Keystroke Inference using Smartphone Kinematics

Oliver Buckley[1], Duncan Hodges[1], Melissa Hadgkiss[2] and Sarah Morris[2]

[1] Centre for Electronic Warfare, Information and Cyber,
Cranfield University, Defence Academy of the United Kingdom,
Shrivenham, Swindon, SN6 8LA, UK
o.buckley@cranfield.ac.uk
d.hodges@cranfield.ac.uk
[2] Cranfield Forensic Institute,
Cranfield University
m.hadgkiss@cranfield.ac.uk
s.l.morris@cranfield.ac.uk

Abstract. The use of smartphones is becoming ubiquitous in modern society, these very personal devices store large amounts of personal information and we use these devices to access everything from our bank to our social networks, we communicate using these devices in both open one-to-many communications and in more closed, private one-to-one communications. In this paper we have created a method to infer what is typed on a device purely from how the device moves in the user's hand. With very small amounts of training data (less than the size of a tweet) we are able to predict the text typed on a device with accuracies of up to 90%. We found no effect on this accuracy from how fast users type, how comfortable they are using smartphone keyboards or how the device was held in the hand. It is trivial to create an application that can access the motion data of a phone whilst a user is engaged in other applications, the accessing of motion data does not require any permission to be granted by the user and hence represents a tangible threat to smartphone users.

1 Introduction

Smartphones are becoming an increasingly significant part of our everyday lives as their popularity continues to grow. Research conducted by the Office of Communications (Ofcom) [1] shows that two thirds of all adults in the UK own a smartphone, compared to only 39% in 2012. The research also highlights how essential smartphones are becoming in our everyday lives as they are now considered to be the most important device for connecting to the Internet, ahead of a laptop. However, this trend of increased smartphone ownership is not limited to the UK as research by the Pew Research Centre [2] shows the global median for smartphone ownership is at 43%.

The increasing popularity of smartphones means that they are now used to manage aspects of our daily lives. A survey conducted by the Pew Research

Centre [3] shows that smartphones are being used for a wide variety of sensitive tasks ranging including online banking, education, social interactions, obtaining information about medical conditions, submitting a job application and using key government services.

In this paper we hypothesise that the motion sensors, such as the accelerometer and gyroscope, within a smartphone can be used to infer keystrokes. We posit that it will be possible to infer the keystrokes on a virtual smartphone keyboard based on the movement of the phone, as recorded by the accelerometer and gyroscope.

The remainder of this paper is structured as follows: Section 2 provides a review of the related work, focusing on keystroke and swipe analysis in smartphones and the use of motion sensors in user identification. Section 3 details the methodology used to conduct the experimentation. Section 4 provides an analysis of the collected data and the results of the study. Finally, in Section 5 we conclude by providing a reflection on our analysis and a discussion of further work in this area.

2 Background

Modern smartphones will typically contain a variety of motion sensors, including a gyroscope that is capable of tracking the rotation of the device and an accelerometer to monitor the movement and orientation of the phone in space. These sensors can be exploited to determine certain information about the user of the phone. For example this includes: recognising the activities that are being performed by the user [4] or identifying an individual based on analysis of their gait [5]. One of the interesting benefits of using these sensors is that they can be run as a background process without the need for explicit approval; therefore it can be possible to covertly capture smartphone motion data without the express permission of the user. In essence, it is entirely possible for a malicious application on a mobile device to be able to freely gather motion data whilst another application is active without first requesting permission from the user. In turn the captured motion data can be used to probabilistically infer the users keystrokes in other applications without their knowledge.

The sensors in smartphones have been used to good effect to infer a wide range of information about an individual solely based on the way that they interact with the smartphone's touchscreen. For example, Bevan et al. [6] used swiping gestures to infer the length of the individual's thumb. The length of the thumb can then be used to infer other physical characteristics such as height. Similarly, Miguel-Hurtado et al. [7] analysed the swiping gestures of users to predict the sex of the individual.

Motion sensors within smartphones have previously been used to attempt to infer a user's keystrokes with promising results. Cai and Chen [8] developed TouchLogger, a smartphone application designed to infer the keystrokes on a soft (or virtual) keyboard based solely on the vibrations recorded by the smartphone's motion sensors. The research was capable of successfully inferring more than 70%

of the keys that were typed using only the device's accelerometer. However, the work focused specifically on inferring the keystrokes from a soft keyboard that contained only numbers. The work we present in this paper will look to infer the keystrokes of an individual that use a standard soft keyboard, which contains both numbers and letters.

Owusu et al. [9] extend the work of Cai and Chen to use a smartphone's accelerometer to infer the characters, both letters and numbers, contained within a user's password, although with a relatively small set of only four participants. The work was capable of extracting the 6 character passwords in as few as 4.5 attempts (median). The work of Owusu et al. focused only on the use of accelerometer readings, in contrast to our own work with also includes analysis of rotational data using the smartphone's gyroscope. When a device is being used by an individual it tends to be held in the hand either unsupported or with the wrists resting on a surface, if the device is being held in two hands with the thumbs for typing the device tends to be held loosely and tilted in the palms in order that the relevant keys are closer to the thumb. If a device is held in one hand the same phenomena occurs however the aim tends to be to reduce the amount the 'pecking' digit has to move. Whilst these movements are relatively subtle they are observable both by the human eye and even more so by the smartphone sensor.

3 Method and experiment

This paper focuses on inferring the keystrokes of individuals as they interact with the virtual keyboard on a smartphone. A data collection framework was created as an Android application, as shown in Figure 1. The application required participants to type a standard paragraph of text twice and then type a different, and dynamically generated, paragraph of text. The participant was asked to type the text using the standard Android on-screen (or 'soft') keyboard, it is worth noting that the auto-complete or predictive text function was disabled. During this activity the motion of the device was recorded using the rotation, gyroscope and acceleration sensors, the times of the key presses were also record. The standard text that participants were required to type contained 132 characters (less than the length of a tweet) and is shown below:

fly me to the moon and let me play among the stars our freedom of speech is freedom or death we have got to fight the powers that be.

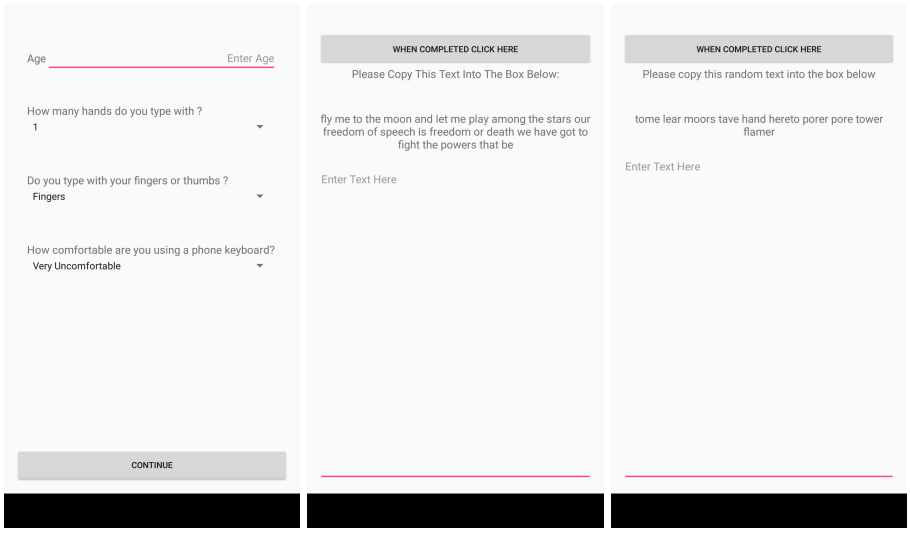
This text was entered twice by all participants of the study, then the final typing activity required the participants to type a paragraph of text that had been dynamically generated. To generate this dynamic text the fixed text, shown above, was segmented into strings of two characters called bigrams. For example, the word *hello* would contain the following bigrams: *he, el, ll, lo*. The dynamically generated text that participants were required to type for the final activity was then generated by searching the Wordnet corpus [10] for words that contained these bigrams. This approach allowed all participants to enter the same set

of training data and then the approach would be validated against the third, dynamically generated set of text.

Additionally, the data collection application also asked participants for a number of biographic questions including:

- Age
- Number of hands used to type
- Whether they type with fingers or thumbs
- How comfortable they were with using a smartphone keyboard — this was ranked as 'Very Uncomfortable', 'Uncomfortable', 'Comfortable' or 'Very Comfortable'.

The study collected data from 25 participants, and all of the participants used the same mobile device (a Nexus 5X) in portrait mode. The use of the same device reduces the risk of any anomalous results based on differences in motion sensors across different devices and indeed across different platforms, for a further exploration of this see the future work.



(a) Metadata Entry (b) Fixed text entry. (c) 'Random' text entry.

Fig. 1: Android application used in this study

4 Analysis and results

The experiment reported in this paper explored how 25 different individuals type on smartphone keyboards, the 25 participants were recruited from Cranfield University staff and students and include a mix of age and gender. The distribution

of the age is shown in Figure 2a, as can be seen the majority of participants are in their 30s and, from the distribution shown in Figure 2b, consider themselves comfortable with using a smartphone keyboard. Whilst debriefing participants following the experiment it became apparent that a number of participants found that they were in fact less comfortable in using a smartphone keyboard without predictive text. In future work, a supplementary post experiment assessment of the participants comfort would be valuable.

The final factor gathered describing the typing was the method of typing, whether the participant used one hand to hold the device and typed with one finger (or one finger and the thumb of the hand holding the device) or the participant used two hands to hold the device and typed with thumbs. There was an even distribution between these two typing methods amongst the participants as shown in the distribution in Figure 2c.

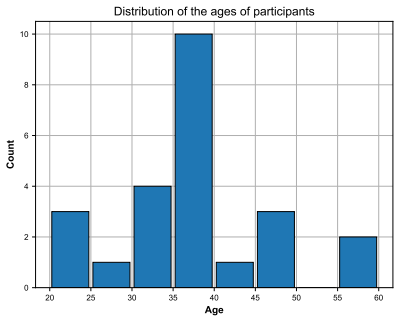
The first task in predicting keystrokes from smartphone motion sensors is the detection of keypresses, in order to identify these presses it is intuitive to consider the acceleration sensor. The act of applying a small force to the device to register a press on a solid surface of the screen causes a small acceleration on the phone, in accordance with the simple laws of motion.

An example trace from the acceleration sensors on the device is shown in Figure 3, where the vertical lines represent keypresses recorded from the keylogger. There are acceleration events caused by initially selecting the box to bring up the keyboard and start typing and other events caused by pressing the button to continue the study. The acceleration traces shown in Figure 3 are broken down to three orthogonal vectors[3]. It is clear that the greatest acceleration is 'into' and 'away from' the phone, this intuitively maps to the pressing down of the soft-keyboard displayed on the screen. The same graph can also be extracted using the magnitude of these three acceleration vectors and this is shown in Figure 4, it is clear from this that the measurements from the smartphone's accelerometer are well correlated with key presses and this correlation for the four different measurements (X, Y, Z and the magnitude of the vector) is shown in Figure 5.

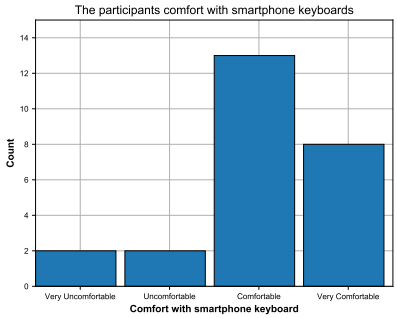
The correlation plots shown in Figure 5 clearly demonstrate that the use of the acceleration in Z or the magnitude of the acceleration vector can be used to extract the keypress times, in our experiment we found that the acceleration in Z produced marginally better results when using a simple threshold.

In order to be able to predict keystrokes we must first build a model for how each user types in this experiment we were primarily interested in the rotation of the device we first consider the initial fixed text typing, it should be noted that this is less than the size of a tweet and represents a relatively singular event (i.e. the text is only written once). We correlate the keystrokes with the acceleration vector in the Z direction in order to identify the optimal threshold for the accelerometer identification of the keystrokes. Once this is performed we now are interested in the rotation of the device between keypresses — this rotation encodes the movement of the device from one keypress to another and

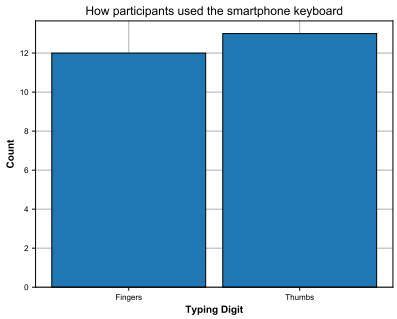
[3] with the phone facing towards the participant X represents left-to-right, Y represents down-to-up and Z represents from behind the phone to the face of the phone



(a) Distribution of participant age.



(b) Distribution of self-described comfort with a soft-keyboard.



(c) Distribution of digit used to type.

Fig. 2: Distributions of participant information

hence encoding the bigram that was typed in the rotation vectors measured by the device.

These rotation vectors are extracted between the keypresses and normalised to a set sample length (in our experiment we chose 1,000 samples), this attempts

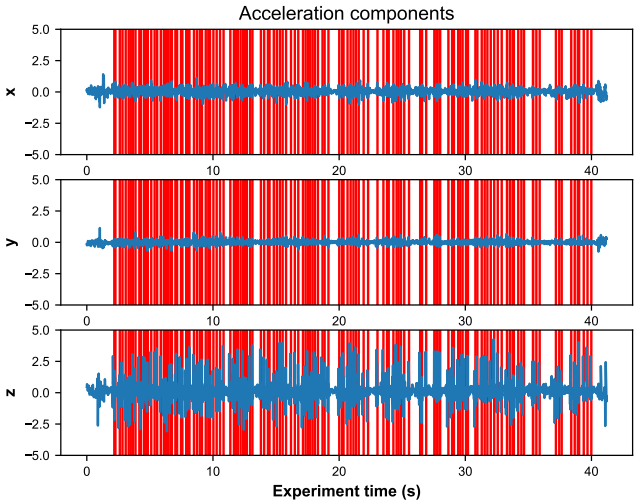


Fig. 3: Individual acceleration vectors during the experiment.

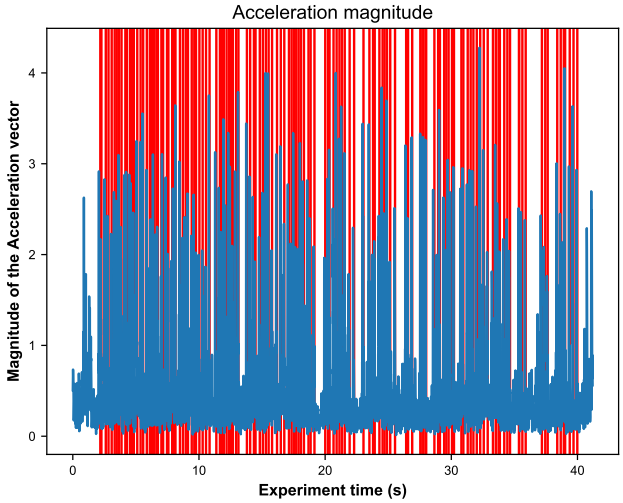


Fig. 4: Acceleration vector magnitude.

to remove the effect of an individual not consistently typing a the same pace. These rotation vectors were then further normalised by removing the average from each vector to form the model for each bigram. This normalisation to a mean of zero attempts to reduce the system memory effects from the previ-

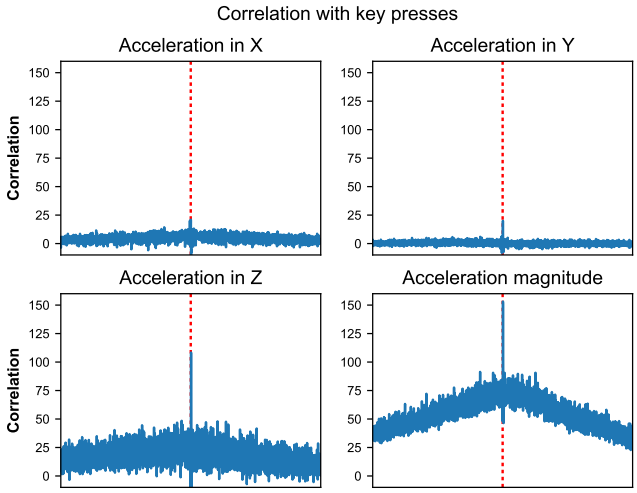


Fig. 5: The correlation of the various accelerometer readings with keypresses as measured by the keylogger. Zero lag is shown by the dotted line.

ous bigrams, as participants do not 'reset' the device to a set position between bigrams.

In this experiment a model was created for each participant, Figure 6 shows the models created for the bigram `FL` for three different participants, one who holds in the right hand and types with the left finger, one who holds in the left hand and types with the right finger and a final participant who holds the phone in both hands and types with their thumbs. It is not surprising that the two 'single-handed' participants result in similar yet inverse models although the extent of the rotation is smaller between two participants. The 'two-handed' participant has a very different trace indicating the centre of rotation closer to the centre of the phone with a more complex rotational vector.

These models were constructed from the fixed text on the first page and then validated against the fixed text in the second page, the model was then used in the final experiment with unseen text. Again it is worth reiterating that the training phase for this experiment was a very short piece of text less than the size of a tweet. The prediction was generated in two different ways using this model:

1. **Naive model:** The acceleration vector was used to identify the key press times, the rotation vectors between these presses was then extracted and the bigram with the lowest RMS error when mapped to this rotation was selected as the proposed bigram.
2. **Bigram model:** This approach was built on the output from the naive model but included the fact that any bigram must start with the end let-

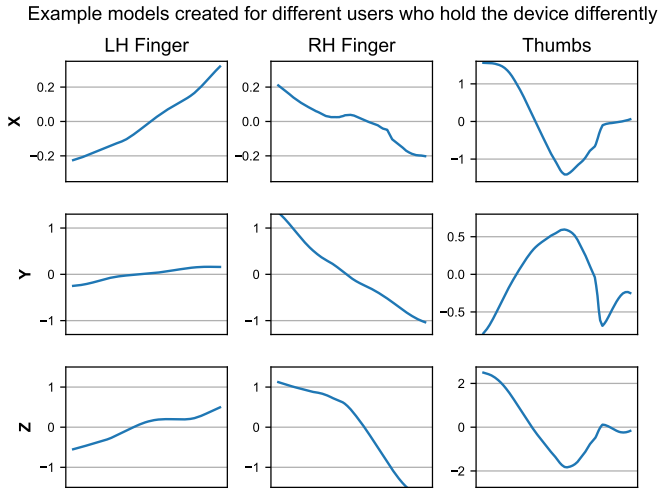


Fig. 6: Example model across three different individuals for the bigram `FL`

ter of the proceeding bigram. In this way the sentence was extracted that minimised the total error whilst maintaining this logical assumption. This made no assumption about the language used, that for example 'er' is a very probable bigram in the English language or that a particular collection of bigrams actually forms a known word, this is covered in the future work section.

The accuracy of these predictions for the 25 participants is shown in Figure 7, the accuracy was the measure of the number of bigrams which were correctly identified normalised to the total number of bigrams in the text, an example of one prediction (achieving 83% accuracy) is:

fly t ato ghe moor ang let me play amongowee stars pof freedom of speech isbfreedom por death we have got go fight the powersbed at be, the underlined bigraphs represent errors

The average accuracies of the naive and bigram model in predicting the training text was 46.9% and 64.7% respectively and the average accuracies on the unseen text was 9% and 16.7%. Whilst these may be considered low, bear in mind that the training process is very short and even with this short training two participants achieve close to 90% accuracy on the repeated texts and close to 50% on the previously unseen text.

Of interest is whether how the participants used the phone has any effect on the accuracy, in order to explore this question we took the accuracy of the bigram model at predicting the fixed text and compared this to the typing method. This is comparison is shown in Figure 8, as can be seen there is little difference in the distribution indeed a two-sided Kolmogorov-Smirnoff test resulted in a

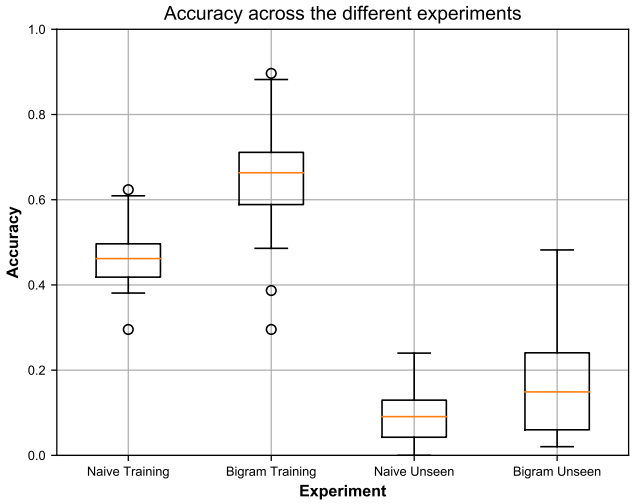


Fig. 7: Accuracy across the experiments.

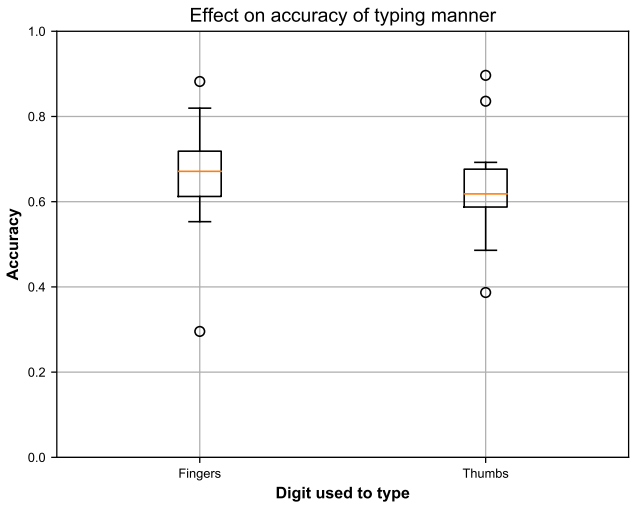


Fig. 8: Effect of typing manner.

Kolmogorov-Smirnoff statistic of 0.269 (p-value of 0.683) indicating that, from this sample, the typing method has no effect on the accuracy.

We can also explore the effect that comfort with a smartphone keyboard has with the accuracy, a boxplot of the participants self-assessed 'comfort' is shown

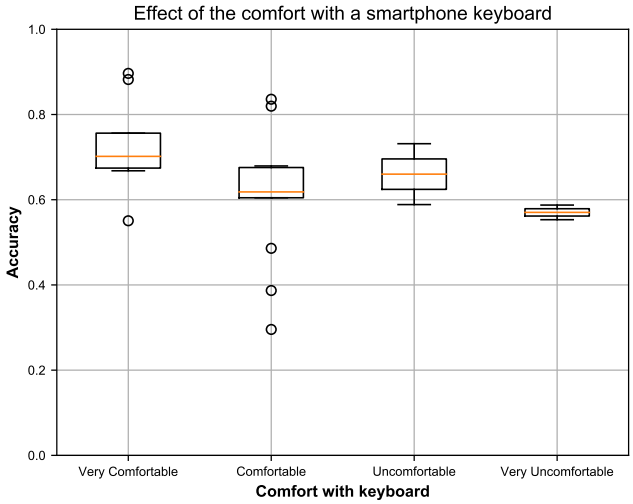


Fig. 9: Effect of typing comfort.

in Figure 9. From this experiment, there is a slight decline in performance, but not a statistically significant one, a Pearsons correlation resulted in a correlation of 0.316 and a p-value of 0.124, this is undoubtedly affected by the relatively few participants who considered themselves 'uncomfortable' or 'very uncomfortable' with the smartphone keyboard.

Since the self-assessment of 'comfort' with a soft-keyboard is a qualitative self-assessment, and as discussed previously a number of participants considered the lack of predictive systems reduced their comfort levels we considered the accuracy as a function of a tangible observable typing characteristic. The most illustrative characteristic in our model is that of flight-time and dwell time, this represents the time from a key-up from one key press to the key-up of the next (so includes the time taken to move from one key in addition to the time for which the key is depressed).

In order to remove the effects of long pauses between presses and purely to focus on the measure of typing speed we extracted the median measure of this characteristic across the two fixed text entries, these are then plotted against the accuracy and this is shown in Figure 10. It is clear from this that for most participants the second attempt was faster than the first attempt and from the linear regression shown in Figure 10 demonstrates no relationship between typing speed and accuracy.

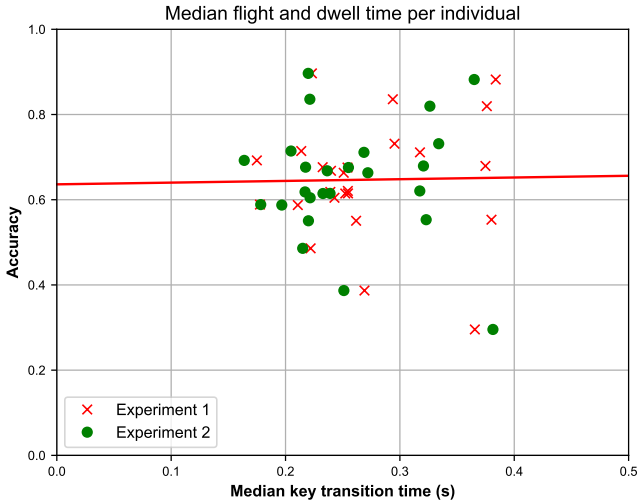


Fig. 10: Effect of typing and timing.

5 Conclusion and Future Work

In this experiment we have deliberately constrained the experiment to use a single device, a Nexus 5X, in portrait mode. Future work will look to explore different sized devices including different sized phones, 'phablets' and tablets, the different physical size and shapes of devices which we would expect to create different motion patterns. Through pushing the app to the appstore we will also have the opportunity to gather greater numbers of participants across multiple languages and devices. This study focused on the ability to train models for prediction from very small amounts of training data, indeed being able to train this device on a timestamped and publicly observable piece of text is a massive opportunity for exploitation. In order to explore the effect of creating a model with larger ngrams than two characters would require larger training sets, of interest in this study would be the degree to which ngrams of three or more characters could be used to generate specific training data.

In order to improve on the bigram model it is possible to leverage the language used on the device to predict the text that has been entered. The language that the device has been configured to can be requested by an application without explicit permission from the user, this would work in combination with the bigram model to predict not only the sentence with the lowest total error but that maximises the number of valid words in the particular language.

The final piece of future work is to generate generic models for the prediction, from manual observation of the models it is clear that there are similarities between the models produced by individuals who type in similar manners. The ability to create generic models would further reduce the amount of training

required and provide an ideal start to a Bayesian approach to predicting keypresses.

We have demonstrated that it is possible to infer the bigrams that are typed on soft-keyboards purely from the rotation of the device, since there is no requirement to ask the user for permission to access the motion sensors of a device this is a covert opportunity for the collection of what is being typed on a smartphone. In our study we trained the models on a small piece of text, shorter than a tweet, even with this limited training data we were able to achieve average performances of 64.7% on text that had been seen before. We have shown that the method that an individual uses to type has no effect on the accuracy of the approach, and whilst how comfortable an individual is using the soft-keyboard does have a small effect it is not statistically significant. In future we look to explore new ways to create the model and inform the predictions to further improve these prediction levels.

References

1. Ofcom. The UK is now a smartphone society. https://www.ofcom.org.uk/about-ofcom/latest/media/media-releases/2015/cmr-uk-2015, 2015. [Online; accessed 26-January-2017].
2. J. Poushter. Smartphone ownership and internet usage continues to climb in emerging economies. http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/, 2016. [Online; accessed 26-January-2017.
3. Pew Research Centre. Smartphone use in 2015. http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/, 2015. [Online; accessed 26-January-2017.
4. J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.
5. T. Iso and K. Yamazaki. Gait analyzer based on a cell phone with a single three-axis accelerometer. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pages 141–144. ACM, 2006.
6. C. Bevan and D. S. Fraser. Different strokes for different folks? revealing the physical characteristics of smartphone users from their swipe gestures. *International Journal of Human-Computer Studies*, 88:51–61, 2016.
7. O. Miguel-Hurtado, S. V. Stevenage, C. Bevan, and R. Guest. Predicting sex as a soft-biometrics from device interaction swipe gestures. *Pattern Recognition Letters*, 79:44–51, 2016.
8. L. Cai and H. Chen. Touchlogger: Inferring keystrokes on touch screen from smartphone motion. *HotSec*, 11:9–9, 2011.
9. E. Owusu, J. Han, S. Das, A. Perrig, and J. Zhang. Accessory: password inference using accelerometers on smartphones. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, page 9. ACM, 2012.
10. G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.