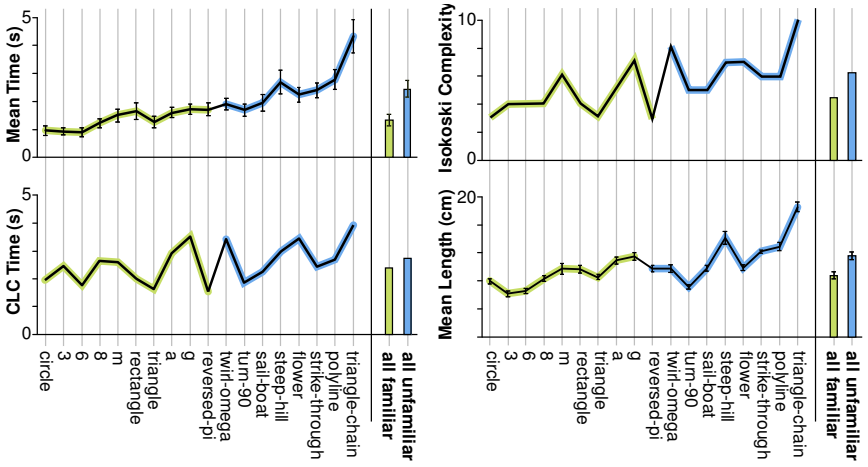


Fig. 4. Visual comparison of the four most promising measures and predictors (y-axes) with actual ascending median gesture *Ranking* (x-axis). A monotonic trend suggests the measure or predictor is a good candidate for estimating *Ranking* (e.g. mean measured time). NOTE: Error bars in all figures represent 95% CI.

There are also significant correlations between descriptors. Production time is correlated with length ($\rho_{(N=18)}$=.89, p=.01) and Isokoski's complexity and production time are correlated with length ($\rho_{(N=18)}$>.70, p=.01). This suggests a partial correlation between these three descriptors, so it is appropriate to test for shared variance. When controlling for production time, the other parameters are no longer significant (p>.05). When controlling for all but production time, production time was still found highly correlated with *Rating* ($\rho_{(N=18)}$=.73) and *Ranking* ($\rho_{(N=18)}$=.67) at p<.01. For familiar and unfamiliar groups, none of the correlations with *Rating* and *Ranking* were significant when either variable was controlled during partial correlations. The t-statistic for comparing coefficients [6] showed a significant difference between coefficients for *Rating* (t(15)=5.92) and *Ranking* (t(15)=4.02) at p<.01.

The poor performance of the CLC predicted production time is somewhat surprising. Previous results found CLC to be highly accurate for first-order predictions when comparing relative ratios of gesture set production times [4,5]. So, we expected it would also perform well with a similar first-order prediction task for execution difficulty *Ranking*, but it has no significant correlations with *Ranking* at all. To investigate further, we directly compared the CLC predicted production times to actual production times. For magnitude, we found a significant, but low correlation (R^2=.37, p=.01). For relative ranking, we also found a significant, but low Spearman correlation ($\rho_{(N=18)}$=.53, p=.05).

Production time is the best indicator of execution difficulty, but the CLC model is not able to accurately predict performance time for our purposes. So, we continue the development of execution difficulty estimation rules based on actual production time, with the assumption (and caveat) that we are at the moment using a post hoc measured value. Later, we show that a small sample of data will provide suitable estimations of production time.

6 Difficulty Estimation Rules

We present two rules for estimating execution difficulty based on production time. The first is a simple rule which compares two candidate gestures according to relative execution difficulty (as *Ranking* does), and the second uses Bayes' rule to classify a gesture into one of five categories of execution difficulty (such as those provided by the *Rating* measure).

Rule 1: Relative Difficulty Ranking
Gesture A is likely to be perceived as more difficult to execute than gesture B if the production time of A is greater than that of B:

$$time(\mathrm{A}) > time(\mathrm{B}) \text{ suggests } Ranking(\mathrm{A}) > Ranking(\mathrm{B})$$

To test this rule, we applied it to each pair of gestures (A,B) out of the (18x17)/2=153 possibilities in experiment 1 using the measured production time and counted how many times the rule was correct out of the total number of classification attempts (*Ranking* accuracy). The rule predicted the relative ranking correctly with 93% accuracy (11 errors out of 153 tests).

Rule 2: Classifying Difficulty Rating
Mapping from production time to one of our five difficulty classes (C_i, i=1..5: *very easy*, *easy*, *moderate*, *difficult*, and *very difficult*) is a pattern classification problem where each gesture is represented by a single feature, in our case production time. A common technique in statistical pattern recognition is Bayes' rule that minimizes classification error [24]. Bayes' rule uses each class-conditional density probability (i.e. the probability for a randomly chosen pattern x to lie in class C_i, denoted $p(x|C_i)$) together with the *a priori* probability of class C_i (or how likely it is to observe a pattern from this class, denoted $p(C_i)$). Using this data, Bayes' rule computes the *a posteriori* probability of x belonging to each class, $p(C_i|x)$, and assigns x to class C_j for which the a posteriori probability is maximum:

$$x \in C_j \Leftarrow j = \arg\max_{i=1,5} \{p(x \mid C_i) \cdot p(C_i)\} \quad (\mathbf{1})$$

In order to apply Bayes' rule, the conditional $p(x|C_i)$ and a priori $p(C_i)$ probabilities must be known for each of our 5 *Rating* classes. Normal parametric models are frequently assumed in practice (equation 2) for estimating the unknown conditional densities $p(x|C_i)$ [24](p.34) for which the parameters (mean μ_i and standard deviation σ_i) can be easily computed from the training set (in our case, data from experiment 1).

$$p(x \mid C_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2} \right) \quad (\mathbf{2})$$

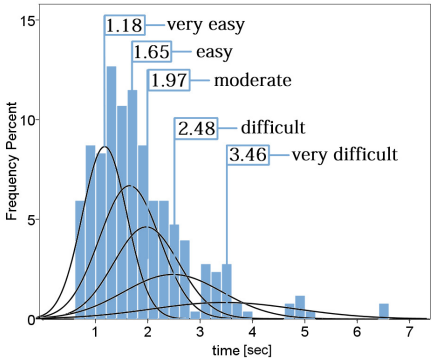
The a priori probabilities $p(C_i)$ are estimated from the training set as the percentage of samples falling into each class [24](p.34-39). In our case, μ_i are the mean production times for each *Rating* class (expressed in seconds); σ_i the standard deviations (seconds); and $p(C_i)$ the percentages of samples belonging to each *Rating* class. Table

5 lists these parameters as computed from our training data (experiment 1) with an illustration of each normal model superimposed over the production time histogram.

Table 5. Left: Bayes' Rule Parameters for the *Rating* Classification Rule. Right: Production time frequency histogram with superimposed time normal models for each *Rating*.

Rating levels	Bayes' Rule Parameters			
	μ_i	σ_i	$p(C_i)$	μ_i simplified[1]
very easy	1.18	0.43	27%	1.0
easy	1.65	0.58	28%	1.5
moderate	1.97	0.62	21%	2.0
difficult	2.48	0.95	16%	2.5
very difficult	3.46	1.38	8%	3.5

[1]simplified values represent *reasonable* approximations for the mean times μ_i; see the text for explanation and results.



We tested Bayes' rule in order to see how good it fits our data. We counted how many times the rule was correct out of 18 classification attemps (the *Rating* accuracy) by applying it to each gesture in our set. The rule achieved an accuracy rate of 83% on its own set (15 gestures were correctly classified to their *Rating* category as indicated by the participants). The three errors occurred for the *strike-through*, *turn-90*, and *sail-boat* gestures, all of which were misclassified to the next lower class. This confirms for now a good model fit for our data while Section 9 will show how the rule applies for new gestures in our validation experiment. The mean production times μ_i for each *Rating* level (see Table 5) could be approximated to more reasonable timestamps such as 1, 1.5, 2, 2.5 and 3.5 seconds (the μ_i simplified column in Table 5). These could represent more intuitive working estimates for each *Rating* class to be used by designers. When using these mean values with the computed standard deviations as before, we also obtained 83% classification accuracy.

7 Estimating Production Time

Applying our rules using measured production time works very well, but we would like designers to estimate production time without running such a formal experiment. Ideally, this could be done with predictive models. However, using times predicted by CLC, *Ranking* accuracy dropped to 67% and *Rating* down to 28%. Although Isokoski does not predict actual time, it can be used for relative *Ranking* where it managed a prediction accuracy of 82%.

Examining the data from experiment 1, we found that individual participant gesture production times are highly correlated with overall mean production times $\rho_{(N=18)}$=.96, p=.01 (min .92, max 1.0). This consistency made us wonder if a designer could estimate difficulty based on only few samples of measured production time. Instead of

a long formal experiment, a few people could perform the candidate gestures a few times in a simple data gathering application. Even more, this data is likely to already exist for training the gesture recognizer [12,20,26].

We first consider the minimal case of gathering data from a single person. Again using data from experiment 1, for each participant, we randomly selected M out of 20 execution samples for each gesture to calculate a mean production time. Using these mean times, we apply our rules and compute the prediction accuracy for *Rating* and *Ranking*. The random selection was repeated 100 times as M varied from 1 to 20: thus 14 participants x 18 gestures x 20 M values x 100 repetitions = 504,000 predictions.

The mean accuracy for *Rating* begins to level out at M=3 near 53%, and *Ranking* also approaches 91% (Fig 5, left). The effect of M over *Rating* is significant ($\chi^2(19)=476.4$, $p<.001$). A Wilcoxon signed-rank test found significant effects between (1,20), (3,20) and (5,20) with a small Cohen effect ($r<.3$). The effect of M over *Ranking* was significant ($\chi^2(19)=4140.54$, $p<.001$) with significant differences between (1,20) ($r=.52$), (3,20) and (5,20) with medium effects ($r<.5$). With 3 samples, mean *Rating* accuracy was 53% (SD 18%) and mean *Ranking* accuracy 89% (SD 3%).

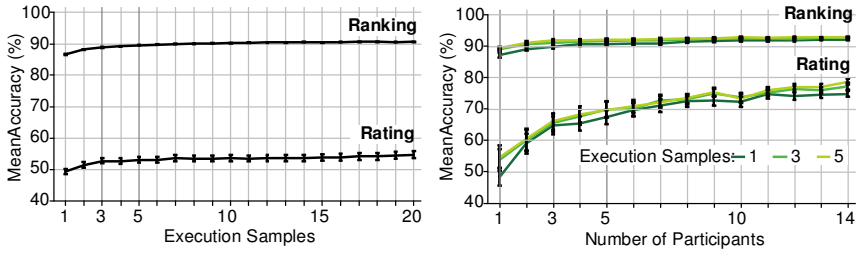


Fig. 5. Left: prediction accuracies for Rating and Ranking vs. number of execution samples. Right: Difficulty prediction accuracies vs. number of participants.

We continue our analysis by varying the number of participants N=1..14 given M=1,3,5 individual gesture execution samples from each. Similar to before, we randomly selected the gesture samples 100 times for each N: thus 14 participants x 18 gestures x 3 M values x 100 repetitions = 75,600 predictions.

The mean accuracy of *Rating* increases from 52% using one participant to 77% (significant, $\chi^2(13)=496.45$, $p<.001$) when data from all participants is used (Fig 5, right). The same trend is observed independently for M=1,3, and 5 executions from each participant. The accuracy of *Ranking* increases from 88% to 93% ($\chi^2(13)=715.1$, $p<.001$). The effect of M was found significant for both *Rating* and *Ranking* (at $p<.001$) but the Wilcoxon signed-rank test showed small Cohen effects between (1,3) and (1,5) $r<.3$ and very small between (3,5) $r<.15$. With 3 participants and 3 execution samples mean *Rating* accuracy was 66% (SD 14%) and mean *Ranking* accuracy 91% (SD 2%). With 5 participants and 3 execution samples mean *Rating* accuracy was 70% (SD 13%) and mean *Ranking* accuracy 92% (SD 2%).

In summary, on average, a designer could estimate a relative *Ranking* of execution difficulty with 89% using 3 gesture execution samples from a single person. To estimate *Rating*, 3 execution samples from 3 or 5 people are needed to achieve mean accuracies of 66% and 70% respectively.

8 Experiment 2: Validation of Difficulty Estimation Rules

A second experiment, similar to the first, was used to validate our execution difficulty rules as well as our simple production time estimation technique. The same apparatus, task, and design were used, but with 20 different gestures (Fig 6) and 11 new participants: 11 x 20 x 20 = 4,400 executions.

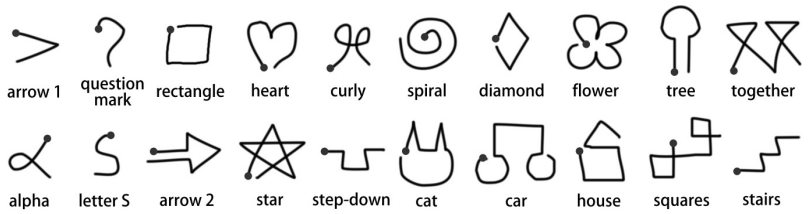


Fig. 6. The validation set of 20 gestures.

Results

We found the same high level of correlation between participants' difficulty *Rating* (Kendall's W=.78, $\chi^2(19)=163.61$, p<.001) and *Ranking* (W=.80, $\chi^2(19)=166.79$, p<.001). *Rating* and *Ranking* were again highly correlated ($\rho_{(N=20)}=.94$, p=.01).

Estimates of Execution Difficulty

We first establish an accuracy upper bound using the actual measured production times logged in the experiment. To test the accuracy of estimating *Ranking* using Rule 1, we ordered the gestures in ascending order of production time, and correlated the resulting ranks with the median participant *Ranking*. Again, there was a strong correlation ($\rho_{(N=20)}=.94$, p=.01). Then, we applied Rule 1 for each pair of gestures (A,B) out of the (20x19)/2 = 190 possibilities, and calculated an accuracy rate (how many times the estimate was correct). In this way, estimating *Ranking* using Rule 1 attained 93% accuracy: 14 errors out of 190 tests. For Rule 2, we used the simplified Bayes parameters generated from Experiment 1 (Table 5). Estimating *Rating* using Rule 2 attained 90%: 18 gestures were correctly classified according to median participant *Rating*. The *rectangle* gesture was classified as *easy* instead of *very easy to execute*, and *tree* was classified as *easy* instead of *moderate* (both were shifted by one *Rating* class).

Next, we tested the accuracy of our rules using an estimate of production time generated from a small number of samples. Based on our analysis in the previous section, we tested N=1,3,5 participants and M=3 gesture execution samples. *Rating* accuracies varied from 66.9% to 79.8% while *Ranking* increased from 89.6% to 91.3%. Table 6 shows the accuracy rates obtained. We also re-tested using CLC and Isokoski for input to the model. CLC still produced a low *Rating* accuracy of 25%, but it performed better for *Ranking* with 75% accuracy. Isokoski did very well with 87% for *Ranking*, but cannot be used to estimate *Rating*. Overall, our rules to estimate difficulty performed well with our validation data, even when using only three samples from three participants as an estimate of production time.

Table 6. Validation experiments results: *Ranking* and *Rating* estimation accuracies using both measured and estimated production times.

Production time		Estimation Accuracy	
		Ranking	*Rating*
Measured		93.0%	90.0%
Estimated (3 executions)	x 1 participant	89.6%	66.9%
	x 3 participants	90.5%	74.6%
	x 5 participants	91.3%	79.8%
Predicted	Isokoski	87.0%	n/a
	CLC	75.0%	25.0%

9 Conclusions and Future Work

Reducing gesture execution difficulty is an often mentioned goal of gesture set design. Our work provides support for this argument with empirical evidence showing that people tend to have similar perceptions of execution difficulty, that it is highly correlated with gesture production time, and that difficulty can be estimated using two simple rules for relative ranking and a classification rating. Because existing models cannot accurately predict the magnitude of production time necessary for our classification rule, we provide evidence that an estimate of production time using only a few execution samples from a few people is good enough. Moreover, this set of estimation data may already exist when designers train a recognizer.

Designers can use our quantitative rules as they are when selecting from candidate single stroke pen gestures. However, we plan to make this process more automatic, by incorporating our difficulty estimation into the popular $1 gesture recognizer [26]. As future work, we also plan to examine how execution difficulty relates to multi-stroke pen gestures and multi-touch gestures.

Acknowledgement

This paper was supported by the project "Progress and development through post-doctoral research and innovation in engineering and applied sciences- PRiDE - Contract no. POSDRU/89/1.5/S/57083", project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

References

1. Appert, C., Zhai, S.: Using strokes as command shortcuts: cognitive benefits and toolkit support. Proceedings of CHI ’09, pp. 2289–2298. ACM Press (2009)
2. Ashbrook, D., Starner, T.: Magic: a motion gesture design tool. Proceedings of the CHI ’10, pp. 2159–2168. ACM Press (2010)

3. Bau, O., Mackay, W.E.: Octopocus: a dynamic guide for learning gesture-based command sets. Proceedings of UIST ’08, pp. 37–46. ACM Press (2008)
4. Cao, X., Zhai, S.: Modeling human performance of pen stroke gestures. Proceedings of CHI ’07, pp. 1495–1504. ACM Press (2007)
5. Castellucci, S.J., MacKenzie, I.S.: Graffiti vs. unistrokes: an empirical comparison. Proceedings of CHI ’08, pp. 305–308. ACM Press (2008)
6. Chen, P., Popovich, P: Correlation: Parametric and nonparametrized measures. Thousand Oaks, Sage (2002)
7. Grange, S., Fong, T., Baur, C.: Moris: a medical/operating room interaction system. Proceedings ICMI ’04, pp. 159–166. ACM Press (2004)
8. Isokoski, P.: Model for unistroke writing time. Proceedings of CHI ’01, pp. 357–364. ACM Press (2001)
9. Kratz, S., Rohs, M.: A $3 gesture recognizer: simple gesture recognition for devices equipped with 3d acceleration sensors. Proc. of IUI ’10, pp. 341–344. ACM Press (2010)
10. Kristensson, P.-O., Zhai, S.: Shark2: a large vocabulary shorthand writing system for pen-based computers. Proceedings of UIST ’04, pp. 43–52. ACM Press (2004)
11. Kurtenbach, G., Moran, T. P., Buxton, W.A.S.: Contextual animation of gestural commands. Computer Graphics Forum, 13(5), pp. 305–314 (1994)
12. Li, Y.: Protractor: a fast and accurate gesture recognizer. Proceedings of CHI ’10, pp. 2169–2172. ACM Press (2010)
13. Long, Jr., A.C., Landay, J.A., Rowe, L.A.: Helping designers create recognition-enabled interfaces. Multimodal interface for human-machine communication, pp. 121–146 (2002)
14. Long, Jr., A.C., Landay, J.A., Rowe, L.A.: Implications for a gesture design tool. Proceedings of CHI ’99, pp. 40–47. ACM Press (1999)
15. Long, Jr., A.C., Landay, J.A., Rowe, L.A., Michiels, J.: Visual similarity of pen gestures. Proceedings of CHI ’00, pp. 360–367. ACM Press (2000)
16. Morris, M.R., Wobbrock, J.O., Wilson, A.D.: Understanding users’ preferences for surface gestures. Proceedings of GI’10, pp. 261–268. Canadian Inf. Processing Society (2010).
17. Nielsen, M., Storring, M., Moeslund, T. B., Granum, E.: A procedure for developing intuitive and ergonomic gesture interfaces for HCI. Proceedings of GW’03, LNCS 2915, pp. 409–420. Springer Berlin / Heidelberg (2003)
18. Pratt, W.: Digital Image Processing, (3rd ed.) John Wiley & Sons, Inc. (2001)
19. Rico, J., Brewster, S.: Usable gestures for mobile interfaces: evaluating social acceptability. Proceedings of CHI ’10, pp. 887–896. ACM Press (2010)
20. Rubine, D. Specifying gestures by example. SIGGRAPH Computer Graphics, 25(4), pp. 329–337. ACM Press (1991)
21. Schomaker, L.: From handwriting analysis to pen-computer applications. IEE Electronics and Communications Engineering Journal, 10(3), pp. 93–102. Institution of Engineering and Technology (1998)
22. Viviani, P., Flash, T.: Minimum-jerk, two-thirds power law, and isochrony: converging approaches to movement planning. Journal of Experimental Psychology: Human Perception and Performance, 21(1), pp. 32–53. APA, Washington, DC (1995)
23. Viviani, P., Terzuolo, C.: 32 space-time invariance in learned motor skills. In Tutorials in Motor Behavior, vol. 1 of Advances in Psychology, pp. 525–533. North-Holland (1980)
24. Webb, A.: Statistical Pattern Recognition. John Wiley & Sons, Inc. (2002)
25. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. Proceedings of CHI ’09, pp. 1083–1092. ACM Press (2009)
26. Wobbrock, J.O., Wilson, A.D., Li, Y.: Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. Proceedings of UIST ’07, pp. 159–168. ACM Press (2007)