# Motor-Impaired Touchscreen Interactions in the Wild

Kyle Montague[1], Hugo Nicolau[1], Vicki L. Hanson[1, 2]

[1]School of Computing
University of Dundee
Dundee, DD1 4HN, Scotland
{kylemontague, hugonicolau}@computing.dundee.ac.uk

[2]Golisano College of Computing and Information Sciences
Rochester Institute of Technology, 20 Lomb Memorial
DriveRochester, NY USA 14623
vlh@acm.org

## ABSTRACT

Touchscreens are pervasive in mainstream technologies; they offer novel user interfaces and exciting gestural interactions. However, to interpret and distinguish between the vast ranges of gestural inputs, the devices require users to consistently perform interactions inline with the predefined location, movement and timing parameters of the gesture recognizers. For people with variable motor abilities, particularly hand tremors, performing these input gestures can be extremely challenging and impose limitations on the possible interactions the user can make with the device. In this paper, we examine touchscreen performance and interaction behaviors of motor-impaired users on mobile devices. The primary goal of this work is to measure and understand the variance of touchscreen interaction performances by people with motor-impairments. We conducted a four-week in-the-wild user study with nine participants using a mobile touchscreen device. A Sudoku stimulus application measured their interaction performance abilities during this time. Our results show that not only does interaction performance vary significantly between users, but also that an individual's interaction abilities are significantly different between device sessions. Finally, we propose and evaluate the effect of novel tap gesture recognizers to accommodate for individual variances in touchscreen interactions.

## Categories and Subject Descriptors

H.5.2 **Information Interfaces and Presentation**: User Interfaces – *Input devices and strategies.* K.4.2 **Computers and Society**: Social Issues – *Assistive technologies for persons with disabilities.*

## General Terms

Measurement, Design, Experimentation, Human Factors.

## Keywords

Touchscreen; Motor-Impaired; In-the-Wild; User Models.

## 1. INTRODUCTION

Touchscreen devices have become the norm for mobile technologies, with smartphones and tablets among the most popular. Companies are increasingly delivering their services and products via touchscreen technologies, meaning that those unable to access them are being excluded and missing out on the advantages that these devices can offer. Touchscreens use gesture recognizers to interpret and respond to a wide variety of touch-based inputs. However, in order to accurately interpret and respond, the gesture recognizers rely on the user being able to consistently perform the touch actions as they have been defined by the device manufacturer and/or application developer.

Previous work investigating mouse interactions by people with motor-impairments demonstrated that consistent performance of interactions was not always possible as abilities were highly variable and erratic [6]. While works have investigated touchscreen interactions by people with motor-impairments, they have relied on a single session laboratory user study design, producing snapshot measurements of touchscreen performances, It is therefore not understood how variable motor abilities impact touchscreen interaction performance over time.

To address this knowledge gap, we conducted a four-week in-the-wild user study, involving nine participants with motor-impairments, to understand their interaction behaviors and measure performance abilities across multiple sessions. To the best of our knowledge, this is the first study of its kind, and offers new insights into the variable performance of touchscreen interactions by people with motor-impairments. The user study applied a novel approach, measuring user abilities of touchscreen interactions from typical device interactions with a Sudoku game. This approach allowed data collection to occur without the need for calibration tasks, and enabled greater number of collection periods than conventional laboratory studies. Using this approach, we ensure that our measurements reflect the true nature of variance associated with motor-impairments, and refrain from simply gathering a snapshot of the individual's abilities.

Our results demonstrate that not only do individuals with motor-impairments vary significantly on tap gesture performance and interaction behaviors, but also that the individual's performance varies significantly between interaction sessions. Based on these findings, we proposed novel methods to individually tailor the tap gesture recognizers to the ever-changing abilities of users. Finally, through simulations of novel tap gesture recognizers, we were able to achieve a recognition accuracy of 97%, significantly greater than the device default recognizer, thus improving touchscreen performance for individuals with variable motor abilities.

## 2. RELATED WORK

We discuss the existing approaches to understand and support interactions by people with motor-impairments, and strategies to conduct longitudinal in-the-wild user studies.

### 2.1 Motor-Impairments and Touchscreens

In recent years, there have been a number of user studies investigating interaction by motor-impaired users of touchscreen technologies. These efforts, aimed at improving accessibility, speed, and accuracy of user interactions. Particularly, authors have proposed novel techniques for user input during text entry


*ASSETS '14,* October 20 - 22 2014, Rochester, NY, USA

http://dx.doi.org/10.1145/2661334.2661362

**Table 1 Participant profile; study id, participant age and gender, prior experience with touchscreen devices, specific impairment and current accommodations to deal with symptoms. Deep Brain Stimulation (DBS)**

| *ID* | *Age* | *Gender* | *Touchscreen Exp.* | *Impairment* | *Current Accommodations* |
|---|---|---|---|---|---|
| **P1** | 55 | Female | Self-service machines | Parkinson's Disease, slight hand tremors | Regular medication to suppress symptoms |
| **P2** | 59 | Male | Tried iPod touch before. | Spinal injury, muscle spasms, hand tremors, Sensitive to light | Regular medication to suppress symptoms |
| **P3** | 57 | Male | Self-service machines | Parkinson's disease, hand tremors | Regular medication to suppress symptoms |
| **P4** | 73 | Female | Tried iPod touch before. | Myalgic Encephalomyelitis. Muscle spasms in arms and hands. | Medication to suppress mobility symptoms, not cognitive |
| **P5** | 63 | Male | None | Parkinson's disease, hand tremors | Regular medication to suppress symptoms |
| **P6** | 21 | Female | Tried iPod touch before. | Essential tremor | Medication when symptoms increase |
| **P7** | 65 | Female | Tried iPod touch before. | Parkinson's disease | Medication. During the study underwent DBS surgery |
| **P8** | 75 | Male | Has an iPod Touch | Parkinson's disease, hand tremors | Medication when symptoms increase |
| **P9** | 74 | Female | Tried iPod touch before. | Essential tremor | Regular medication to suppress symptoms |

tasks, and recommendations for screen layouts, target sizes, and interactions styles [3, 10, 14, 13].

Guerreiro et al., [3] measured the performance abilities of tetraplegic people using tasks of common touchscreen interactions. The stimulus application included gestures such as tapping, directional swipes, swipes crossing targets, and swipes exiting the screen. The results showed that the optimal interaction method was tapping with target sizes of at least 12mm. Later Wacharamanotham et al., [14] compared tapping with Swabbing, an alternative input method of target selection for people with tremors, and found that Swabbing was able to reduce target selection error rates.

Nicolau and Jorge [10] investigated text-entry on virtual keyboards by elderly users. They reported a strong correlation between error rates and users' tremors. Furthermore, they demonstrate that applying personal touch offset models to users' inputs could significantly reduce error rates.

While these works have investigated the interaction characteristics and abilities of motor-impaired users, the laboratory study designs meant that the measurements were obtained from a single session. Therefore, it is unknown if the participants performances would remain consistent when measured for longer periods of time, or if their abilities would be subject to high variable and erratic change, as observed by Hurst et al. [6], with mouse interactions.

## 2.2 In-the-Wild Studies

Laboratory-based evaluations allow researchers to control for external factors that can influence participant interaction performance. Typically, these studies tailor situations to remove distraction and interruption, thus ensuring users' attention on the task and relative precision in interaction accuracy. While highly controlled laboratory experiments provide clean measurements with minimal errors, interaction behaviors captured within natural settings differ from those captured within the laboratory [2]. Additionally, laboratory-based evaluations impose time restrictions on user studies. Characteristically lasting no more than an hour at a time, they restrict the potential for capturing the performance changes that naturally occur throughout daily usage as a result of fatigue or situational constraints. During the *Dynamic Keyboard* evaluations participants were asked to provide typing samples at various points throughout the day to begin to understand these changes, their findings revealed that typing performances could vary erratically, gradually or for some users remain constant [11].

Hurst et al. [6] conducted in-the-wild user evaluations to investigate the pointing performance of individuals with motor impairments in natural usage conditions. The initial phase of the evaluation required participants to complete baseline calibrations using the IDA [8] software suite, based on Fitts' Law clicking tasks. Beyond this initial phase, participants were free to login to the system and play games, or use other applications such as word processing. Using application interaction models, the authors were able to infer user intent from the mouse input, allowing measurements of overlapping button clicks, slips, accidental clicks, direction changes and excess distance travelled similar to the type of measurements possible within the controlled laboratory setting [7]. Hurst et al. [6] reported that participant performance was highly variable both between and within sessions, further supporting Trewin's early findings that individuals' performance can fluctuate due to medication, progression of a disease, or as a symptom of impairment [12]. Hurst et al. [6] argue that user evaluations with less control and constraints can help to reduce the risk of fatigue and stress by allowing participants to dictate their own break and interaction schedules.

# 3. USER STUDY

The purpose of this study was to capture in-the-wild touchscreen performance of motor-impaired participants in order to understand how individuals' abilities may vary over time.

## 3.1 Participants

Nine participants with motor-impairments, four male and five female, took part in the four-week in-the-wild user study. They were recruited through local Parkinson's UK support groups. Ages ranged from 21 to 75 (M=60, SD=17) years old. Table 1 provides details of the participants' individual abilities and medical conditions.

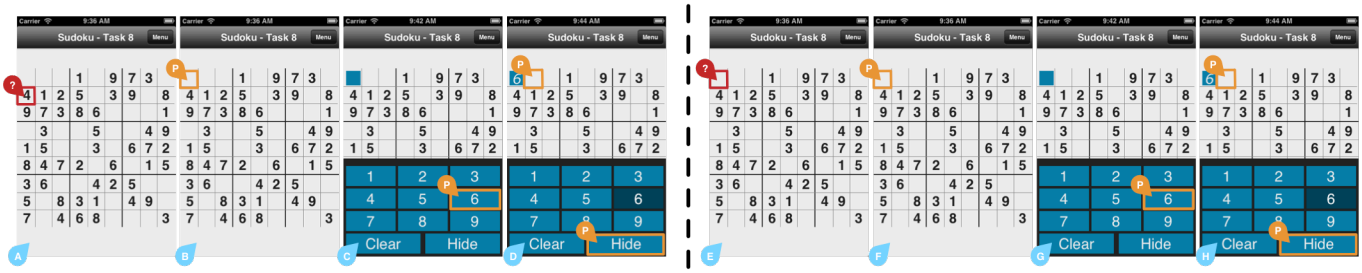


**Figure 1 The Sudoku Game model refining the target intent for a wrong target error (left) and refining the gesture type of an unrecognized gesture error (right).**

## 3.2 Apparatus

The high-level structure of the apparatus mirrors that of an earlier laboratory evaluation [9], whereby the participants were provided with a 4th Generation iPod touch device, preloaded with the stimulus application to be used in-the-wild. The Sudoku stimulus was designed to mirror typical interactions of mainstream touchscreen applications, and was embedded with the SUM framework, a data collection framework to capture the participants' interactions. The mobile device was connected to the university server via the participants' home WiFi connection, allowing the application to transmit the interaction data throughout the four-week in-situ study. The SUM framework enabled the collection of the following touch interaction features:

**Touch Location (X, Y):** represents the horizontal and vertical location of the user's finger when it is lifted from the screen. These locations are absolute values measured in relation to the physical screen dimensions.

**Touch Offset (X, Y):** captures the user's *x* or *y* offset between the touch begin (finger down) and end (lifting the finger off) states.

**Touch Duration:** captures the time duration between the first and final state of a touch gesture.

**Absolute Touch Movement:** measures the total Euclidean distance between all of the touch states of a gesture.

**Straight-line Touch Movement**: measures the Euclidean distance between the first and last touch states of a gesture, the combines touch x and y offset.

**Relative Touch Movement:** calculated as the ratio of *straight-line movement* to *absolute movement* to measure the amount of additional or unintentional movement within the gesture.

**Movement Direction Changes (X, Y):** measures the number of direction changes within the horizontal or vertical axis during the touch movement states.

**Target Offset (X, Y):** captures the user's *x* or *y* offsets from the center of the target interacted with during the touch gesture.

## 3.3 Experimental Application: Sudoku

This study aimed to understand the touchscreen interaction abilities of people with motor impairments; therefore, the application performed no interface adaptations or personalization. We used custom gesture recognizers to record all of the application interactions, they relied on the behaviors of the device's default recognizers to interpret and respond to the participants' inputs. A Sudoku game was selected as the stimulus application due to the appeal of mobile gaming; its logical gameplay strategy required that participants enter particular values for each cell to solve the puzzle; and roughly 40-70 precise tapping interactions could be captured from playing a single game. Furthermore, the design of the Sudoku board meant that the tapping interactions would occur throughout all of the screen locations, giving an understanding of the participants interactions across the entire screen. Participants could interact with Sudoku application in the following modes. *New game,* this mode would ask the player to select a difficult setting depending on their skill level, and then a new Sudoku puzzle would be generated and displayed on screen. *Task game*, from which they could select one of the 14 predefined puzzles.

To interact, players had to tap on an empty cell to select it for editing, then enter the desired number using the onscreen number pad as illustrated in Figure 1. When the correct value was entered for a cell, the player could either select another cell to edit, or tap the *Hide* button to remove the number pad and reveal the entire board again. Alternatively, if a number were incorrectly entered, the selected cell would highlight this error by making the cell background *red.* Players could resolve errors by either entering another number, or by tapping the *Clear* button to remove the cell value. To complete the game the players had to solve the puzzle and enter the correct value into each of the empty board cells.

### *3.3.1 Touch Intent Discrimination*

Within controlled laboratory user studies, it is relatively straightforward to establish a user's intended actions. Typically the design of the study is such that users have a clear goal, thus error identification is easy. For example, their brief would be to tap the onscreen targets as quickly as they can with their dominant hand. The resulting dataset would contain user touch information where the intended gesture and target are known. However, when conducting in-situ user studies it is unreasonable to assume that each user interaction carries intent, or that the device correctly interpreted the user's intentions. Therefore, it is vital to apply methods to discriminate between actions with and without intent. We leverage the logical gameplay strategy of Sudoku to aid this discrimination of the touch data.

**Extract Intent from Sudoku Interactions.** To successfully complete a Sudoku puzzle, the player must correctly position the numbers 1-9 into each empty cell, ensuring that no column, row or 3x3 block contains duplicate numbers. Since there is only one correct number per cell, once an empty cell has been selected we can infer the correct target for that cell. Therefore, we can discriminate between interactions that are accurate and intended, and those that are not, in the following way using our Sudoku *Game Model*.

*Wrong Target.* One possible scenario for intent correction is when the participant taps a target that does not respond to the tap gesture, implying that a nearby target would have been the intended target. The *Game Model* captures such scenarios in the following way, illustrated in Figure 1 (left). The participant taps the cell containing the number '4', which is not an interactive object and therefore no interaction feedback is provided. However, SUM records the tap gesture marking this object as the target. The participant next taps a nearby empty cell target that is

interactive. This was potentially the intended target for the previous interaction, however there is still uncertainty. The next probable moves are either tapping the 'hide' button to remove the number pad, signifying that the user is happy with their number selection. Alternatively, they may tap a new cell and enter the next number. If either of these possible interactions occur then the game model marks the cell (B) as complete and the number '6' as committed. At this point the intended target for the original tap gesture (A) is refined to the empty cell above (B), moreover the other tap gestures are confirmed as intended tap gestures and targets.

*Unrecognized gesture.* Another common interaction error occurs when the participant performs a touch gesture that is unrecognized by the device. Possible reasons for a gesture being unrecognized is due to timing or movement values outside of the acceptable parameters for the tap gesture. The following example details how the Sudoku game modeler handles unrecognized gesture errors, to refine user intent, illustrated in Figure 1 (right).

The participant attempts to perform a tap gesture in the empty cell (E), but the tap duration exceeds the maximum duration parameter of the tap gesture recognizer. No interaction feedback is provided to the participant, but the gesture is captured and recorded by the SUM framework as an unrecognized gesture. Next, the participant repeats the action (F), this time it is recognized by the device and the cell receives the tap gesture. The scenario then plays out as detailed above for the *Wrong Target*. The Sudoku modeler can then refine the gesture type of the original interaction (E) from being an unrecognized gesture to being an intended tap gesture.

*Unrecognized gesture and wrong target*. As the name suggests, this error occurs when the participant performs a gesture that is unrecognized with a target that is not interactive. While it would be possible to use the steps detailed above to attempt to refine and correct the intent for these interactions, it was decided not to infer intent for these interactions due to the compound errors.

### *3.3.2 Validating The Sudoku Game Model*

Participants were asked to complete data copying tasks during the study. Using the *task* game mode, they could complete one of the 14 predefined Sudoku puzzles using the solution sheets provided by the researchers. The predefined puzzles and solution sheets meant that the participants could copy the correct values for each cell without having to solve the puzzle themselves. Thus, any incorrect values entered were the result of an interaction error, such as tapping the wrong target. Leveraging the task sheets, we were able to obtain refined measurements for the participants' intended actions. When comparing the agreement between the intent classifications from the Sudoku Game Model with the copy task data, we found that the results were statistically similar, kappa $\kappa$=.896, $z$=-1.524, $p$=.127.

## 3.4 Procedure

At the beginning of the user study, the participants met for an initial training session and informal discussion with the researchers. This initial session took roughly 30-minutes to introduce the purpose of the study and to provide basic training of the stimulus application. Participants P1, P3, P5 had never used smartphone touchscreen devices before, and were provided further training on the basic device functionality and controls within this session. Once the participants felt confident enough to operate the device and the application on their own, the researcher entered the unique login details for that participant and activated the data collection capabilities. Participants were provided with printed copies of the Sudoku solutions for the task puzzles, and encouraged to complete a number of the Sudoku task puzzles during the four-week study. Finally, the participants were provided with information to assist in connecting the devices to their home Wi-Fi network to ensure that the captured interaction data could be synchronized with the data collection server.

It was vital that the devices were able to regularly communicate with the data collection servers to return interaction logs, which allowed the researchers to verify that the devices were operating as intended. Additionally, participants were asked to keep a brief diary of their experience for the duration of the study. This was aimed at supporting the interpretation of the interactions, in particular providing a better understanding of extreme outliers in user performance. Since the system automatically recorded timestamps for all interactions, participants were only encouraged to take note of the unusual or out of the ordinary behaviors, such as feeling poorly or experiencing extreme symptoms.

# 4. RESULTS

Our goal was to understand how people with variable motor abilities interact with touchscreen devices in-the-wild. Firstly, we summarize the dataset of captured touchscreen interactions, and the extracted intent measurements. Then we describe the interaction performance abilities and relate them to the touch features and the accuracy of the device tap gesture recognizer.

## 4.1 Dataset Summary

Using the Sudoku application embedded with the SUM framework, a dataset containing over 244 interaction sessions, consisting of 23,474 touchscreen gesture interactions (taps, swipes and unrecognized gestures) was collected from the nine participants throughout the four week in-situ user study, show in Table 2. We breakdown and summarize the recorded touch gestures during the user evaluation in Table 3. 17,092 (72.8% of all touchscreen inputs) gestures were assigned user intent and target classifications using the Sudoku game model. These classified instances are used to test the classification accuracy of the device gesture recognizer

**Table 2 Summary of participant gestures captured from the Sudoku application during the in-situ user study**

| *Participant* | *Taps* | *Swipes* | *Unrecognised* |
|---|---|---|---|
| **P1** | 4275 | 43 | 1726 |
| **P2** | 2491 | 97 | 1053 |
| **P3** | 1094 | 22 | 491 |
| **P4** | 2529 | 25 | 86 |
| **P5** | 3467 | 13 | 37 |
| **P6** | 489 | 6 | 26 |
| **P7** | 3547 | 9 | 445 |
| **P8** | 250 | 5 | 15 |
| **P9** | 1026 | 19 | 173 |
| **Total** | **19182** | **239** | **4053** |

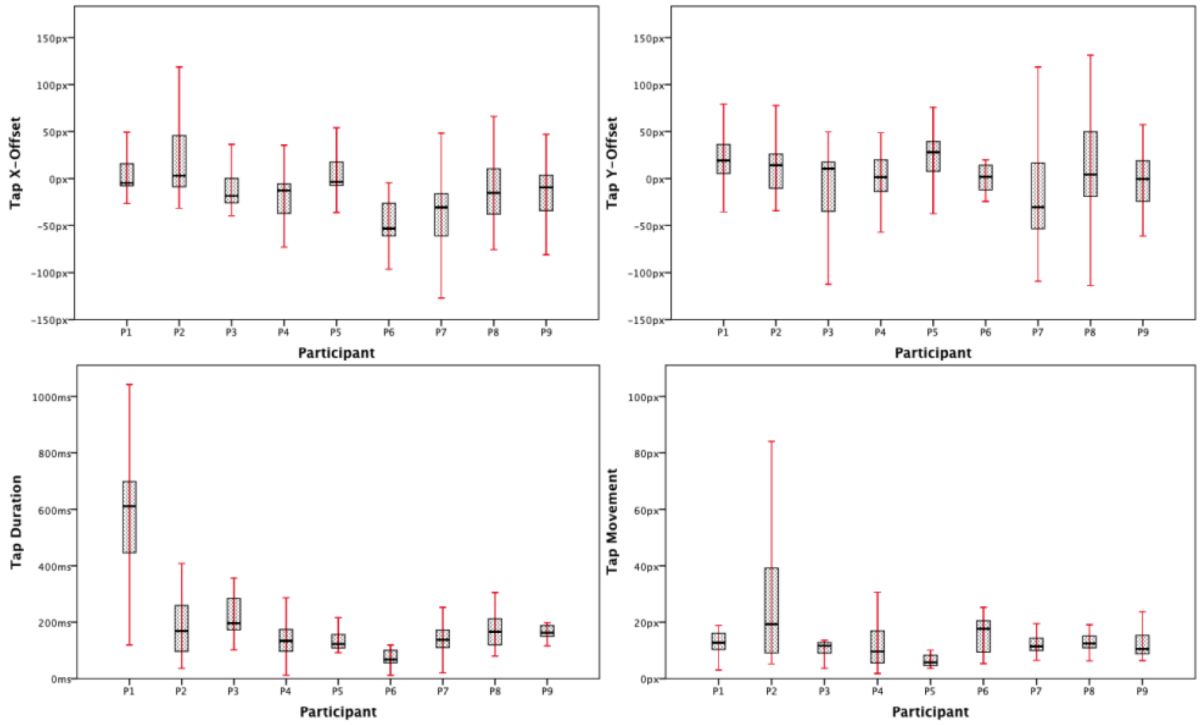


**Figure 2 Boxplots of the overall tap x-offset, y-offset, duration and movement of each participant.**

Table 3 Summarizes the 17,092 touchscreen gestures with intent measurements, showing the number of gestures that were recognized or unrecognized, and whether or not they were associated with the correct target. Overall, the default tap gesture recognizer was able to classify the users' input with 82.9% accuracy. The breakdown shows that 769 (39.9% of all 1928 unrecognized gestures) were in fact intended tap gestures on the correct target, however, the participants were unable to perform the tap gesture in line with the device tap gesture recognizer timing and or movement parameters. Furthermore, 2140 (14.1% of the 15,154 successful tap gestures) were recognized as the wrong target. These types of errors can be caused by involuntary movement when performing the tap gesture, or difficulties with target acquisition.

**Table 3 Breakdown of recognized gestures (from default recognizers) and the resulting intent measurements using the Sudoku Game model**

| | ***Unrecognised*** | ***Recognised*** | |
|---|---|---|---|
| ***Correct*** | 769 | 13014 | **13783** |
| ***Incorrect*** | 1159 | 2140 | **3299** |
| | **1928** | **15154** | |

## 4.2 Touchscreen Performance

We can see from the recognized tap gestures and corresponding intent measurements, that the participants were experiencing errors due to the selection of wrong targets and because of difficulties adhering to the fixed timing and movement constraints of the recognizers. However, it is unclear if these errors could be resolved by simply applying new fixed values for touch offsets, timing and movement parameters specific to this population, or is touchscreen performance dependent on an individual's abilities. To answer this question Kruskal-Wallis tests were run to determine if there were differences in touch interaction characteristics between participants. We found that touch interaction characteristics were statistically significantly different between participants for, touch x-offset $\chi^2(8)=87.393$, $p<.001$; touch y-offset $\chi^2(8)=39.116$, $p<.001$; duration $\chi^2(8)=126.987$, $p<.001$; and movement $\chi^2(8)=91.970$, $p<.001$, shown in Figure 2. Our results suggest that x-offset, y-offset, duration and movement are the factors that were affecting touchscreen performance.

Our primary goal of this study was to identify how touchscreen performance behaves across time; *do performance abilities differ between sessions?* We applied Kruskal-Wallis tests to individuals' interaction characteristics to identify differences between sessions. We found that the these touch interaction characteristics were also statistically significantly different between interaction sessions for participants *P2, P3, P4* and *P7,* ($p<.001$). No significant differences were observed in touch movement of tap gestures for participants *P1, P5* and *P8* between sessions, ($p>.05$). However, the touch x and y-offsets were statistically significantly different between sessions, ($p<.001$). Statistical differences in tap duration and movement only, were observed between sessions for participants *P6 and P9* ($p<.001$).

Figure 3 illustrates the individual participant's daily average x-offset behaviors when performing tap gestures. It is clear from these figures and the aforementioned results that for most participants these interaction characteristics vary dramatically and erratically between sessions, making it unrealistic to predict a users current abilities based on previous sessions alone.

# 5. DISCUSSION

Our goal was to investigate the touchscreen performance abilities of people with motor impairments to understand the barriers to access these technologies. We collected interaction data from a four-week in-the-wild study with nine participants on a set of

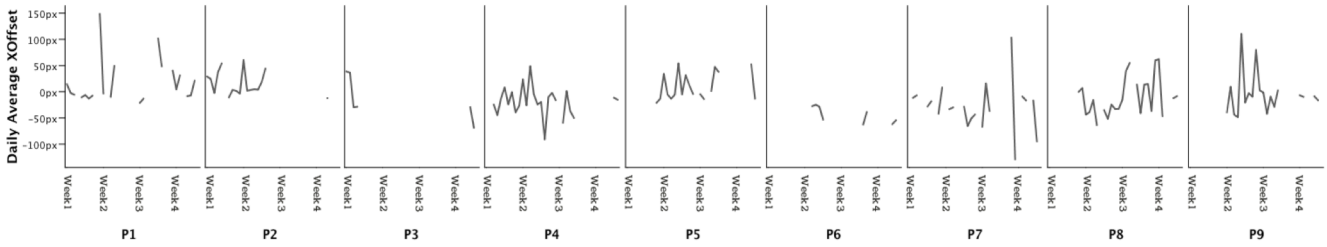


**Figure 3 Line graphs showing the daily average tap x-offset of each participant.**

eight touch features. Although we only collect data from a single application and analyze only the tap gesture recognizer, results show that performance widely varies between users. Moreover, we show that an individual's touch behaviors also vary between sessions, resulting in serious implications for the design of future gesture recognizers. In fact, the device's default tap recognizer was unable to deal with users' varying abilities and misinterpreted almost one in five interactions. Therefore, alternative gesture recognition methods would be required to accommodate these variances in performance, such as personalized touch models.

We can see from the gesture recognizer breakdown (Table 3) that 73.5% of the tap recognizer errors were due to the wrong target being selected, thus related to users' accuracy when tapping. A common strategy to improve accuracy when tapping targets on touchscreens, is to use touch offset models that can correct for the offsets of users' input, by shifting the intended touch *x, y* location a fixed amount. However, this strategy assumes that the user input is consistently offset thereafter; otherwise, this touch offset would need to be continually recalculated.

# 6. SESSION SPECIFIC MODELS

To mitigate the between session differences of user performance, we propose a novel approach leveraging measurements of the initial session interactions to predict current performance abilities of the user and adapt the tap gesture recognizer to match those abilities, by using probabilistic distribution functions to parameterize the tap gesture recognizer success criteria.

## 6.1 Probabilistic Gesture Recognition

Currently within mobile touchscreen interactions, the method of classifying *tap* gestures is the use of the x, y location (either touch begin, or touch end); and fixed movement threshold (movement between the touch begin and end states). However, we opted to use tap gesture recognizers that are not defined by fixed parameter boundaries. Instead, our tap recognizers use statistical probability to account for the variations in gesture performance between instances. The tap recognizers used in this evaluation applied Gaussian functions to define the attributes for gesture classification. Gaussian functions allow the tap recognizers to perform classifications based on probability of an action given a series of parameters, as opposed to relying on definitive parameters. For example, Gaussian functions are capable of resolving common touch offset errors whereby the touch occludes two or more possible targets. The target with the highest probability is suggested as the intended target. Similarly, they can account for variances within user performance, such as timing, rather than using a fixed maximum value to threshold all touches above this. The Gaussian function would simply return a lower probability. If the probability of the touch being intended were greater than the probability of it being unintended, then the tap would be recognized. However, the traditional fixed threshold model would not be recognized if the touch were even 1ms over the threshold tap duration.

We defined the parameters of the tap gesture recognizer using example data of successful, intended tap gestures from our participants to obtain the mean (μ) and standard deviation ($\sigma$) values required by the probability density functions. The features included were: *x, y* location, *duration* and *movement*. In addition, the models also parameterized the *x, y* offset. Typically the x, y offset is handled by touch offset models defined on a per-device nature, shifting the user's touch input location by a fixed Euclidean vector. Previous studies have proposed user specific touch offsets models, reporting significant improvements in the precision of touch input [1, 4, 5, 15]. However, these works were not evaluated with motor-impaired users, spanning several sessions and long periods. Our results demonstrated that user performance varies significantly between sessions for people with motor-impairments. Therefore, we propose applying touch models that are specific to the abilities of the user on a per session basis.

## 6.2 Classifying Session Performances

Session features, based on the touch features of the gesture recognizers, were used to measure the variances of user interaction across sessions and cluster the individual sessions based on performance abilities. To ensure the sessions were clustered independent of any stereotypical groups, the features were selected based on the low-level touch interactions. Each session feature captures the mean and standard deviation of the corresponding touch feature, and are grouped by the gesture type they represent, e.g. *Tap* or *Swipe*.

The session features were then normalized using the standard score formula to aid the distance measurement process. The similarity of two sessions is based on the Euclidean distances of each normalized session feature, as shown in Equation 1.

**Equation 1 Session distance formula; where $F^{X}_{i}$, represents the feature value of the current session, $F^{S}_{i}$ is the feature value of the comparison session.**

$$d = \sum_{i=0}^{n} \left| F_i^x \quad F_i^s \right|$$

## 6.3 Evaluation of Touch Models

The purpose of this evaluation is to simulate the effects of applying our touch models and tap gesture recognizers on the Sudoku gameplay. We used the interaction data collected within the in-situ user study for both the training and testing data, it is possible to simulate the behavior of the tap gesture recognizer and measure the classification accuracy against the extracted user intention values. Our simulations explored the effect of the model subject, by using data from an *individual* user vs. a *group* model,

built using data from all other participants excluding the current user. Furthermore, we explored the effect of *user specific* vs. *session specific models*; these are user specific models built each session to accommodate for the user's current performance abilities.

### 6.3.1 Training and Testing Data

*Testing Data:* each simulation required 200 tap gesture instances with intent measurements. These tap gestures were sourced from the user's touchscreen interactions within the Sudoku application. Touch gestures were selected randomly from any of the user's Sudoku sessions whereby the gestures had an associated intent measurement.

*Training Data:* depending on the selection method of the user model condition, training data was defined as 300 touch gesture instances. For the *user specific* models, data was randomly selected from all available sessions. However, for *session specific* models, we defined a subset of sessions with similar interaction behaviors to the users current abilities, based of their session distances, from which the training data were selected. In both methods, the available sessions were subset by the subject condition of the model, i.e. *individual* contained only data from the current user. Furthermore, to ensure that the training data used to build the touch models was also not being used to evaluate the model's accuracy, the 200 testing instances were selected first, and excluded from the available dataset of training data.

### 6.3.2 Validation Method

Baseline performance scores were obtained for the device default configuration by measuring the number of recognized user interactions that match the previously extracted touch intent values. Each model was then scored against these baseline measurements, values greater than zero determined that user models correctly recognized more instances of user intent. In order to reduce the variability of the user model performance measurements, 30-fold cross-validation was applied to each model evaluation. Sessions were excluded if fewer than 10 touchscreen interactions were captured. Likewise, any model dataset that did not meet the required number of training ($n$=300) and testing ($n$=200) instances was excluded from the evaluation.

### 6.3.3 User Specific vs. Session Specific

A Kruskal-Wallis test was used to determine whether there were differences in the accuracy of the gesture recognizers between the *baseline* (*Mdn* = 85%), *user specific* (*Mdn* =79.7%), and *session specific* (*Mdn* = 95.1%) touch model conditions. Tap gesture recognizer accuracy showed a statistically significant difference between the touch models, $\chi^2(2) = 18.763$, $p < .001$. Pairwise comparisons were performed with a Bonferroni correction ($p$<.0167) for multiple comparisons and Post-hoc analysis revealed statistically significant differences in tap gesture recognizer accuracy between the *session specific* and *baseline* ($p$=.006), and *session specific* and *user specific* ($p$<.001) touch model conditions, but not between the *baseline* and *user specific* ($p$ = .118).These results suggest that the *session specific* models have an effect on the performance of the tap gesture recognizers. Specifically, these results demonstrate that *session specific* user models can improve the touch recognition accuracy of touchscreen devices for individuals with motor-impairments, as illustrated in Figure 4.

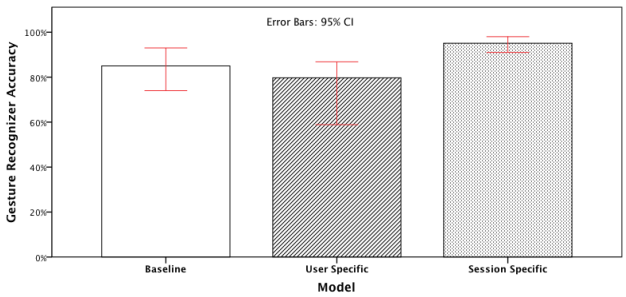


**Figure 4 Classification accuracy of gesture recognizers for touch model conditions**

### 6.3.4 Group vs. Individual Models

Mann-Whitney U tests were run to determine whether there were differences in tap gesture recognizer classification accuracy between the *Group* and *Individual* subject conditions. The median classification accuracy were not statistically significant between the stereotypical and individual user models for the *user specific*, $U$=48, $z$=1.156, $p$=.211 or *session specific*, $U$=19, $z$=-1.640, $p$=.101 models, illustrated in Figure 5. These results suggest that the subject of the dataset does not affect the accuracy of our touch models, therefore permitting the creation of touch models from the interactions of other users. However, we have found that when applying the session similarity measurements it is actually more beneficial to share data between users, with the results increasing from the *session specific individual* (*Mdn*=93.6%) to *group* (*Mdn*=97%) models. In contrast, the *user specific group* (*Mdn*=59%) decrease in accuracy from the *individual* (*Mdn*=82.6%). This shows that contemporary *user specific group* models cannot outperform *individual* models. However, by leveraging the similarity of sessions between users the *session specific* models can take advantage of larger interaction datasets to locate data that closely matches the user's current behaviors and abilities to define tap gesture recognizers that can improve the recognition accuracy of their interactions.

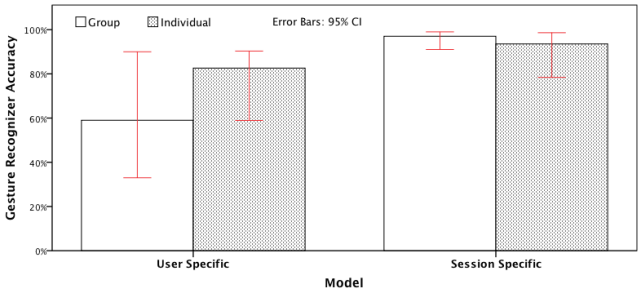


**Figure 5 Classification accuracy of tap gesture recognizers for touch models by subject condition**

Our results demonstrate that touch gesture recognizers using personalized user models can outperform the classification accuracy of the device default gesture recognizers. However, the *user specific* models relied on random selection of training data producing highly variable results that would not consistently improve touch recognition accuracy. Using the *session specific* models to create personalized tap gesture recognizers resulted in a more consistent performance, with significantly better recognition accuracy. Furthermore, the subject of the training data had no significant effect on the recognizer accuracy. This result was true for both *user specific* and *session specific* models. Moreover, the *session specific* models achieved higher levels of accuracy with the group condition than with the individual's own data.

# 7. CONCLUSIONS

We conducted a four-week in-the-wild investigation into the mobile touchscreen interaction performances of nine users with motor-impairments. We believe this study to be the first of its kind. Measurements of participants' touchscreen abilities were captured from the 23,474 touchscreen interactions made within our Sudoku stimulus application. Leveraging the logical strategy used to solve the Sudoku puzzles allowed us to obtained refined intent classifications for the captured interactions, thus removing the need for participant to complete semantically meaningless calibration tasks. Analysis of this dataset revealed that 39.9% of the unrecognized gestures were intentional taps on the correct targets, which were misclassified by device's default gesture recognizer due to the interactions exceeding timing and movement parameters. Similarly, 14.1% of recognized taps were associated with the wrong target, leading to the participants experiencing unintended actions.

Our results showed that touchscreen interaction characteristics varied significantly not only between participants, but also for the participants' own sessions. Based on these findings, we introduced and evaluated *session specific* gesture recognizers that accommodate for the variances of individuals' touchscreen performances by leveraging measurements from their current abilities. Participants' sessions were clustered using their interaction characteristics, allowing new models to be constructed from session data that closely matches the individuals' performance abilities at that point in time. Applying *session specific* gesture recognizers, we were able to achieve 95.1% recognition accuracy, significantly outperforming the device default recognizer. Finally, we investigated the effect of producing models from both the group and the individual's data only. While *session specific* models using an individual's data provided 93.6% accuracy, the models trained using data from the other participants was able to achieve 97% accuracy. These results have demonstrated that by sharing interaction data between users and accounting for their variable abilities, we can improve touchscreen performance for individuals with motor-impairments.

# 8. ACKNOWLEDGMENTS

Support for this project was provided by RCUK Digital Economy Research Hub EP/G066019/1 – SIDE: Social Inclusion through the Digital Economy. We thank our participants who provided so many insights and Marianne Dee who helped us locate participants for the research.

# 9. REFERENCES

[1] Buschek, D., Rogers, S. and Murray-Smith, R. 2013. User-specific touch models in a cross-device context. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services* (MobileHCI '13). ACM, New York, NY,USA,382-391

[2] Chapuis, O., Blanch, R. and Beaudouin-Lafon, M. (2007). Fitts' law in the wild: A field study of aimed movements. LRI Technical Repport Number 1480 (December 2007). Orsay, France: Universite de Paris Sud.

[3] Guerreiro, T., Nicolau, H., Jorge, J. and Gonçalves, D. 2010. Towards accessible touch interfaces. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility* (ASSETS '10). ACM, New York, NY, USA, 19-26.

[4] Henze, N., Rukzio, E. and Boll, S. 2012. Observational and experimental investigation of typing behaviour using virtual keyboards for mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12). ACM, New York, NY, USA, 2659-2668.

[5] Holz, C. and Baudisch, P. 2011. Understanding touch. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). ACM, New York, NY, USA, 2501-2510.

[6] Hurst, A., Mankoff, J. and Hudson, S.E. 2008. Understanding pointing problems in real world computing environments. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility* (Assets '08).ACM, New York, NY,USA,43-50.

[7] Hurst, A., Trewin, S., Hudson, S.E. and Mankoff, J. 2008. Automatically detecting pointing performance. In *Proceedings of the 13th international conference on Intelligent user interfaces* (IUI '08). ACM, New York, NY, USA, 11-19.

[8] Koester, H.H., LoPresti, E. and Simpson, R.C. 2005. Toward Goldilocks' pointing device: determining a "just right" gain setting for users with physical impairments. In *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility* (Assets '05). ACM, New York, NY, USA, 84-89.

[9] Montague, K., Hanson, V.L. and Cobley, A. 2012. Designing for individuals: usable touch-screen interaction through shared user models. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility* (ASSETS '12). ACM, New York, NY, USA, 151-158.

[10] Nicolau, H. and Jorge, J. 2012. Elderly text-entry performance on touchscreens. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility* (ASSETS '12). ACM, New York, NY, USA, 127-134.

[11] Trewin, S. 2003. Automating accessibility: the dynamic keyboard. In *Proceedings of the 6th international ACM SIGACCESS conference on Computers and accessibility* (Assets '04). ACM, New York, NY, USA, 71-78.

[12] Trewin, S., Keates, S. and Moffatt, K. 2006. Developing steady clicks:: a method of cursor assistance for people with motor impairments. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility* (Assets '06). ACM, New York, NY, USA, 26-33.

[13] Trewin, S., Swart, C. and Pettick, D. 2013. Physical accessibility of touchscreen smartphones. *the 15th International ACM SIGACCESS Conference*.(2013), 19–8.

[14] Wacharamanotham, C., Hurtmanns, J., Mertens, A., Kronenbuerger, M., Schlick, C. and Borchers, J. 2011. Evaluating swabbing: a touchscreen input method for elderly users with tremor. (New York, NY, USA, May 2011), 623.

[15] Weir, D., Rogers, S., Murray-Smith, R. and Löchtefeld, M. 2012. A user-specific machine learning approach for improving touch accuracy on mobile devices. In *Proceedings of the 25th annual ACM symposium on User interface software and technology* (UIST '12). ACM, New York, NY, USA, 465-476.