TYPE AND LEAK YOUR ETHNICITY ON SMARTPHONES

Hadiyattullahi Tanko Aliyu, and Yogachandran Rahulamathavan

Institute for Digital Technologies
Loughborough University London, London, UK
e-mails: {H.Aliyu-Tanko-17@student., y.rahulamathavan@}lboro.ac.uk

ABSTRACT

This paper provides some preliminary results for a possible novel side channel attack on Android smart phones. This attack collects accelerometer readings when users type on soft keyboards. The work has the following two parts: 1) differentiate users based on sensor readings and 2) identify whether the user belongs to a particular nationality i.e., Chinese in this work. This work uses a novel signal processing technique along with random forest machine learning algorithm to extract unique features belong to Chinese nationalities. We collected more than 2000 keystrokes data from six users where three of them are Chinese nationals. Our model has correctly identified 86% of the sensor data to classify Chinese nationality. Since any apps installed on Android device can listen to the accelerometer sensor data, the side channel attack presented in this work demonstrates another potential privacy vulnerability which could be exploited by malicious apps for targeted activities such as advertisements.

Index Terms— keystrokes inference, motion sensors, machine learning, Random forest algorithms, Android.

1. INTRODUCTION

In the last few decades, the smartphones have become an essential part of every human daily life, the smartphone equipped with high computational capabilities, easy access to the internet, storage capacity and portability has led to the surge use of the devices for important transactions, like inputting PINs, credit card information.

As todays smartphone get more smarter with addition of some sophisticated tools and sensors such as the camera, microphone and motion sensors (accelerometer, gyroscope, e.t.c) to improve performance and UX, the question of privacy and security was raised. To demonstrate this side channel attack, we developed a malicious application called SensorReader, that has access to the motion sensors readings. The application stealthily monitors the motion sensors variations as the user is tapping on the touchscreen. Figure 1 shows the sensor readings of a user who typed letter 'A' and 'L' multiple times. As shown in Fig. 1, the sensor readings are clearly separated in three-dimensional space.

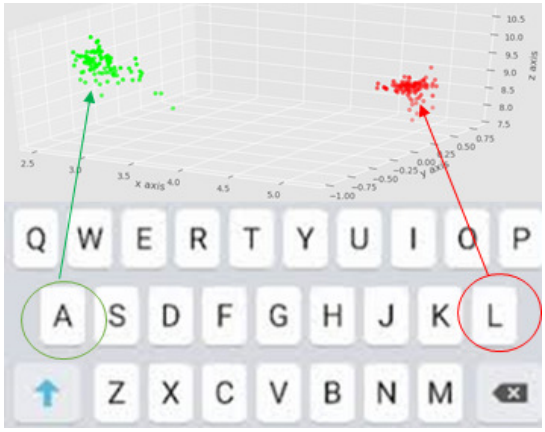

Fig. 1. Accelerometer sensor readings of letter A (green) and L (red).

Several research works have exploited the feasibility of keystrokes inferencing based on the motion sensors variation readings, this research work was based on individual keystrokes pattern. We analysed the sensor readings from 6 volunteers, the analysis showed that the keystrokes pattern for each user is 90+ percent unique to each individual, with some level of overlapping among users from the same nationality (China). the overlap between same nation users led to a further analysis based on nationalities, we showed that, the keystrokes pattern of people from the same nation (China) has significant different from people of other nations. To be precise, we showed that, the characters (I, N, O, P, T, U and M) show the difference in the pattern. To my knowledge this is the first work that looked at the possibility of inferencing keystrokes patterns based on nationality.

As smartphones are getting light in weight, it is now easy to interact with one hand even while standing. This work is focused on a controlled scenario where each user is standing and using the smartphone with one hand (right hand).

The objective of the work is; The different and similarity of keystrokes pattern analysis among users and also the possibility of nationality (China) prediction based on the keystroke

This work was supported by the UK-India Education Research Initiative (UKIERI). Ref No. UGC-UKIERI-2016-17-019..

pattern.

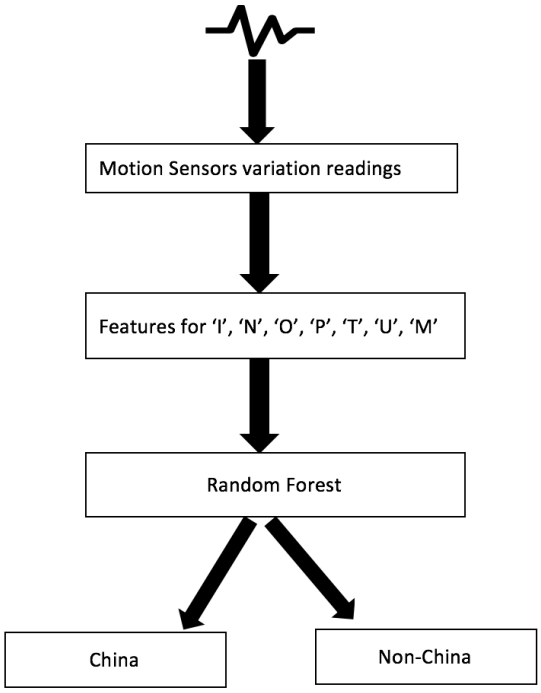

Fig. 2. The accelerometer sensor data goes thorough a few filters to extract features of key characters followed by classification.

2. RELATED WORK

Previously, smartphone sensor readings have been used as a unique identifier to authenticate different devices [6, 7, 8, 9]. These works clearly shows the sensitivity of on-board sensors. There are works on this domain that performs active attack on smartphone by exploiting and inserting malware programs within applications [10, 11]. However, the focus of this paper is on reviewing passive attacks.

Several research works have shown that, the motion sensors (e.g. accelerometer, gyroscope) which are considered insensitive sensors by smartphone manufacturers, are susceptible to side channel attacks. Different types of attacks were explored, from location inferencing to keystrokes inferencing [2, 3, 4, 5].

Jun et al., gave feasibility of location inferencing based on the accelerometer sensor readings on a smartphone, since the accelerometer measures the non-gravitation velocity, using the accelerometer trajectory variation readings, they were able to narrow down the possible movement of the user with the smartphone accelerometer readings [2].

One of the first research work that focuses on the smartphone motion sensors variation readings to infer and analyse keystrokes pattern was presented in [3]. Philip et al., developed an application (sp)iPhone which uses motion sensors in an iPhone 4 placed on a table to infer the keystrokes of a nearby laptop, as the user was typing, the tapping caused vibrations that move through the table and rattled the motion sensor (accelerometer) and it caused variation [3].

Similarly, Emmanuel, et al., showed that, accelerometer sensor readings in a smart phone are a powerful side channel attack feature in inferring users password [4].

Ahmed, et al focused on addressing the question, which available sensors can perform best in the context of the inference attack [5]. They considered all the available sensors and the integration of all the sensors data in a single dataset and compared each sensors performance in relation to keystrokes inference attack.

In most of the previous works, the possibility of the side channel attacks was the main objective, this work was focused on the keystrokes pattern rather than the attacks, in addition, the distinction and similarity of the keystrokes patterns among users based on their nationalities was emphasized.

3. METHODOLOGY

This work involves designing a malicious application that captures the motion sensor data readings when user types on QWERTY soft keyboard on an Android device, the variation sensors readings was extracted along with other features from the data and a machine learning algorithm was developed to classify the data into different classifications for analysis. Fig. 2 shows the methodology at high-level.

3.1. Tools and machine model

The range of tools to be use for this project include:

- **An htc one A9s Android smartphone** with an on-board motion sensor. For this work, the accelerometer and gyroscope sensors was used.
- **The application:** an android application is created with interfaces that allow users to type the inputs. the application stealthily captures the readings, it saves the data to be transferred into a workstation.
- **Weka machine learning suite tool:** is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is used for data mining, classification, regression and feature selections.
- **Random forest algorithm:** random forest is a supervised ensemble machine algorithm that build multiple decision trees to get better accuracy. The decision trees node is selected at random, in the end, the highest number of outcomes from all the decision trees in the forest is voted as the most likely class.

1. Randomly select a feature K among all the features of 26 characters.
2. From the K feature, calculate the nodes best separation of the feature.
3. Split the nodes into other nodes using the best separation.
4. Repeat the A to C steps until all the other features nodes has been created.
5. Build the forest by repeating steps A to D for n number times to create n number of trees.

3.2. Type of data

Type data used for this research work;

- Accelerometer sensor: monitors the acceleration of the device when the device is in motion in three axes: left-right(x-axis), forward-backward(y-axis) and the up-down(z-axis). For example, the readings will be positive when the device is accelerating in each of the directions.
- The gyroscope sensor: monitors the rotation or twisting of a smartphone with respect to gravity. It is calibrated in three axes. It measures angular velocity. It has three axes: When the device rotates along the Z-axis (perpendicular to the screen plane) azimuth angle changes in [0,360], when the device rotates along the X-axis (pitch angle) changes in [180,180), when the device rotates along the Y-axis (roll angle) changes in [90,90).

3.3. Data collection

The data collection is done in a controlled manner where users are asked standing and holding the smartphone with the right hand. For this project the focused is placed on the right-handed people. **Data collection from other users:** Data was collected from six (6) users for further to analysis. Each user is asked to input some random words into the application in the controlled manner. The random words contain 114 words making up of 626 characters (626 characters or 20 to 25 characters per each alphabet). The volunteers are from different nationalities, three are from China and the rest making up of different nationals. Also four of them are males and two of them are females.

3.4. Feature Extraction

In this work, we consider three readings for each tap for each direction. For example, for x-axis direction, x_1 denotes the reading before the tap, x_2 denotes the reading during the tap and x_3 denotes the reading after the tap. Similarly, we consider y_1, y_2, and y_3 for y-axis and z_1, z_2, and z_3 for z-axis. We use these raw values to define the following six features for each direction, totalling eighteen features. Let us define the six features in x-axis as follows:

min_x: this is the minimum value in x direction i.e., $min_x = minimum\{x_1,\ x_2,\ x_3\}$.

max_x: this is the maximum value in x direction i.e., $max_x = maximum\{x_1,\ x_2,\ x_3\}$.

$mean_x$: the average value of three outputs in the x-axis direction i.e., $mean_x = (x_1 + x_2 + x_3)/3$.

$median_x$: is the middle of the three outputs.

std_x: denotes the standard deviation. It is the amount of variations of each output from the corresponding average.

$$std_x = \sqrt[2]{\frac{\sum_{i=1}^{3}(x_i - mean_x)^2}{3}}$$

$skewness_x$: is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.

$$skewness_x = \frac{3(mean_x - median_x)}{std_x}$$

4. EVALUATION

In this section, we perform two different tests on the collected dataset. In the first test, we build a machine learning model to identify a user. In the second test, we build machine learning model to identify whether a user belongs to a particular national not i.e., Chinese national or non-Chinese.

4.1. User identification

As mentioned before, the collected dataset contains sensor readings from six volunteers. Hence the user identification problem becomes a six-class machine learning problem. For training and testing, we use 10-cross validation technique i.e., use 90% of each user data for training and the remaining 10% of the data testing. This process is repeated at least ten times.

For training and testing, we consider random forest machine learning algorithm. We noticed that random forest algorithm outperforms other popular algorithms such as support vector machines. The results are presented as a confusion matrix in Table 1. Table 1 shows the unique nature of keystrokes pattern for each user. We believe this is linked to the physical

Table 1. Confusion Matrix per Participants

Users	Male				Female		
	A	B	C	D	A	B	Accuracy
Male A	**585**	1	0	0	0	0	99.83%
Male B	0	**566**	0	0	0	0	100%
Male C	0	0	**603**	0	0	0	100%
Male D	0	0	0	**584**	40	1	93.44%
Female A	0	0	0	27	**611**	6	94.87%
Female B	0	0	0	1	7	**532**	98.52%
Average Accuracy							**97.67%**

structure of users and also we believe that the way each individual is holding the phone and the force of the keystrokes are unique (similar to fingerprints). However, this hypothesis may require further experiments for validation.

4.2. Ethnicity identification

From the classification matrix in Table 1, it showed that there is some similarity of the keystrokes pattern from the last three users, whom happened to be from the same nation (China).

Given that the dataset is made up of 50% of the data from Chinese nationals, the dataset was grouped into two: China and non-China. Hence becomes a two-class machine learning problem.

In order to train and test, we used the 66% of the dataset consisting of data from two Chinese and two non-Chinese users as training set and used the remaining as a test set. The confusion matrix for this initial test is shown in Table 2.

Table 2 shows both the training and test results. During the training, more than 97% of the characters are correctly divided into two classes i.e., Chinese and non-Chinese. However, during the testing phase, only 71% of the characters are correctly classified into right class. To improve the accuracy we conducted another test based on specific characters which is described in the next subsection.

Table 2. Confusion matrix per nationality when all characters are used for training and classification.

	Training		Testing	
	Chinese	Non-Chinese	Chinese	Non-Chinese
Chinese	1244	25	540	0
Non-Chinese	39	1112	329	274
	Training Accuracy: 97.36%		**Testing Accuracy: 71.22%**	

4.3. Ethnicity identification using specific characters

The confusion matrix in Table 2 shows a 100 percent accuracy for the user from China and only 45 percent for the user from non-China nation. This suggests that, there are some specific characters among the 26 characters that distinguished the nature of how people from the China tap on the soft keypad with other people from different nation.

After further analysis of each character, we noticed that there are seven characters (I, N, O, P, T, U and M) which are correctly classified into right class compared to other characters. Hence, we filtered the sensor data for these characters build another model. Again, we used 66 percent of the filtered data for training and the remaining for testing. The confusion matrix for this experiment is shown in Table 3. The model has an accuracy of 86.62 percent.

Table 3. Confusion matrix per nationality when only characters I, N, O, P, T, and U are used for training and classification.

	Training		Testing	
	Chinese	Non-Chinese	Chinese	Non-Chinese
Chinese	409	19	191	1
Non-Chinese	33	373	54	165
	Training Accuracy: 93.76%		**Testing Accuracy: 86.62%**	

5. DISCUSSION

This research focused only on Chinese nationals since the majority of the data is collected from one nation (China). User identification analysis showed that each user has a unique typing pattern. This may be due to the fact that the length of the fingers, strength of the hand grip and force of the keystrokes of the touch-screen vary for each user.

The dataset was grouped into two, China and non-china, Further analysis between the two groups, some unique characters, (I, N, O, P, T, U and M) that differentiate the typing pattern between the two groups with high accuracy. Based on the model created from these unique characters, with an accuracy of 86 percent, it is possible to predict the users nation between the two groups.

This diversity analysis shows the possibilities of division of how people type on the smartphone screens due to some factors like the length of fingers, how consistent different people use certain characters, like in English, the most common used characters are the vowels (A E I O U). This is open for further research.

6. CONCLUSION

This work showed that it is possible to distinguish different people or even their ethnicity from smartphone sensor readings with high accuracy. Since this work shows that it is possible to infer some sensitive information of a user indirectly, this work can be classified as a side channel attack on Android device. This attack could be used by any applications on Android to exploit the user with tailored advertisement.

7. REFERENCES

[1] Miluzzo, E., Varshavsky, A., Balakrishnan, S. and Choudhury, R.R., 2012, June. Tapprints: your finger taps have fingerprints. In Proceedings of the 10th international conference on Mobile systems, applications, and services (pp. 323-336). ACm.

[2] Han, J., Owusu, E., Nguyen, L.T., Perrig, A. and Zhang, J., 2012, January. Accomplice: Location inference us-

ing accelerometers on smartphones. In Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on (pp. 1-9). IEEE.

[3] Marquardt, P., Verma, A., Carter, H. and Traynor, P., 2011, October. (sp) iPhone: decoding vibrations from nearby keyboards using mobile phone accelerometers. In Proceedings of the 18th ACM conference on Computer and communications security (pp. 551-562). ACM.

[4] Owusu, E., Han, J., Das, S., Perrig, A. and Zhang, J., 2012. Accessory: Keystroke inference using accelerometers on smartphones. Proc. of HotMobile.

[5] Al-Haiqi, A., Ismail, M. and Nordin, R., 2013. On the best sensor for keystrokes inference attack on android. Procedia Technology, 8, pp.947-953.

[6] Amerini, I., Becarelli, R., Caldelli, R., Melani, A. and Niccolai, M., 2017. Smartphone fingerprinting combining features of on-board sensors. IEEE Transactions on Information Forensics and Security, 12(10), pp.2457-2466.

[7] Dey, S., Roy, N., Xu, W., Choudhury, R.R. and Nelakuditi, S., 2014, February. AccelPrint: Imperfections of Accelerometers Make Smartphones Trackable. In NDSS.

[8] Bojinov, H., Michalevsky, Y., Nakibly, G. and Boneh, D., 2014. Mobile device identification via sensor fingerprinting. arXiv preprint arXiv:1408.1416.

[9] Amerini, I., Bestagini, P., Bondi, L., Caldelli, R., Casini, M. and Tubaro, S., 2016. Robust smartphone fingerprint by mixing device sensors features for mobile strong authentication. Electronic Imaging, 2016(8), pp.1-8.

[10] Idrees, F., Rajarajan, M., Conti, M., Chen, T.M. and Rahulamathavan, Y., 2017. PIndroid: A novel Android malware detection system using ensemble learning methods. Computers & Security, 68, pp.36-46.

[11] Rahulamathavan, Y., Moonsamy, V., Batten, L., Shunliang, S. and Rajarajan, M., 2014, July. An analysis of tracking settings in Blackberry 10 and Windows Phone 8 Smartphones. In Australasian Conference on Information Security and Privacy (pp. 430-437). Springer, Cham.