Comparing Human and Machine Recognition of Children's Touchscreen Stroke Gestures

Alex Shaw
Department of CISE
University of Florida
Gainesville, FL, USA 32611
alexshaw@ufl.edu

Jaime Ruiz
Department of CISE
University of Florida
Gainesville, FL, USA 32611
jaime.ruiz@ufl.edu

Lisa Anthony
Department of CISE
University of Florida
Gainesville, FL, USA 32611
lanthony@cise.ufl.edu

ABSTRACT

Children's touchscreen stroke gestures are poorly recognized by existing recognition algorithms, especially compared to adults' gestures. It seems clear that improved recognition is necessary, but how much is realistic? Human recognition rates may be a good starting point, but no prior work exists establishing an empirical threshold for a target accuracy in recognizing children's gestures based on human recognition. To this end, we present a crowdsourcing study in which naïve adult viewers recruited via Amazon Mechanical Turk were asked to classify gestures produced by 5- to 10-year-old children. We found a significant difference between human (90.60%) and machine (84.14%) recognition accuracy, over all ages. We also found significant differences between human and machine recognition of gestures of different types: humans perform much better than machines do on letters and numbers versus symbols and shapes. We provide an empirical measure of the accuracy that future machine recognition should aim for, as well as a guide for which categories of gestures have the most room for improvement in automated recognition. Our findings will inform future work on recognition of children's gestures and improving applications for children.

CCS CONCEPTS

• **Human-centered computing→Touch screens** • **Human-centered computing→ Gestural input**

KEYWORDS

Gesture recognition, touchscreen, crowdsourcing, children

ACM Reference format:

Alex Shaw, Jaime Ruiz, and Lisa Anthony. 2017. Comparing Human and Machine Recognition of Children's Touchscreen Stroke Gestures. In *Proceedings of the 19th International Conference on Multimodal Interaction, November 2017 (ICMI '17).* ACM, New York, NY, USA, 9 pages.
https://doi.org/10.1145/3136755.3136810

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ICMI '17, November 13-17, 2017, Glasgow, United Kingdom
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5543-8/17/11…$15.00
http://dx.doi.org/10.1145/3136755.3136810

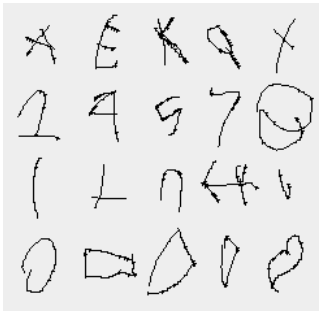


Figure 1. Examples of the gestures drawn by several 5-year-olds from the dataset we use [33]. The gestures correspond to the types in Figure 2.

1 INTRODUCTION

While many children use mainstream commercial touchscreen technology [12,13], it has been established that children exhibit different interaction behaviors on these devices than do adults [6,11,21,31,33]. For example, when touching on-screen targets, children have slower response times and higher miss rates than adults [33]. Prior work has examined the ways children use touchscreen devices to provide design guidelines to help application designers tailor their interfaces toward children [6,20,22,33]. In addition, several studies have shown that children's gesture interactions are different from those of adults [6,25,33]. Fig. 1 shows examples of some gestures produced by 5-year-old children from the dataset we use. Prior studies have found that children's gestures are recognized with lower accuracy than adults', with younger children's gestures being recognized with much less accuracy than those of older children [25,33]. For example, Woodward et al.'s [33] study of 5- to 10-year-olds reported a rate of 64% accuracy for gestures produced by 5-year-olds, and up to 91% accuracy for gestures produced by 10-year-olds. In the same study, adults' gestures were recognized with 94% accuracy. Shaw and Anthony [25] analyzed articulation features of children's and adults' gestures and found significant differences that could explain the recognition differences. For example, children had more variation in length error than adults did [25].

In gesture interfaces, users produce a gesture and a recognizer attempts to classify the gesture based on a defined set of template or training gestures. While the computer recognition algorithms employed by these interfaces can achieve high accuracy, they are not perfect, and they sometimes incorrectly classify gestures that

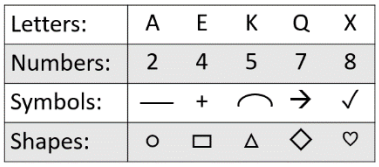


Figure 2. Gesture set from the data corpus we use [33].

could be correctly identified by humans. Thus, we set out to establish the baseline recognition of children's gestures that can be achieved by humans, to identify a target for recognition improvements. To get an idea of how accurately humans are able to recognize gestures, we wanted to get a large number of participants to classify the data. We chose a crowdsourcing [18] study to quickly reach a large, diverse pool of participants.

In this paper, we present a crowdsourcing study wherein 131 human participants classified a series of children's gestures via an online survey. The gesture dataset we used was collected by Woodward et al. [33], and consists of 2,600 examples of 20 gesture types (Fig. 2), written by 26 children ages 5 to 10 years old. We computed the recognition rates of the human survey participants and compared these rates to those of a popular recognition algorithm, $P [28]. While we choose to compare human recognition rates to $P recognition rates, our human recognition rates can be used as a benchmark in evaluating any recognition algorithm. We found a significant difference between human (90.60%) and machine (84.14%) recognition accuracy. These results identify the accuracy benchmark that future recognition algorithms should aim for in recognizing children's gestures.

Our work contributes to the growing body of knowledge on children's gesture interactions in several ways. First, we quantify human recognition accuracy of children's touchscreen gestures (90.60%), providing a baseline by which work in the area can be evaluated. We also show that the category of the gesture (letter, number, symbol, or shape) significantly affects the recognition difference between humans and machine, with humans performing much better at recognizing letters and numbers. We also investigate the gestures most often confused by humans to better understand problems in children's gesture articulation and recognition, which can help motivate the design of future recognizers. This work will inform the development of more accurate recognition of children's touchscreen stroke gestures, and will ultimately lead to improved experiences for children when using gesture-based applications.

2. RELATED WORK

We group our review of related work into three main categories: (1) recognition and analysis of children's gestures, (2) error tolerance in gesture recognition, and (3) crowdsourcing of recognition tasks, particularly using Amazon Mechanical Turk.

2.1 Children's Gestures

Several prior studies have investigated recognizing children's gestures, particularly in the context of comparing recognition accuracy to that of adults' gestures. Anthony et al. [5] examined recognition accuracy of gestures produced by children ages 7 to 11 compared to adults using the $N-Protractor recognizer [9], finding 81% accuracy for children's gestures and 90% for adults' gestures. Woodward et al. [33] compared recognition rates of gestures by children from 5 to 10 as well as adults using $P, finding 64% accuracy for the 5-year-olds, increasing by age up to 94% for 10-year-olds and 96% for adults. Children's gestures have also been used for other types of classifications; for example, a study by Kim et al. [17] identified a child's developmental level and gender from their gestures. That study did not examine recognition accuracy. Shaw and Anthony [25] compared time- and distance-based features of children's and adults' gestures and found significant differences, indicating variations in articulation patterns that may be important in understanding recognition [25].

2.2 Error Tolerance

One of the main purposes of our study is to determine an empirical accuracy level that can be used as a goal for future work in gesture recognition based on human ability to recognize children's gestures. Prior work has attempted to establish such a goal by finding the error rate that participants report as tolerable. These studies have focused on the domain of long-form handwriting recognition. One such study was that of Read et al. [24], which found that 7- to 8-year-old participants were satisfied with 91% or higher recognition accuracy. A similar study by LaLomia [19] found that adults were less tolerant of error, reporting being satisfied with 97% or higher accuracy. While these provide useful measures of ideal accuracy levels in a different interaction domain, we believe that the gesture interaction domain is similar enough to handwriting that these are still valuable goals. However, these levels are quite high compared to existing recognition rates, indicating a need to improve gesture recognition. Some gestures may be unrecognizable to both humans and machine. Establishing the accuracy with which humans can recognize gestures will help future recognition work by providing a target accuracy in classifying children's gestures.

2.3 Crowdsourcing & Human Computation

Crowdsourcing refers to the distribution of a task over a large body of participants through some online mechanism [18]. A popular platform for running crowdsourcing studies is Amazon's Mechanical Turk [3], which connects requesters with workers who perform tasks in exchange for monetary compensation. The service is used for a wide variety of purposes, such as distributing freelance work like audio transcription or sentiment analysis. The distributed nature of the system allows requesters to reach a large and diverse pool of participants. Researchers can use the system to reach more participants than possible with traditional means.

Several studies have employed crowdsourcing for recognition. For example, Novotney and Callison-Burch [23] found that Mechanical Turk participants could recognize and transcribe spoken sentences with nearly the same accuracy as a professional transcription service. In the domain of sketch recognition, Eitz et al.'s [15] study of human sketch recognition found that adults recognized complex sketches from other adults with 73.1%

accuracy. However, no prior studies have examined human recognition of children's gestures through a crowdsourcing tool.

Human computation was defined by von Ahn as "a paradigm for utilizing human processing power to solve problems that computers cannot yet solve" [1]. In our case, the problem is correctly classifying children's gestures. In human computation studies involving object classification, a certain percentage of participants must agree on the classification for a result to be associated with that object [2,18]. The researchers set the threshold based on the needs of their study. An example of a classification study using human computation is that of von Ahn and Dabbish [2], which used crowdsourcing to identify potential labels of images. Labels on which multiple participants agreed were accepted. We employ a similar methodology in our study.

3. METHOD

We now present the methodology we used in our study, including the data corpus we used and the experiment procedure.

3.1 Data Corpus

The gesture data corpus used in this experiment was collected by Woodward et al. in a study examining the effect of interface complexity on children's touchscreen interactions [33]. The gesture types, shown in Fig. 2, were chosen based on a survey of psychological and developmental literature [10], and included letters, numbers, symbols, and shapes. In that study, 30 children ages 5 to 10 produced 6 samples of each gesture type on each of two applications. One application presented a simple, abstract interface asking the children to draw gestures, whereas the other was a more complex game-like interface in which an animated bird asked the children to draw gestures. Because that study found no significant effect of interface complexity on gesture recognition, we use only the gestures collected in the abstract application. The paper also reported that six of the children were excluded from analysis due to missing data, but this number included gestures from both apps. We removed only those children who were missing data from the abstract app, a total of four. Thus, we used data from 26 children. In the rest of this paper, we refer to these 26 children as writers.

The first gesture of each type produced by each writer was considered practice and not used for recognition, leaving 5 gestures of each type produced by each writer. This gave us a total of 26 writers x 20 gesture types x 5 samples = 2,600 gestures. We used these gestures in our experiment. Of the 26 writers, there were three 5-year-olds, four 6-year-olds, three 7-year-olds, seven 8-year-olds, five 9-year-olds, and four 10-year-olds.

3.2 Experiment

In our experiment, we made image captures of each of the 2,600 gestures in the corpus, scaling them such that they had the same physical onscreen size as when they were produced. For each gesture, we created a single online survey question showing the image of the gesture and asking which of the 20 possible gesture types the gesture most resembled. Participants were required to select one of the options, and could guess if they had no idea.

Because it is not practical for each participant to answer 2,600 survey questions, the questions viewed by each participant were randomly selected such that each participant saw exactly 20 gestures from each age group, and of those 20 gestures, each represented exactly one of the distinct gesture types. Thus, each participant was asked a question about each gesture type for 5-, 6-, 7-, 8-, 9-, and 10-year-olds, a total of 6 age groups x 20 gesture types = 120 questions, in a randomized order. Participants were not told the age of the creators of the gestures, and they did not know that they would see an equal number of each gesture type. In fact, participants did not even know the gestures were produced by children. The random selection of gestures for the survey was such that each of the gestures was seen approximately the same number of times as other gestures from the same age group across the study. Each gesture in our corpus was evaluated by at least 3 participants. To ensure this, we initially recruited a small sample of approximately 50 people, then looked at which gestures still had fewer than 3 responses. We redeployed the survey with only those questions that still needed to be evaluated to reach at least 3 participants. We repeated this process until all gestures had been evaluated by at least 3 participants (max: 32 due to some categories having a small number of gestures). No participant was allowed to take more than one version of the survey. Table 1 shows that similar accuracy rates were found for the gestures evaluated by a smaller number of participants (3-5) and a larger number (6+), and a paired t-test found no significant difference in accuracy between the two categories ($t(5) = 1.3$, *n.s.*).

All of the participants in our study were recruited through Amazon's Mechanical Turk crowdsourcing platform. Participants were paid $2.00 for completing the full 120-question survey, which was designed to take approximately 15 minutes (a rate of $8.00 per hour). Some later participants completed fewer questions to help balance the dataset, in which they were paid the same per-question rate as previous participants, about two cents per question. In total, there were 131 unique participants. Our protocol was approved by our Institutional Review Board.

We also ran automatic recognition experiments using $P [28], a multistroke point-matching based algorithm that is used by both developers and researchers due to its high accuracy and ease of implementation. $P is scale, direction, and translation invariant [28]. In machine recognition experiments, a number of training examples must be provided for the recognizer to use as reference when classifying gestures; in our experiment, we vary the number of training samples from the minimum (1 per gesture) to the maximum (in our case 4, since there are 5 of each type of gesture per writer and 1 must be chosen for testing). We then averaged

Table 1. Human accuracy of gestures seen by a small number of participants versus a larger number.

	% Human Accuracy (SD)	
Age (Years)	**3-5 Participants**	**6+ Participants**
5	74.99 (39.56)	75.30 (34.89)
6	84.59 (29.80)	82.09 (31.08)
7	90.52 (19.44)	92.31 (20.41)
8	90.99 (21.19)	91.74 (17.52)
9	94.28 (16.89)	87.05 (21.43)
10	97.51 (9.47)	94.08 (12.81)

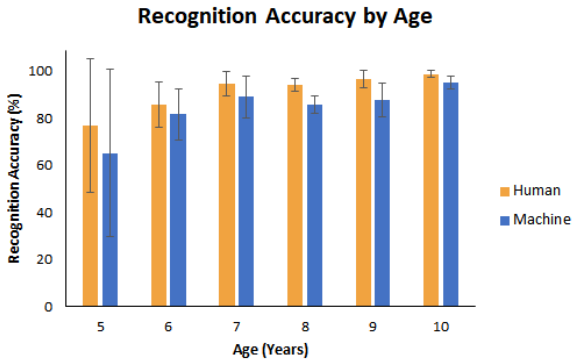


Figure 3. Effect of age and recognizer (human vs. machine) on accuracy. Error bars represent the 95% confidence interval.

the recognition accuracy for each of the different values of number of training samples, giving an average for each gesture for each writer in the dataset.

In this study, we use user-dependent tests for machine recognition. In user-dependent tests, all training examples are taken from a single writer's gestures, and the test gesture is taken from that same writer. We compare human recognition to user-dependent accuracy, because user-dependent accuracy is generally much higher than other methods [4,33]. This allows us to compare human accuracy to the best-case recognition results for the machine; user-independent accuracy would be lower.

4. ANALYSIS AND RESULTS

To understand how well humans recognize children's gestures, we first determined whether each individual gesture was recognized correctly. To do this, we looked at the survey takers' categorization of each individual gesture from each writer. If at least half of the survey takers who saw that gesture identified it correctly, it was counted as a correct recognition, otherwise it was incorrect. Ties were broken by having an additional participant classify the gesture as described above. We thus represented each gesture in our writer corpus as either correctly recognized (labeled 1) or incorrectly recognized (labeled 0). For each child in the writer dataset, we computed the average accuracy for each type of gesture produced by that writer, then the overall accuracy of recognition of that writer's gestures. We compared the per-writer average recognition rate by humans to the per-writer average recognition rate by machine. The majority-based method we employ has been used in previous crowdsourcing studies [16,26].

We now discuss the results of our analyses. All of the factors in our study were analyzed using the same three-way repeated-measures ANOVA on recognition accuracy with a between-subjects factor of *age* and within-subjects factors of *recognizer type* (human vs. machine) and *gesture category* (the general type of each gesture; discussed further in section 4.2). All references to our ANOVA refer to this single test.

4.1 Recognition by Recognizer Type

For overall accuracy including all age groups, the human recognition rate was 90.60%, compared to 84.14% for machines. Our ANOVA found a significant main effect of *recognizer type* (human vs. machine) on recognition accuracy ($F_{1,20} = 42.197$, $p < 0.001$). The same test found a marginal effect of *age* ($F_{5,20} = 2.485$, $p = 0.066$) and no significant interaction between age and recognizer type ($F_{1,48} = 0.761$, *n.s.*). Fig. 3 illustrates the overall recognition rates for each of the age groups in our study, showing a higher accuracy for each group in the case of human recognition. The significant gap between human and machine recognition is illustrative of the potential for improvement of machine recognizers.

A Tukey post-hoc test on the interaction between *age* and *recognizer type* (human vs. machine) found a significant difference in recognizer type for 5-year-olds, 8-year-olds, and 9-year-olds ($p < 0.05$). There was also a marginal difference for 6-year-olds ($p < 0.1$). The mean difference in recognition accuracy between human and machine is greatest for the youngest group, the 5-year-old writers. For the gestures produced by writers in this age group, the machine recognizer achieved only 65.30% accuracy, compared to 74.63% accuracy for humans, a difference of nearly 10%. While the human recognition accuracy is higher than the machine accuracy, even the human accuracy for 5-year-olds is far below what children report as an acceptable error level, 91% [24]. However, our work provides a realistic target for future work in machine recognition. Later work could focus on achieving higher accuracy for machine recognition. Because age was only marginally significant, we focus our discussion on gesture category and recognizer type (human vs. machine).

4.2 Recognition by Gesture Category

To better understand the types of gestures humans are best able to recognize compared to machines, we also examined recognition rates by category. We grouped each of the 20 gesture types in our analysis into one of four general categories of gestures: letters (A, E, K, Q, X), numbers (2, 4, 5, 7, 8), shapes (circle, square, triangle, diamond, heart), and symbols (line, plus, arch, arrowhead, checkmark). Each row in Fig. 2 represents a different category. These categories are consistent with those described by the gesture set's creators [28].

We hypothesized that children would be more familiar with letters and numbers than shapes and symbols, and that therefore they may be less skilled at creating the latter two categories of

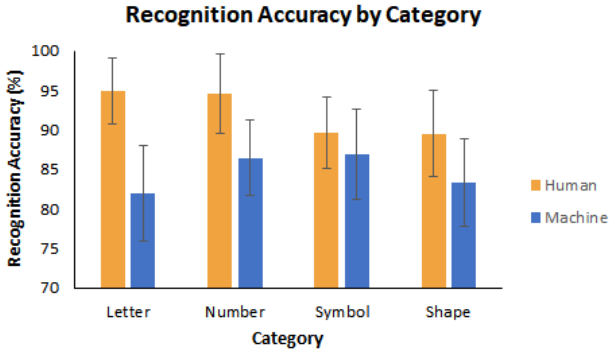


Figure 4. Effect of recognizer (human vs. machine) and category on recognition accuracy. Error bars represent the 95% confidence interval.

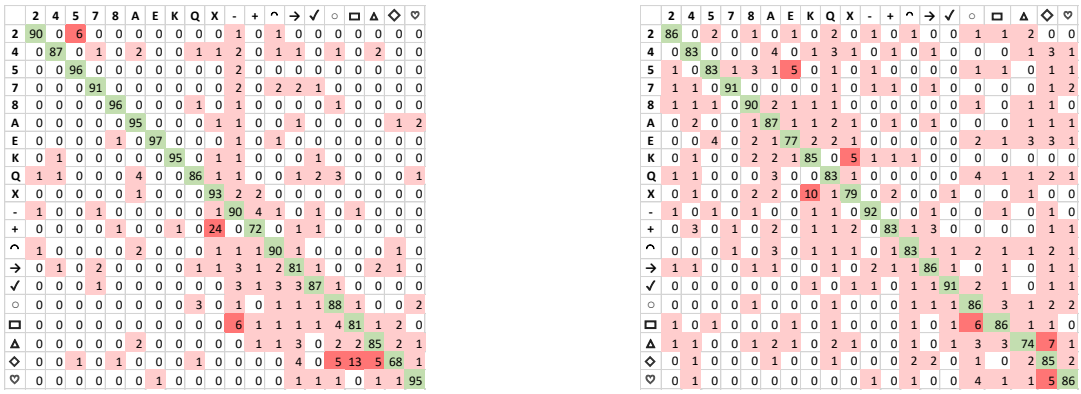


Figure 5. Confusion matrix for human recognition (left) and machine recognition (right) of the gestures in the dataset. Each cell represents the percentage of times the gesture in the row label was recognized as the gesture in the column label. Values are rounded to the nearest integer.

gestures. The younger children in the dataset, in particular, may have still been developing their ability to produce shapes and symbols [10]. Humans are constantly presented with different variations of letters, numbers, symbols, and shapes. Therefore, humans should have superior ability at recognizing our gesture set than machines, which are often trained on a limited number of examples for each gesture (e.g., up to 4 examples in our study). Thus, we hypothesized that, though both humans and machines would be affected by deviations (e.g., because of children's articulation issues [25]), machines would be more affected.

Our ANOVA found a significant main effect of *gesture category* on accuracy ($F_{3,60} = 3.093$, $p < 0.05$). We also found a significant interaction between *recognizer type* (human vs. machine) and *gesture category* ($F_{3,72} = 5.745$, $p < 0.05$). Fig. 4 shows the average recognition rates for each gesture category by recognizer type. Humans were even more accurate in classifying letter and number gestures than the other gesture categories as compared to the machine. We believe there are two reasons for this: first, the children tend to learn to draw letters and numbers before learning shapes and symbols, meaning they have more practice in creating them. Secondly, the ubiquity of letters and numbers means that humans will inherently have more practice interpreting letters and numbers drawn by others than shapes and symbols. Prior work has shown significant differences in the articulation features of gestures, based on whether the user was familiar or unfamiliar with the gesture [7,29,30]. For example, Vatavu et al. [29] found that users had a lower level of variability in speed of execution of gestures that they were familiar with compared to unfamiliar gestures. Thus, humans may be better able to recognize gestures that are more familiar to the writer due to the writer's increased fluency in creating those gestures.

The wide gap between human and machine recognition rates, particularly for letters and numbers, shows the importance of improving the accuracy of recognition of children's gestures. As previously mentioned, Read et al.'s [24] prior work shows that children are not satisfied with rates below 91% accuracy, a target attained by humans in our study but not machines. By showing that humans can achieve this accuracy, we show that it is a worthwhile goal to work toward in machine recognition. If humans were not able to meet this mark, we would be unsure if it is a reasonable target for accuracy. Thus, we establish that current recognition rates should not be accepted as optimal, but instead that future work should push toward achieving higher accuracy.

4.3 Confusion Matrices

To dig into specific examples of the issues that humans and machines had in classifying the children's gestures in our study, we created and examined confusion matrices. A confusion matrix [27] (p. 209) is used to visualize the frequency with which each pair of gestures is recognized as one another. The dataset had 20 gestures, so our confusion matrix has 20 rows and 20 columns. The rows of the matrix represent the correct answer, and the columns represent the recognition response. The value in each cell represents the percentage of times that all gestures of that corresponding row label were identified as the corresponding column label. Thus, each row's values should sum to 100%; we round to the nearest integer in our matrix for simplicity, so the sum is ±4% for some rows.

Fig. 5 shows the confusion matrices for the human (left) and machine (right) recognition accuracy in our experiment. The cells in the main diagonal (highlighted in green) represent correct recognitions. These cells generally contain high values, while other cells generally contain low values since incorrect recognitions are not as common as correct recognitions. We provide the machine confusion matrix for reference, but we focus our discussion on the human confusion matrix.

In the human confusion matrix, only six cells outside of the main diagonal have at least 5% confusion (darker red shading). Each of these provides interesting insight into recognition mistakes made by humans, so we now discuss each of these commonly confused types. For the rest of this section, we use the term *gesture* to refer to one of the 20 types shown in Fig. 2, whereas we use the term *instance* to refer to a specific example of one of these gestures produced by a writer.

"Plus" Confused for "X". The "plus" gesture was confused for the "X" gesture 23% of the time. In total, this misrecognition occurred a total of 148 times over 72 different instances. The confusion is likely due to the resemblance between the two

gestures (the same instance could be seen as an "X" or a "plus" depending on the viewer's orientation). Conversely, the "X" gesture was only confused for the "plus" gesture 1% of the time. One possible explanation for this discrepancy could be that children have difficulty drawing the lines perfectly vertically or horizontally on a touchscreen, causing them to appear slanted, thus resulting in many participants classifying them as "X" gestures. In some cases, shown in Fig. 6a, the reason for the misrecognition is quite apparent. The "plus" gestures are slanted and appear much more like "X" gestures. Other cases, however, are much less clear cut, with mostly vertical and horizontal lines, yet they were still often misclassified. Another reason for this could be that the "X" option was the last choice in our survey, so it may have "jumped out" at users more readily than the "plus" option, which was in the middle of the choices.

"2" Confused for "5". The "2" gesture was confused for the "5" gesture 6% of the time. In total, this misrecognition occurred 41 times over 6 different instances, and 37 of those 41 misrecognitions were of the same writer's instances. Examining these instances, indicated in Fig. 6b, it becomes quite clear why this was such a common mistake. Although these are meant to be "2" gestures, it appears to a human that the writer drew "5" gestures. We believe that this mistake was either due to the child drawing the "2" gestures backwards (mirror writing [14]), causing it to resemble a "5", or by the child simply drawing the wrong gesture. Interestingly, this child's "5" gestures resembled standard gestures of the same type and were well recognized.

"Rectangle" Confused for "Line". The "rectangle" gesture was confused for the "line" gesture 6% of the time. In total, this misrecognition occurred 39 times over 6 different instances. Fig. 6c shows the "rectangle" instances that were most confused for "line" gestures. In most of these cases, the child drew the rectangle without vertical lines, causing the instance to resemble an equals sign. Since this was not an option in the survey, the most logical next choice for many of the survey participants was to select the line option, since an equals sign is composed of two lines. The distance between the two lines may have caused participants to perceive the instance as two pieces rather than a single unit.

"Diamond" Confused for "Circle". The "diamond" gesture was confused for the "circle" gesture 5% of the time. This misrecognition occurred a total of 41 times over 6 instances, most of which were produced by 5- and 6-year-olds. These younger children may have been unfamiliar with the diamond shape compared to most of the other gestures in the set. A circle, rectangle, or triangle is more basic than a diamond, and children are likely to have more experience drawing these shapes. Supporting this idea, previous work using the same gesture set identified the diamond gesture as being difficult for children to draw [4]. Fig. 6d shows examples of some of the "diamond" instances that were most commonly confused for "circle" gestures. While none of these look very similar to a circle, they do have more "rounded" sides than other gestures, which may have led some participants to classify them as "circle" gestures.

"Diamond" Confused for "Rectangle". The "diamond" gesture was confused for the "rectangle" gesture 12% of the time. This misrecognition occurred a total of 81 times over 48 different instances, spanning all age groups. Again, the high level of confusion may be due to the relative unfamiliarity of the "diamond" gesture compared to the other shapes in our corpus. Geometrically, a diamond and a rectangle are similar in that they are both quadrilaterals, which could partially account for some of the confusion we see in this case. Fig. 6e shows examples of some of the most commonly confused instances in this set. While most of these instances are quadrilaterals that do resemble the traditional definition of a diamond, it is not unreasonable to think that they could be interpreted as slanted rectangles. A participant rushing through the survey may very well see the "rectangle" option before the "diamond" option, and decide to select that option without carefully considering the other options. In the case of the instance in the middle of the left column of the figure, the intent of the user is less clear. The long, straight line is likely to be disregarded by the participant as an error, leaving a rectangle.

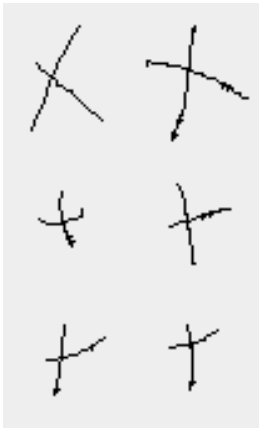
"Diamond" Confused for "Triangle". The "diamond" gesture was confused for the "triangle" gesture 5% of the time. This misrecognition occurred a total of 81 times over 19 different instances. As in the prior two cases, the confusion of this pair may be due to the relative shape complexity compared to the other shapes in our gesture set. Fig. 6f shows examples of "diamond" instances that were commonly confused for "triangle" gestures. Most of these instances look very similar to triangles, indicating that the child either drew the wrong gesture or did not know how to properly draw the diamond. In the last case, the child drew a diamond, but added a line through the middle. The extra line likely led many participants to interpret the instance as two triangles rather than a single diamond, leading to confusion of the gestures.

5. DISCUSSION

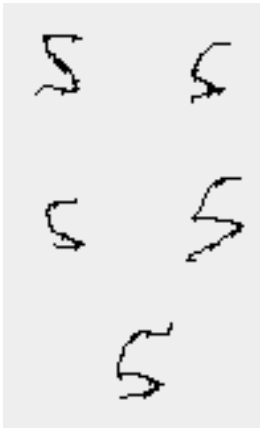
We now discuss our finding and their implications, as well as the limitations and conclusions of our work, and conclude with a short discussion of our future work.

5.1 Human vs. Machine Recognition

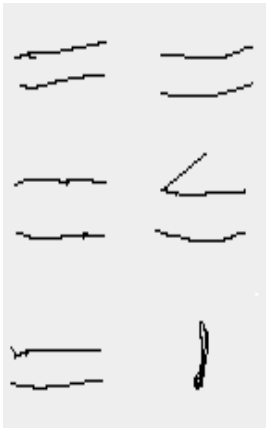
Our work shows a significant difference between machine and human recognition of children's gestures. We also found a significant effect of gesture category on recognition, and a significant interaction between gesture category and recognizer type. Our study provides an empirically established threshold by which future recognition algorithms can be judged: 90.60% for 5- to 10-year-olds by humans, compared to 84.14% by machines. We also show that letters and numbers are the gesture categories with the biggest gaps between human and machine recognition, indicating the potential for larger increases in recognition of these categories of gestures. Our experiment provides a good idea of the range of accuracy we can expect for various age groups from 5 to 10 and the accuracy that new recognizers should aim for. That said, as Read et al. [24] and LaLomia's [19] work shows, users are willing to tolerate a certain level of error in recognition. Interestingly, the 90.60% accuracy achieved by humans in our study is very similar to the 91% reported as acceptable by children in Read et al.'s study [24]. However, while we achieved an overall accuracy rate of 90.60% for children's gestures, human accuracy



(a) "Plus" instances commonly confused for "X"

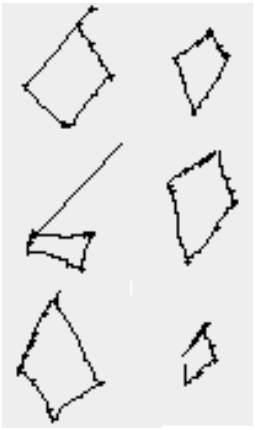


(b) "2" instances commonly confused for "5"

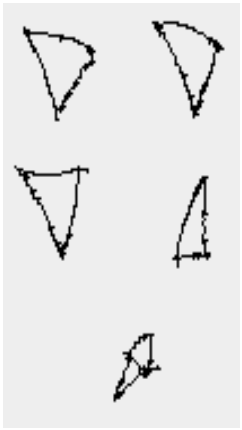


(c) "Rectangle" instances commonly confused for "line"

(d) "Diamond" instances commonly confused for "circle"



(e) "Diamond" instances commonly confused for "rectangle"



(f) "Diamond" instances commonly confused for "triangle"

Figure 6. Images of the gestures most commonly confused for others by humans.

was much lower for the younger children's gestures than older children's. Human accuracy was 80.97% for 5-year-olds, 83.20% for 6-year-olds, 95.67% for 7-year-olds, 91.52% for 8-year-olds, 93.80% for 9-year-olds, and 98.24% for 10-year-olds. While humans are able to recognize gestures by children of all ages better than machines, human accuracy for 5- and 6-year-olds' gestures is under the 91% tolerance level reported by Read et al. [24]. The low recognition rates of the youngest children's gestures by humans is likely due to the children's articulation features, which affects both human and machine recognition. These findings suggest a need to explore new recognition techniques that can better handle the inconsistencies in these children's gestures.

Our human recognition results for each individual gesture in the dataset are available at http://init.cise.ufl.edu/downloads.

5.2 Commonly Confused Pairs

Since the types of gestures that were most confused varied, it is difficult to draw general conclusions about what particular habits of children cause these recognition errors. For example, confusion of "rectangle" gestures for "line" gestures appears to be mainly due to one child's drawing the rectangles without vertical lines, causing viewers to interpret them as two horizontal lines.

In the case of "plus" being confused for "X", a greater number of writers' gestures were involved. Overall, humans recognized "plus" gestures with only 71% accuracy, compared to 83% for machine (this gesture is an example of one that was recognized more accurately by machine than human, although on average over all gestures human recognition was higher). The machine algorithm we use, $P [28], can detect differences between rotationally similar gestures like "plus" and "X", but not all recognizers can do so [8,32]. Also, in our study humans had more difficulty distinguishing between pairs of rotated gestures. This high level of confusion in humans may represent a bias toward the "X" gesture over "plus", but the machine recognizer has no such bias. We saw a similar pattern for the "diamond" gesture: depending on rotation, it may resemble a rectangle, and it was recognized with only 67% accuracy by humans versus 85% by machine. These types of confusions may represent a bias in humans toward one gesture over another when they are rotationally similar (e.g., the more common gesture is chosen first). Based on this finding, we recommend that gesture set designers carefully consider whether using gestures that are rotationally similar in an application is necessary: if the children will likely exhibit inconsistent rotation behavior, the recognition algorithm chosen may or may not be able to detect the difference.

The fact that the "diamond" gesture was so widely confused points to another interesting issue. No other gestures were commonly (more than 5% of the time) confused with more than one other gesture, yet diamond was confused with three different shapes at least 5% of the time each. We speculate that many children know that a diamond is a shape, but are not familiar with exactly how to draw it, leading their gesture to resemble other shapes in the survey. Another potential explanation is that participants may have expected diamond shapes to be more like "gem" or "jewel" shapes when taking the survey. In any case, the

confused pairs provide interesting insight into the differences between human and machine recognition that result from children's gesture articulation patterns. Based on these findings, we recommend that gesture set designers carefully consider how familiar children will be with certain gestures, as children of a young age will not yet have developed fluency with basic shapes.

5.3 Limitations

While our results have important and interesting implications for future recognition algorithms, there are several limitations that must also be considered. We chose $P [28] for comparison in this work, but it may also be useful to compare to other, potentially more accurate recognizers. Another limitation is that we focus on only a subset of children, specifically 5- to 10-year-olds. While this is certainly an important age group (due to the cognitive and motor developmental changes in this age group [10] and the fact that the youngest children in this group are just starting school), our work shows that the gap between human and machine recognition is higher for younger children, and as such we believe that the gap would be even wider for children under 5. Because even children much younger than 5 often use touchscreens, we believe this is an important topic for further investigation.

Another limitation, discussed in our section on confusion matrices, is that some of the gestures appear to have been poorly drawn by the writer in such a way that they resembled a different gesture in the set. The poorly drawn gesture, in turn, leads to less accurate recognition since the gesture will likely be classified as what it most resembles rather than the correct gesture. Fig. 6b shows an example of this type of case: the child was prompted to draw "2" gestures, but instead drew what appears to be "5" gestures. This so-called "mirror-writing" occurs in the majority of children between 3 to 7 years old [14]. It is certainly important to consider these types of responses in terms of evaluating results, as it is a realistic limitation that must be considered in real systems. Users, particularly children, may sometimes produce the "incorrect" gesture, but there is no way of knowing whether the error is due to a misunderstanding on the part of the writer in how the gesture is drawn or due to the writer drawing the wrong type of gesture. Thus, it is only fair that we include these gestures in our analysis, since excluding them would unfairly inflate recognition rates. To examine the extent to which this limitation affected our data, we examined each of the 2,600 gestures and labeled them as either (a) correct, (b) wrong (e.g., could be another gesture type in our set), (c) malformed, or (d) illegible. We found 91.15% fell in category (a), 2.28% in (b), 3.87% in (c), and 2.28% in (d). Category (b) made up only 2.28% of the gestures, so it could not be the main cause of the 9.40% error for humans.

Another potential limitation of our study is that we did not randomize the order of choices in our survey, so some choices may have stood out more than others based on their location. We chose not to randomize the order so that the participants would be able to learn the position of each of the choices and locate them quickly, allowing them to finish the survey with less frustration.

One final limitation of our study is the inherent possibility that some Mechanical Turk participants may try to 'scam' the system by, e.g., randomly picking a response for each question to quickly complete the survey and receive compensation. We planned to address this issue by excluding results from participants who scored lower than two standard deviations from the mean accuracy, but none of them fell below this threshold. However, it is possible that some of the participants in our study rushed through at least a portion of the survey.

5.4 Future Work

In our future work, we plan to use the empirical recognition accuracies found in this study as a standard by which to evaluate our future recognition algorithms. We are particularly interested in developing improved recognition for young children, and understanding human ability to recognize the gestures provides a benchmark by which these algorithms can be judged. We are interested in extending this study to younger children to see how large the discrepancies between human and machine rates are for very young children in the age range of 2 to 5 years old.

Furthermore, our confusion matrices (Fig. 5) help establish some of the most commonly confused pairs of gestures in both the human and machine case. We found that rotationally similar gestures, like "X" and "plus", were often confused, especially by humans. We also found that both humans and machines had a high level of confusion within the shape category, which leads us to conclude that children have more trouble drawing these gestures than other categories. Our examination of these frequently confused pairs helps establish the types of errors that are most prevalent based on children's gesture patterns, allowing developers to better predict which types of gestures will be confused and improve gesture set design. Future work in recognition can also use this information about common mistakes to create algorithms tailored toward children's gestures.

6. CONCLUSION

We presented a comparison of human versus machine ability to recognize gestures produced by children ages 5 to 10. We found a significant difference between human (90.60%) versus machine (84.14%) accuracy. We also found a significant effect of gesture category (letter, number, symbol, or shape) on accuracy, as well as a significant interaction between recognizer type and category. Machine accuracy was consistently lower than human accuracy for all age groups. The difference between human and machine recognition was more pronounced for younger children than older children, indicating considerable potential for improvement in recognition of younger children's gestures. Our work will inform the design and development of improved recognition algorithms for children's gestures, and ultimately lead to improved gesture interaction experiences for children.

ACKNOWLEDGMENTS

This work is partially supported by National Science Foundation Grant Awards #IIS-1433228 and #IIS-1552598. Opinions, findings, and conclusions or recommendations in this paper are those of the authors and do not necessarily reflect these agencies' views.

REFERENCES

[1] Luis von Ahn. 2005. Human Computation. *PhD thesis, School of Computer Science, Carnegie Mellon University.*

[2] Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. *Proceedings of the International Conference on Human Factors in Computing Systems (CHI '04)*, ACM Press, 319–326. http://doi.org/10.1145/985692.985733

[3] Amazon. Mechanical Turk. Retrieved from www.mturk.com

[4] Lisa Anthony, Quincy Brown, Jaye Nias, and Berthel Tate. 2013. Examining the need for visual feedback during gesture interaction on mobile touchscreen devices for kids. *Proceedings of the International Conference on Interaction Design and Children (IDC '13)*, ACM Press, 157–164. http://doi.org/10.1145/2485760.2485775

[5] Lisa Anthony, Quincy Brown, Jaye Nias, Berthel Tate, and Shreya Mohan. 2012. Interaction and recognition challenges in interpreting children's touch and gesture input on mobile devices. *Proceedings of the ACM international Conference on Interactive Tabletops and Surfaces (ITS '12)*, ACM Press, 225–234. http://doi.org/10.1145/2396636.2396671

[6] Lisa Anthony, Quincy Brown, Jaye Nias, and Bethel Tate. 2015. Children (and Adults) Benefit From Visual Feedback During Gesture Interaction on Mobile Touchscreen Devices. *International Journal of Child-Computer Interaction.* 6, 17–27. http://doi.org/10.1016/j.ijcci.2016.01.002

[7] Lisa Anthony, Radu-Daniel Vatavu, and Jacob O. Wobbrock. 2013. Understanding the Consistency of Users' Pen and Finger Stroke Gesture Articulation. *Proceedings of Graphics Interface (GI '13)*, Canadian Information Processing Society, 87–94.

[8] Lisa Anthony and Jacob O. Wobbrock. 2010. A lightweight multistroke recognizer for user interface prototypes. *Proceedings of Graphics Interface (GI '10)*, Canadian Information Processing Society, 245–252.

[9] Lisa Anthony and Jacob O. Wobbrock. 2012. $N-protractor: a fast and accurate multistroke recognizer. *Proceedings of Graphics Interface (GI '12)*, Canadian Information Processing Society, 117–120.

[10] Keith E. Beery, Norman A. Buktenica, and Natasha A. Beery. 2004. *The Beery-Buktenica Developmental Test of Visual-Motor Integration, 5th Edition.* Modern Curriculum Press, New Jersey.

[11] Robin Brewer, Lisa Anthony, Quincy Brown, Germaine Irwin, Jaye Nias, and Berthel Tate. 2013. Using gamification to motivate children to complete empirical studies in lab environments. *Proceedings of the International Conference on Interaction Design and Children (IDC '13)*, ACM Press, 388–391. http://doi.org/10.1145/2485760.2485816

[12] Francette Broekman, Jessica Piotrowski, Hans Beentjees, and Patti Valkenburg. 2016. A parental perspective on apps for young children. *Computers in Human Behavior* 63, 142–151. http://doi.org/10.1016/j.chb.2016.05.017

[13] Cynthia Chiong and Carly Shuler. 2010. *Learning: Is there an app for that? Investigations of young children's usage and learning with mobile devices and apps.* The Joan Ganz Cooney Center at Sesame Workshop, New York, NY. Retrieved August 23, 2012 from http://dmlcentral.net/resources/4496

[14] James M Cornell. 1985. Spontaneous Mirror-Writing in Children. *Canadian Journal of Psychology* 39, 1, 174–179. http://doi.org/10.1037/h0080122

[15] Mathias Eitz, James Hays, and Marc Alexa. 2012. How do humans sketch objects? *ACM Transactions on Graphics* 31, 4, 1–10. http://doi.org/10.1145/2185520.2335395

[16] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining human and machine intelligence in large-scale crowdsourcing. *AAMAS '12 Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS '12)*, 467–474.

[17] Hong-hoe Kim, Paul Taele, Stephanie Valentine, Erin McTigue, and Tracy Hammond. 2013. KimCHI: a sketch-based developmental skill classifier to enhance pen-driven educational interfaces for children. *Proceedings of the International Symposium on Sketch-Based Interfaces and Modeling (SBIM '13)*, ACM Press, 33–42. http://doi.org/10.1145/2487381.2487389

[18] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies With Mechanical Turk. *Proceedings of the Interational Conference on Human Factors in Computing Systems*, ACM Press, 453–456.

[19] Mary LaLomia. 1994. User acceptance of handwritten recognition accuracy. *Extended Abstracts of the ACM SIGCHI Conference on Human Factors in Computing Systems*, ACM Press, 107–108. http://doi.org/http://doi.org/10.1145/259963.260086

[20] Lorna McKnight and Brendan Cassidy. 2010. Children's Interaction with Mobile Touch-Screen Devices. *International Journal of Mobile Human Computer Interaction* 2, 2, 18 pp. http://doi.org/10.4018/jmhci.2010040101

[21] Lorna McKnight and Daniel Fitton. 2010. Touch-screen Technology for Children: Giving the Right Instructions and Getting the Right Responses. *Proceedings of the International Conference on Interaction Design and Children (IDC '10)*, ACM Press, 238–241. http://doi.org/10.1145/1810543.1810580

[22] Vicente Nacher, Javier Jaen, Elena Navarro, Alejandro Catala, and Pascual Gonzalez. 2015. Multi-touch gestures for pre-kindergarten children. *International Journal of Human Computer Studies* 73, 37–51. http://doi.org/10.1016/j.ijhcs.2014.08.004

[23] Scott Novotney and Chris Callison-Burch. 2010. Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '10)*, Association for Computational Linguistics, 207–215.

[24] Janet C. Read, Stuart MacFarlane, and Chris Casey. 2003. "Good enough for what?": acceptance of handwriting recognition errors by child users. *Proceedings of the International Conference on Interaction Design and Children*, ACM Press, 155–155. http://doi.org/10.1145/953536.953565

[25] Alex Shaw and Lisa Anthony. 2016. Analyzing the Articulation Features of Children's Touchscreen Gestures. *Proceedings of the International Conference on Multimodal Interaction (ICMI '16)*, ACM Press, 333–340. http://doi.org/10.1145/2993148.2993179

[26] Rion Snow, Brendan O Connor, Daniel Jurafsky, Andrew Y Ng, Dolores Labs, and Capp St. 2008. Cheap and fast - but is it good? Evaluating non-expert annotiations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 254–263. http://doi.org/10.1.1.142.8286

[27] Kai Ming Ting. 2010. Confusion Matrix. In *Encyclopedia of Machine Learning*, Claude Sammut and Geoffrey I Webb (eds.). Springer US, Boston, MA.

[28] Radu-Daniel Vatavu, Lisa Anthony, and Jacob O. Wobbrock. 2012. Gestures as point clouds: a $P recognizer for user interface prototypes. *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '12)*, ACM Press, 273–280. http://doi.org/10.1145/2388676.2388732

[29] Radu-Daniel Vatavu, Lisa Anthony, and Jacob O. Wobbrock. 2013. Relative accuracy measures for stroke gestures. *Proceedings of the ACM International Conference on Multimodal Interaction*, ACM Press, 279–286. http://doi.org/10.1145/2522848.2522875

[30] Radu-Daniel Vatavu, Daniel Vogel, Géry Casiez, and Laurent Grisoni. 2011. Estimating the perceived difficulty of pen gestures. *INTERACT*, Springer-Verlag, 89–106.

[31] Radu Daniel Vatavu, Lisa Anthony, and Quincy Brown. 2015. Child or adult? Inferring Smartphone users' age group from touch measurements alone. *Proceedings of the International Conference on Human-Computer Interaction (INTERACT '15)*, 1–9. http://doi.org/10.1007/978-3-319-22723-8_1

[32] Jacob O. Wobbrock, Andrew D. Wilson, and Yang Li. 2007. Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '07)*, ACM Press, 159–168. http://doi.org/10.1145/1294211.1294238

[33] Julia Woodward, Alex Shaw, Annie Luc, Brittany Craig, Juthika Das, Phillip Hall, Akshay Hollay, Germaine Irwin, Danielle Sikich, Quincy Brown, and Lisa Anthony. 2016. Characterizing How Interface Complexity Affects Children's Touchscreen Interactions. *Proceedings of the ACM International Conference on Human Factors in Computing Systems (CHI '16)*, ACM Press, 1921–1933. http://doi.org/10.1145/2858036.2858200