International Journal of Pervasive Computing and Comm

TDAS: a touch dynamics based multi-factor authentication solution for mobile devices

Pin Shen Teh Ning Zhang Andrew Beng Jin Teoh Ke Chen

TDAS: a touch dynamics based multi-factor authentication solution for mobile devices

Pin Shen Teh and Ning Zhang
School of Computer Science, University of Manchester, Manchester, UK
Andrew Beng Jin Teoh
College of Engineering, Yonsei University, Seoul, Republic of Korea, and
Ke Chen
School of Computer Science, University of Manchester, Manchester, UK

Received 1 July 2015
Revised 4 October 2015
Accepted 1 February 2016

Abstract

Purpose – The use of mobile devices in handling our daily activities that involve the storage or access of sensitive data (e.g. on-line banking, paperless prescription services, etc.) is becoming very common. These mobile electronic services typically use a knowledge-based authentication method to authenticate a user (claimed identity). However, this authentication method is vulnerable to several security attacks. To counter the attacks and to make the authentication process more secure, this paper aims to investigate the use of touch dynamics biometrics in conjunction with a personal identification number (PIN)-based authentication method, and demonstrate its benefits in terms of strengthening the security of authentication services for mobile devices.

Design/methodology/approach – The investigation has made use of three light-weighted matching functions and a comprehensive reference data set collected from 150 subjects.

Findings – The investigative results show that, with this multi-factor authentication approach, even when the PIN is exposed, as much as nine out of ten impersonation attempts can be successfully identified. It has also been discovered that the accuracy performance can be increased by combining different feature data types and by increasing the input string length.

Originality/value – The novel contributions of this paper are twofold. Firstly, it describes how a comprehensive experiment is set up to collect touch dynamics biometrics data, and the set of collected data is being made publically available, which may facilitate further research in the problem domain. Secondly, the paper demonstrates how the data set may be used to strengthen the protection of resources that are accessible via mobile devices.

Keywords Touch dynamics, Biometrics, Benchmark data set, Keystroke dynamics, Mobile authentication

Paper type Research paper

1. Introduction

Mobile devices have become an integral part of our routine activities. This is particularly the case with the rapid growth and widespread use of smartphones and digital tablets. The processing capability of these devices has advanced up to the point

This research work was supported by the University of Manchester. The authors would also like to thank all the people who have contributed their time and effort to the creation of this data set, and also to the reviewers for their valuable time and insightful comments.

International Journal of Pervasive Computing and Communications
Vol. 12 No. 1, 2016
pp. 127-153
© Emerald Group Publishing Limited
1742-7371
DOI 10.1108/IJPCC-01-2016-0005

that most digital activities that can be accomplished on workstations or laptops can also be performed on these portable devices. Routine activities, such as personal and corporate e-mail communications, on-line banking transactions, accessing paperless prescriptions services, route navigation, etc., can also be carried out ubiquitously with these devices.

According to a forecast by networking giant Cisco (2015), there will be approximately 11.5 billion mobile-connected devices by 2019, and the global mobile data traffic will increase nearly tenfold between 2014 and 2019, reaching 24.3 exabytes per month by 2019. This shows our increasing reliance on mobile devices, and also implies that our private and sensitive data will increasingly be handled, managed and processed by these devices. Therefore, the security of accessing mobile devices and accessing data, services and other resources through the mobile devices is of a prime concern. More stringent security services (or measures) should be embedded in mobile devices. One of these services is user authentication, i.e. how to securely verify a claimed identity.

Authentication is the first-line of defence in any computer system or device, as it is a pre-requisite for several other security services such as authorisation and accountability. In a mobile device context, authentication is typically achieved via a knowledge-based authentication method, and with this method, a user proves their identity by demonstrating the knowledge of a secret. This secret could be a PIN (personal identification number), a password, a shared secret (which is similar to a password, but with higher entropy) or a private key corresponding to a public key certified in a digital certificate. Authentication by using a low entropy secret (such as a PIN or a password) is vulnerable to a number of security attacks, e.g. theft of a mobile device, shoulder spoofing and brute force attacks. Authentication by using a high entropy secret (using a shared secret or a private key) usually requires some means to store the secret securely and this may hinder usability and/or introduce a higher cost. To address these issues and to make the authentication service more secure, while, at the same time, without hindering usability, we have been working on integrating a biometric-based (touch dynamics) authentication method with a knowledge-based (PIN) authentication method, called a touch dynamic-based multi-factor authentication solution (TDAS).

This paper reports our effort on the creation and use of a touch dynamics data set to investigate the benefits of integrating touch dynamics with a PIN-based authentication method. It describes the process of collecting a comprehensive reference data set consisting of two sets of input PINs from 150 subjects, the extractions of feature data from the data set and the use of the extracted feature data as a second authentication factor. The extracted feature data are timing, finger touch size and pressure feature data. The data are classified by using three light-weight classification algorithms, and are used to support the identification of a user in addition to using the PIN-based authentication method. The experiments show that this two-factor approach to authentication can make impersonation attacks much more difficult, significantly increasing the assurance level of the authentication process. The paper has also discussed various factors that may impact the accuracy performance of TDAS, such as the timing resolution of feature data type, combinations of feature data types, subject size and input string length.

The structure of this paper is as follows. The next section gives an overview of touch dynamics biometrics. Section 3 explains the experimental setup and the methods and procedures used in collecting (i.e. acquiring) the touch dynamic biometrics data. Section 4 describes the properties of the data set. Section 5 describes potential feature data that can be extracted from the data set and, for proof-of-concept, it illustrates how the captured feature data may be used for authenticating users in a mobile context and the level of improvements in terms of accuracy performance (Section 6). Section 7 critically analyses the related work in the problem context, compares our data set with other public data sets and outlines open issues and potential research directions. Finally, Section 8 concludes the paper.

2. Background

2.1 Overview of touch dynamics

Prior to the emergence of touch dynamics, one of the earliest research work on using keystroke dynamics (i.e. the patterns of interactions between human input and a physical keyboard) to identify users was by Gaines *et al.* (1980). They attempted to recognise six professional secretaries by analysing the way they typed three passages of text consisting of 300 to 400 words each. Since then, keystroke dynamics either on workstations (Bleha *et al.*, 1990; Obaidat, 1995) or in a Web-based environment (Cho *et al.*, 2000; Stewart *et al.*, 2011) have been the major topics of research. With the rapid development of mobile communication technologies, recent research efforts in the area have been focusing on mobile devices with physical keypads (Campisi *et al.*, 2009; Clarke and Furnell, 2007). According to the literature survey of related works in the topic area of keystroke dynamics (Teh *et al.*, 2013), since 2007, there have been growing efforts on examining the possibility of applying the concept of keystroke dynamics to user authentication on mobile platforms. More recently, the research activities are largely carried out in the context of touchscreen mobile devices (Saravanan *et al.*, 2014; Dhage *et al.*, 2015).

Touch dynamics refers to the process of measuring and assessing human touch rhythm on mobile devices, such as digital tablets, smartphones or touchscreen panels. When a human interacts with a mobile device, a digital signature is generated. The signatures generated by different individuals are believed to be rich in discriminative properties, which hold potentials as personal identifiers. The availability of higher-resolution sensors in recently released mobile devices provides added opportunities to the development of touch dynamics biometrics, as these sensors allow the extraction of more discriminative feature data types. Touch dynamics biometrics can be integrated with an existing knowledge-based authentication method to form a so-called multi-factor authentication solution. Such a solution can make unauthorised accesses to mobile devices harder, thus strengthening the security level of mobile devices. Using touch dynamics to identify a user has its unique advantages, and, at the same time, it also introduces challenging issues. The following highlights the advantages and challenging issues.

2.2 Advantages

- *Transparency*: A touch dynamics-based authentication method requires little or no additional interventions from a device user. This is because the capture and processing of touch patterns can be carried out in the background while the user is using the device. Users may not even be aware that they are being authenticated periodically or are being protected by an extra layer of authentication. This is in a stark contrast to other biometrics based authentication methods that usually requires explicit alignment of a biometrics feature to a specific sensor. For example, in the case of iris recognition, a user is required to look straight into an intra-red camera to take an iris image, and in the case of finger based authentication, a user needs to put one of his/her fingers on the fingerprint sensor.

- *Familiarity*: The touch dynamics feature data used for authentication are collected during mobile users' routine input activities. This is a process which mobile users are already familiar with, so the feature data collected tend to have a gentler learning curve with a higher usability level.
- *Revocability*: The touch dynamics feature data can be replaced should a passcode associated to a touch dynamics pattern be compromised, as a new touch dynamics template can easily be associated to a new passcode. However, this is not the case for other physiological biometrics, e.g. for iris or face biometrics, once they are compromised, there will be no replacement, and for fingerprints, the number of replacements is limited (there are only ten fingers to use after all).
- *Non-dependency*: The feature acquisition of touch dynamics is less sensitive to environmental factors. Therefore, it is more suited to, and can be more easily deployed in, a mobile context. A mobile device is usually operated in an on-the-go manner, so the values of some environmental factors, such as the screen lighting level and background noise level, are constantly changing. Other biometrics features such as iris or voice biometrics are sensitive to these environmental factors.
- *Cost-effectiveness*: In contrast to other physiological biometrics systems such as iris and fingerprint recognitions, which typically require the use of some specialist hardware, touch dynamics recognition only requires the use of build-in mobile sensors. This can reduce device costs and it is ideal for large-scale deployments.

2.3 Challenging issues

- *Algorithm and communication costs*: Computational capabilities of mobile devices are typically lower than desktop computers. This means that certain criteria such as algorithm complexity, communication cost and authentication delay are important and should be considered in the design of touch dynamics-based authentication solutions. In other words, computational and communication costs introduced as the result of deploying this authentication means should be minimal.
- *Energy consumption*: Mobile devices, unlike their desktop counterparts, are typically battery-powered, so the less the energy an application consumes, the longer the device can operate. Though communication is the major consumer of the power of a device, the number and usage frequencies of various sensors embedded in a mobile device, which are used to extract touch feature data, also have a direct impact on the mobile device battery consumption level. Various measures, such as reducing the sampling rate or frequency of data sensing Niu and Chen, 2012), or performing complex computations only when a device is being recharged (Crawford *et al.*, 2013), have been proposed to reduce power consumption of a mobile device.

- *Non-permanence*: Usually, human behavioural characteristics change more frequently than physiological characteristics. Similarly, a user's touch dynamics may change gradually, as the user is getting more familiar with the passcode, input method, device and other external factors. Adaptive approaches (Xu *et al.*, 2014) have been proposed in literature to take into account any gradual changes in touch patterns.

2.4 Operational process

Figure 1 shows a typical touch dynamics biometrics authentication system. From the figure, we can see that the operation of this system can largely be captured in two major phases:

(1) user enrolment, where touch pattern feature data are collected, processed and stored as a reference template; and

(2) user authentication, where a test sample is compared against the stored reference template(s) to determine the similarity.

The two operational phases are accomplished by a number of functional blocks (i.e. architectural components), each of which performs a well-defined function or operation. These components and their respective operations are described as follows:

- *Data acquisition*: It is an operation by which raw touch feature data are collected. This is usually carried out as the first step and during the setup stage of a touch dynamics biometrics authentication system. The collected raw feature data typically consist of repetitive input samples or collections of input samples over a specified period. Devices commonly used in the data acquisition experiments are commercial off-the-shelf smartphones (Buschek *et al.*, 2015) or, on some occasions, digital tablets (Saravanan *et al.*, 2014).
- *Data pre-processing*: Once the raw feature data are collected, a data pre-processing operation is carried out to remove outliers in the raw feature data, improving data quality and accuracy performance. Additionally, to optimise computational efficiency on resource-limited mobile devices, a dimension reduction technique may also be used to ensure that the selected raw feature data remain small yet representable (de Mendizabal-Vazquez *et al.*, 2014).

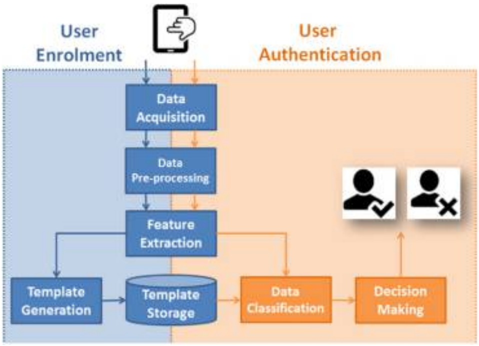

Figure 1.
Touch dynamics biometrics authentication framework

- *Feature extraction*: Feature extraction is another mandatory operation which is carried out in both the enrolment and authentication phases. The main task of this operation is to identify and extract distinctive features common to a subject from the collected raw data. These features will later be used for template generations.
- *Template generation*: Depending on the classification algorithm selected, the template generation can be an operation that is used to transform the extracted touch feature into a compact form or an operation that trains, selects and stores classification models that uniquely represent each user's touch dynamics characteristics.
- *Data classification*: This is the major operation for most biometrics systems, where feature data are categorised and compared against reference templates or models. The outcome of this operation is normally associated with a matching score used for decision-making.
- *Decision-making*: This is the last operation carried out by a touch dynamics authentication system in authenticating a user. It is to determine if a presented touch pattern has indeed originated from the target subject. The decision is usually made by comparing the similarity or dissimilarity score generated from a classifier against a predefined threshold (Bo *et al.*, 2014).

3. Experiment setup

This section describes and justifies the experiment setup in which raw touch feature data are collected.

3.1 Deployment and working modes

A touch dynamics biometrics system may be deployed in one of the two modes, identification and verification. Each of these modes functions uniquely and serves a different purpose and use-case scenario. The purpose served by the identification mode is to classify unknown touch dynamics samples. This mode is typically deployed for forensic investigations and intrusion detections. Its use on mobile devices is rather limited. The verification mode, on the other hand, is typically used to prove or verify a claimed identity.

The verification mode of touch dynamics biometrics can be further divided into two working modes: a static and a dynamic working mode. In the static working mode, a user is authenticated at the initial instance of, or, at some predefined intervals, during, the user-to-system interaction. Unlike the case for the static working mode, in the dynamic working mode, a user may be authenticated at any instant of a user-to-system interaction or for every service access (i.e. continuously) throughout a service access session (in addition to the initial authentication). The functions performed in both modes are complementary, which means that they can be deployed alongside each other to enhance the security of mobile devices or the security of service access using mobile devices. Our experiment is conducted under the assumption that the static working mode of the verification mode is used.

3.2 Data collection devices

As mentioned above, there are different types of mobile devices, e.g. featured phones, smartphones, digital tablets and laptops. Most research works in relation to touch dynamics are carried out using various mobile phones. In comparison with mobile phones, digital tablets have a relatively larger screen resolution, which means that a higher subject input variation, and therefore a better feature discrimination, can be captured (Saravanan *et al.*, 2014). For this reason, we have chosen to use a digital tablet as our data collection device. The device is a commercial off-the-shelf Samsung Galaxy Tab 10.1 (GT-P7510) digital tablet. It has a 10.1" widescreen, and is powered by 1-GHz dual-core processor and equipped with a 1-GB RAM. The device runs under Android 4.0.4 (Ice Cream Sandwich) and a data collection tool that was developed using Java and Android API Level 15. The entire data collection process was performed using this tablet. The justification for using a predefined device, rather than the devices chosen by the subjects, is to remove any uncontrolled variables such as subject preferences, program compatibility and functionality differences. In this way, the results obtained from the experiment can better capture the discriminative power of touch dynamics feature data and the classification algorithm used.

3.3 Data collection environments

There are three main environments in which raw touch feature data may be collected:

(1) while the subjects carry out their activities as usual;

(2) under a controlled laboratory environment in a fixed location; or

(3) in multiple fixed locations.

The first option is expensive, as due to the ubiquitous nature of mobile devices, subjects are likely to be on-the-move and following the subjects while collecting the data may not be convenient and could be costly. The second option is least costly in terms of setting up and running the experiments, but the data collected may not give a true reflection of real-world scenarios. To balance costs/feasibility with real-life situations, we have chosen to use the third option, i.e. we let subjects to choose their preferred locations where their touch dynamics are extracted. The locations used included offices, homes, inside cars, classrooms, cafés and public areas.

3.4 Collection method

Data should be collected when subjects are in a stable state, i.e. after they are familiar with the device input facility and the data collection procedure, as, otherwise, data collected may not properly capture subjects' input features. Data captured improperly may increase false acceptance and false rejection rates when they are used to authenticate the subjects. According to Ngugi *et al.* (2011b), input patterns, styles or speeds can vary and stabilise over time. To ensure data are collected after the patterns, styles and speeds are stabilised, and to reduce the effect of subjects' unfamiliarity with the input facility and procedure, one of the two approaches may be used. The first one is to divide a data collection session into multiple sub-sessions and the sub-sessions are separated by a selected time frame. For example, we may have four sub-sessions each with 1 week apart, and the data collected in multiple sub-sessions are cumulatively merged into a single set. This approach provides a good level of accuracy, but may suffer from a high dropout rate (De Luca *et al.*, 2012), as the data collection process is usually carried out on a voluntary basis and we cannot expect all the participants to take part in all the sub-sessions. As a result, the sample size of the data set may be reduced. The second approach is to collect data in a single session, but the subjects are asked to familiarise themselves with the input facility and procedure as many times as necessary before a data collection process actually takes place. The entire data collection process takes an average of 15 to 20 min (in addition to the time taken to familiarise the input device and procedure). This latter approach is commonly used in experiments reported in the literature (Chong *et al.*, 2010; Kim *et al.*, 2010; Loy *et al.*, 2007). So in our experiments, we have chosen to use the second approach.

4. Data set

4.1 Subject size

The subject size refers to the number of subjects from whom data are collected. Typically, a subject size of greater than 100 subjects is regarded as a large subject size (Teh *et al.*, 2013). Using a larger subject size can provide more data to verify the scalability of a chosen classifier, as mentioned by Clarke and Furnell (2007). Most of the relevant works in the touch dynamics domain were carried out with a subject size smaller than this value. Only a handful of published works (Gascon *et al.*, 2014; Serwadda *et al.*, 2013; Trojahn *et al.*, 2013) use a subject size greater than 100 subjects. However, in the latter group of works, the subjects involved were restricted to a certain profession and the data sets were not made publically available for evaluation. At the time of this writing, our data set contains touch dynamics data of 150 subjects with a diverse range of professions. The data set is grouped in three different packages. The first package contains data from 50 subjects, the second package contains data from 100 subjects and the third package contains data from all of the 150 subjects. In this way, the data sets may be used, in different ways, for comparisons between different subject sizes within the same subject grouping.

4.2 Subject demography

Experimental subjects are typically selected based on some criteria. The commonly used criteria are age distribution, population mixture and profession diversity. Subjects from different age groups, with different backgrounds and/or different professions, tend to use their devices at different frequencies. Therefore, if the subjects are not selected properly, there may be unintended bias in the experimental results. To reflect real-world situations as much as possible, the demography of the subjects taking part in the data collection should be as diverse as possible. In other words, people from different age groups, of different genders and with different device usage frequencies should be represented as much as possible.

Most of the published works are based on subjects that are selected from:

- a narrow age distribution (i.e. 19-26) (Antal and Szabó, 2014);
- confined to only people within the same organisation (i.e. within research institute) (Giuffrida *et al.*, 2014); or
- restricted to limited profession (i.e. students) (Draffin *et al.*, 2014).

These options are frequently selected because they are less costly and may be easier to conduct the experiments. Data collected in this way may not properly capture the biometric features from a wider community. El-Abed *et al.* (2014), Kolly *et al.* (2012) and Trojahn *et al.* (2013) are the few pieces of work we are able to find in the literature, which recruited subjects from dissimilar age groups, and diverse population groups and professions. Inspired by these works, we have made our best effort to reach out to the general public within our resource budget. Table I summarised the demography of the subjects recruited in our experiment.

4.3 Input type

Input string types are an important experimental variable that should be considered in touch dynamics biometrics research. This is because the feature used for touch dynamics biometrics is extracted from a subject's input string. Generally, experimental subjects are required to provide character-based, numerical-based (PIN) or other non-specific touch events. PIN input has been the most widely used authentication method for mobile devices, so we first used a 4-digit numerical input ("5560"), and then a 16-digit numerical input ("1379666624680852"). The use of two different PIN lengths allows us to evaluate the effects of different input string lengths experimentally. These two predefined numbers were carefully chosen with the following key positioning combination strategies:

- *Apart*: Keys are separated by at least one key apart.
- *Repetition*: Reoccurrence of identical key.
- *Adjacent*: Keys located diagonally to each other.
- *Sequence*: Keys situated horizontally or vertically to each other.

These positioning strategies were used to spread the variety of input strings. A graphical illustration of the approach is depicted in Figure 2. Requiring all the subjects to use a single predefined input string (for each string length) allows us to increase the number of impersonation samples available for our testing phase (Section 6.2) without the need for collecting additional data.

Properties		Details	
Subjects	50	100	150
Population			
Academia	9	11	18
Public	41	88	132
Age			
<20	8	13	28
20-40	27	49	66
>40	15	38	56
Gender			
Male	12	24	45
Female	38	76	105
Usage frequency			
Rare	16	32	49
Average	12	20	32
Often	22	48	69
Hand preference			
Left-handed	5	8	14
Right-handed	45	92	136

Table I. Subject demography of our data set

4.4 Sample size

The input sample size affects the accuracy, robustness and outcome of an experiment (Sen and Muralidharan, 2014). A larger number of samples used would allow us to gain a better representation of a subject's touch dynamics behaviour, and, as a result, achieve a higher level of accuracy (Tasia *et al.*, 2014). Though it is desirable to have a large number of sample data, it is not always practical to expect subjects to repeat the input for a large number of times during the enrolment stage. This is because subjects may not be available for a long stretch of time, or may feel uncomfortable with a lengthy acquisition procedure (Tasia *et al.*, 2014). Therefore, selecting an optimal input sample size for each data acquisition session is necessary.

According to the work reported in literature, the benchmark for the number of samples collected per subject per session is somewhere between 10 to 20 repetitions for a fixed input type (Buschek *et al.*, 2015; Sen and Muralidharan, 2014; Trojahn *et al.*, 2013). Therefore, in our data collection process, subjects were required to repeat each input string for 10 consecutive times, resulting in 20 samples per subject (10 for the short digit sample and 10 for the long digit sample). In terms of error handling, any input mistake made by a subject was automatically discarded and the subject was prompted to repeat that particular input sample, and this is a common practice, as explained in the literature (Campisi *et al.*, 2009; Maiorana *et al.*, 2011; Robinson *et al.*, 1998).

4.5 Raw feature representation

A number of application programming interfaces (APIs) have been used to capture a subject's feature data. In detail, each single-screen touch event (finger touching down or lifting up from the touchscreen) is detected by the onTouchListener API. The timestamps of each key press and release are logged by invoking the nanoTime() API. This API returns the most precise time (in nanoseconds) that is available on the device. We also use the API functions getSize() and getPressure() under the MotionEvent class to retrieve the values of finger circumference and pressure, respectively. These functions return a normalised decimal value between 0 and 1. However, we have noticed that getPressure() always returns a value of 1.0. We have tried to resolve this issue but with no success. However, as some other devices also encounter the same problem, we anticipate that this problem will be resolved by the mobile operating system's provider in their subsequent API version update. Each completed touch event on a key generates two timestamps (t_{press} and $t_{release}$), a finger touch size (*ps*) and a pressure value (*pv*). The data collected from these events will go through feature extraction and template generation process. For each repeated input sample (*r*) of raw touch dynamics data, the feature data and the particular key press (k_{press}) and key release ($k_{release}$) are recorded using the format shown below and the recorded data are stored in a separate file for each subject:

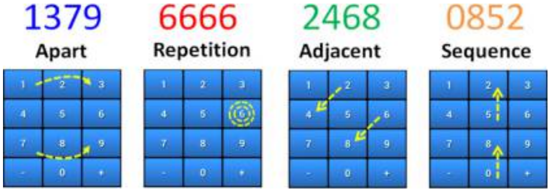

Figure 2.
Four different key positioning strategies

$$r, k_{press}, t_{press}, k_{release}, t_{release}, ps, pv$$

5. Methodology

5.1 Feature extraction

Two types of features are captured. These are timing data and finger touch size (*ps*). Both are captured during the subject-to-device interactions with the input keys. *ps* is obtained directly from the returned value of an Android API function without further customisation. However, for timing data acquisition, some manipulation to the touch event timestamps is required.

A timing data can be extracted in different feature length. The shortest feature length is known as a uni-graph, which are the timing data extracted between touch event timestamp values of the same key. Subsequently, the timing data extracted from two or more keys are called a di-graph and an n-graph, respectively. Figure 3 shows the different n-graph sizes for a given sample input string. In our experiments, we have chosen to extract timing data of a uni-graph and a di-graph.

The timing data extracted can be further divided into two categories:

(1) dwell time (*DT*), i.e. the time duration for the touch action of the same key (also known as interval, press or hold time); and

(2) flight time (*FT*), i.e. the time interval between the touch actions on two successive keys (also known as latency).

As shown in Figure 4, there are four variants of *FT* .

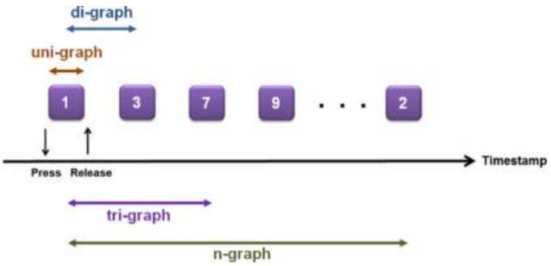

Figure 3. The different timing data feature length

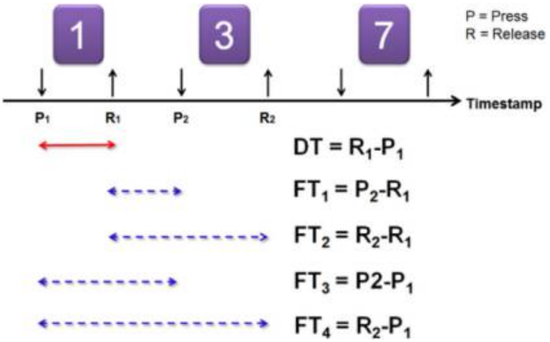

Figure 4. Types of timing feature data extracted

5.2 Template generation

Template generation is a process by which a subject's touch feature samples are combined and transformed into a compact yet representative structure. A subject's template should uniquely capture the subject's touch feature. For each subject, we use six feature data types, and, for each feature data type, template data are generated. These template data consist of two items, a mean (μ) value and a standard deviation (σ) value. The equations below show how the template data for a feature data type, DT, are calculated. For example, given a training sample set of n number of DT, the template data for DT are calculated as:

$$\mu = \frac{1}{n} \times \sum_{i=1}^{n} DT_i$$

$$\sigma = \sqrt{\frac{1}{n} \times \left(\sum_{i=1}^{n} DT_i^2 - \frac{\left(\sum_{i=1}^{n} DT_i \right)^2}{n} \right)}$$

The same procedure and computation are applied to other feature data types.

5.3 Classifier

We use three matching functions to, respectively, compute and compare the likeliness of a test sample against a reference template feature. The likeliness, which is measured in terms of a similarity score (s), is computed by feeding the test sample value (τ) of a feature vector element of position (i) and the value of μ and σ from the reference template into each function, i.e. $s_i = f(\tau_i, \mu_i, \sigma_i)$. The three matching functions used are Gaussian estimation (GE), z-score (ZS) and standard deviation drift (SD), as given below:

$$f_{GE}(\tau_i, \mu_i, \sigma_i) = e^{-\frac{(\tau_i - \mu_i)2}{2\sigma_i^2}}$$

$$f_{ZS}(\tau_i, \mu_i, \sigma_i) = \frac{|\tau_i - \mu_i|}{\sigma_i}$$

$$f_{SD}(\tau_i, \mu_i, \sigma_i) = e^{-\frac{|\tau_i - \mu_i|}{\sigma_i}}$$

We calculate a similarity score for each individual element within the intended feature vector. Then we compare these scores against an empirical threshold (φ) to make a partial decision (D_i) for each feature data element of position (i) in that feature vector, i.e.:

$$D_i = \begin{cases} 0, & s_i \le \varphi \\ 1, & s_i > \varphi \end{cases}$$

Here, 1 and 0, respectively, indicate acceptance and rejection. A final decision is then made to determine if a test sample belongs to the reference template, and this is done by using the formula:

$$D_{final} = \begin{cases} accept, & \frac{\sum_{i=1}^{n} D_i}{n} \geq 0.5 \\ reject, & \frac{\sum_{i=1}^{n} D_i}{n} < 0.5 \end{cases}$$

where n refers to the total elements in the feature vector considered, and D_{final} is the final acceptance or rejection decision of a given test sample.

6. Performance evaluation

6.1 Evaluation criteria

To evaluate the accuracy level of a biometrics authentication system, three main metrics are commonly used. These are the false acceptance rate (FAR), false rejection rate (FRR) and equal error rate (EER). FAR is the percentage ratio of the number of illegitimate trails that are falsely accepted against the total number of illegitimate trials. A lower FAR indicates fewer illegitimate trails are falsely accepted, thus the higher the security level of the biometrics authentication method. FRR is the percentage ratio of the number of legitimate trials that are falsely rejected against the total number of legitimate trials. A lower FRR indicates fewer legitimate trails are falsely rejected, thus the higher the usability of the method. The third performance metric, EER, is a single-number performance metric, which is commonly used to measure and compare the overall accuracy of different biometrics systems. EER is obtained by first plotting a graph for each of FAR and FRR against a matching threshold and then taking the interception point of the two graphs. Typically, the lower the FAR and FRR values, the lower the EER value. A lower EER value indicates a better performance of the biometrics authentication method. However, it is impractical to lower both FAR and FRR simultaneously, as FAR and FRR are negatively correlated. Therefore, in practice, we usually choose the threshold value to achieve a required security level.

6.2 Training and testing setup

For each subject recruited, two sets of ten input samples are collected, one set for the 4-digit input and the other for the 16-digit input. For each of the two input length categories, seven out of ten are used for training (i.e. for generating a template for the subject) and the remaining three for testing (i.e. for estimating the FAR and FRR values). To reduce intra-session variations, the seven samples are selected randomly from the ten-sample set. The samples used for training are not reused for testing, so that the performance assessment can be made independent of the model development. In a FAR test, a subject's template was compared against all the other subjects' testing samples. This process was reiterated for all the subjects' templates. As there are a total of 150 subjects recruited and each subject has three testing samples, the total number of illegitimate trials conducted is $150 \times (150 - 1) \times 3 = 67{,}050$. In a FRR test, a subject's template is compared against the subject's own testing samples. As there are a total of 150 subjects and each subject has three testing samples, the total number of legitimate trials is $150 \times 3 = 450$.

6.3 Results discussion

6.3.1 Timing feature data scaling. The accuracy of timing feature data extracted from timestamps is influenced by the timing resolution used, so it is important to choose an appropriate timing resolution before extracting the timing feature data. As discussed in Section 5.1, timing feature data are extracted by subtracting the timestamps of different touch actions (i.e. a key press or release). By default, the timestamps recorded by a mobile device's nanoTime() API has a timing resolution in the order of nanoseconds. This resolution is inappropriate because a human's tapping speed is usually at a much slower pace than this order. To solve this problem, an original timestamp (t) should be normalised by a scaling factor (s) to create a normalised timestamp (T_{norm}) with the chosen resolution, and this can be done by using the following formula:

$$T_{norm} = t \times e^{-s}$$

To investigate the effects of using different scaling factor values, we have calculated the EER values of timing feature data extracted from timestamps normalised with different scaling factor values. Figure 5 shows the EER values of the timing feature data type FT_4. The same trend has been observed for all the other timing feature data types. As can be seen from the figure, the EER values are higher when the scaling factor uses a very small or a very big value; the EER values for both input string lengths decrease when the scaling factor value increases from 1 to approximately 4, plateau from 4 to 8 and then they start to increase sharply. This result may be explained as follows. When timestamps are normalised by using a smaller or moderate scaling factor, the timing feature data gradually resemble better a typical human tapping speed, and so is the ability to better represent a human's touch pattern. However, when the scaling factor of 9 is used, the timestamp value becomes so small that it conveys little useful timing

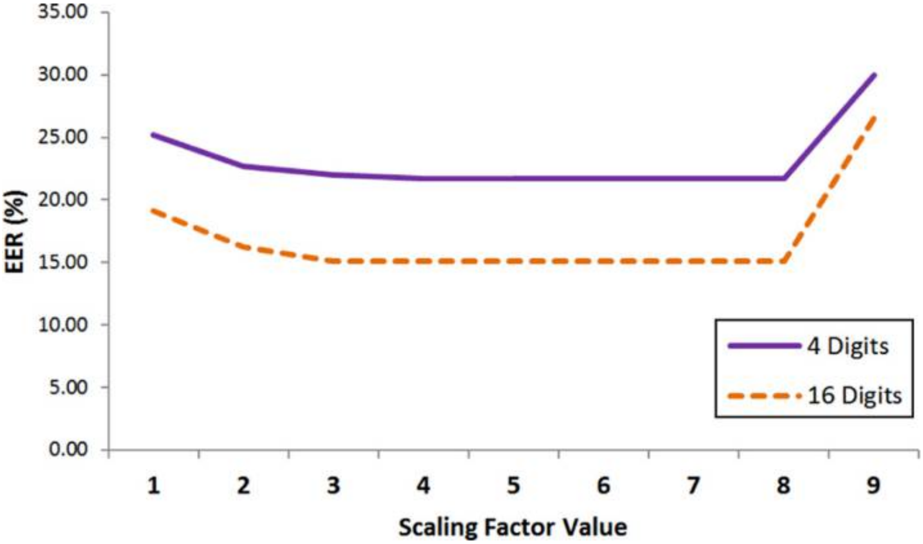

Figure 5.
Accuracy performances of different scaling factor values

information. Therefore, when an extreme scaling factor value is used, the EER value increases sharply. In this experiment, we have chosen to use a moderate scaling factor value of 5.

6.3.2 Single feature data types. Experiments have been conducted to investigate the accuracy performances of different feature data types. As shown in Figure 6, the *ps* feature outperforms all timing-related feature data types for both the 4-digit and 16-digit input strings. This may be due to the fact that *ps* could capture more properties from a subject's touch pattern. For example, the amount of force used, finger arrangement, touch angle and finger thickness. The mixture of these properties establishes a distinctive pattern, which can better capture the uniqueness in each subject's touch pattern.

As for timing-related feature data types, the accuracy performance of any variant of *FT* is slightly better than *DT*. This implies that the time taken for a subject's finger to traverse from one key to another has more discriminative power than how long a key is held down. Throughout the data collection process, we observed that the key combinations entered by different subjects are different even if the input strings entered are identical. This may be due to the fact that breaking up an input string into multiple smaller subsets (i.e. chunks) can make it easier to memorise (Ngugi *et al.*, 2011a). The information embedded within the natural short pauses between different chunks may have increased the uniqueness of *FT*. This is particularly the case for strings with a longer string length, due to the increased number of possible chunk combinations, as shown in Figure 7.

6.3.3 Combining different feature data types. Although *ps* proved to be the best feature data type in terms of EER, its EER performance of 15.98 per cent (the 4-digit string) and 12.44 per cent (the 16-digit string) are still rather unsatisfactory. To improve this performance, we have combined *ps* with different combinations of timing feature data types. So, in a given authentication instance, multiple feature data types are used, and the decision made for the authentication instance is made by combining the decisions made on each chosen feature data type using the AND voting rule. Table II

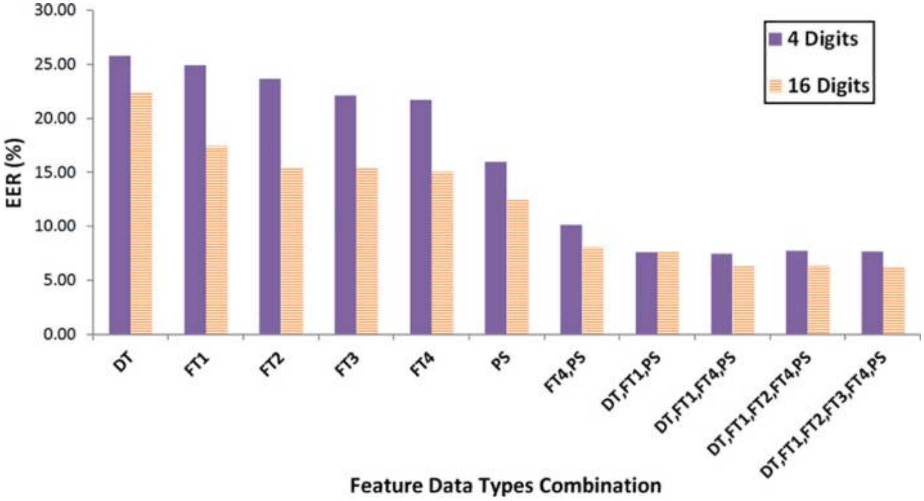

Figure 6. Accuracy performances of single versus different feature data type combinations

shows the best EER values of different feature data type combinations. The values in brackets indicate the accuracy performance gains for using different feature data type combinations against a single feature data type. As can be seen from the table, the more feature data types used, the lower the EER value, the better the accuracy performance. The lowest EER value is achieved when all the six feature data types are used. In this case, the EER values for the 4-digit and 16-digit input strings are, respectively, 7.71 per cent and 6.27 per cent, which are less than half of the respective EER values when *ps* is used alone.

From the results shown in Figure 6, it is clear that combining two or more feature data types can lead to a significantly better accuracy performance than using a single feature data type. This indicates that when more feature data types are used to train the model, the ability to distinguish between a legitimate and an illegitimate user sample also increases.

6.3.4 Input string lengths. Input string lengths may also affect the accuracy performance of a biometrics authentication system. To investigate the effect, we have calculated the EER values using the three matching functions and two sets of PINs with respective string lengths of 4 and 16 digits. The 4-digit set represents a short-input string case, while the 16-digit set represents a long-input string case. The analysis results are plotted in Figure 8. As can be seen from the figure, a longer input string leads to a lower EER value, which indicates a better accuracy performance. This result can be explained as follows. When the input string length increases, feature data samples within each input string, and the number of different chunk combinations also increase, and so is the ability to better capture a subject's touch pattern. In addition, when the input string length increases, the number of illegitimate feature data samples required to match that of a legitimate reference template will also increase. Therefore, the longer the

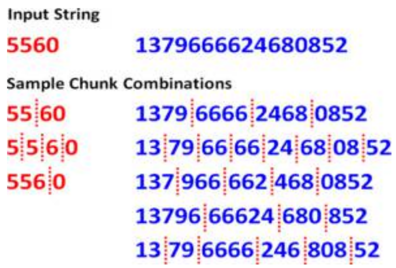

Figure 7.
The different possible chunk combinations of different input string

Table II.
EER values of different feature data type combinations

Feature data types	No. of feature	EER 4-Digit	 16-Digit
PS	1	15.98	12.44
FT_4, PS	2	10.17 (+36.36%)	8.12 (+49.19%)
DT, FT_1, PS	3	7.63 (+52.25%)	7.67 (+52.00%)
DT, FT_1, FT_4, PS	4	7.48 (+53.19%)	6.39 (+60.01%)
DT, FT_1, FT_2, FT_4, PS	5	7.75 (+51.50%)	6.33 (+60.39%)
$DT, FT_1, FT_2, FT_3, FT_4, PS$	6	7.71 (+51.75%)	6.27 (+60.76%)

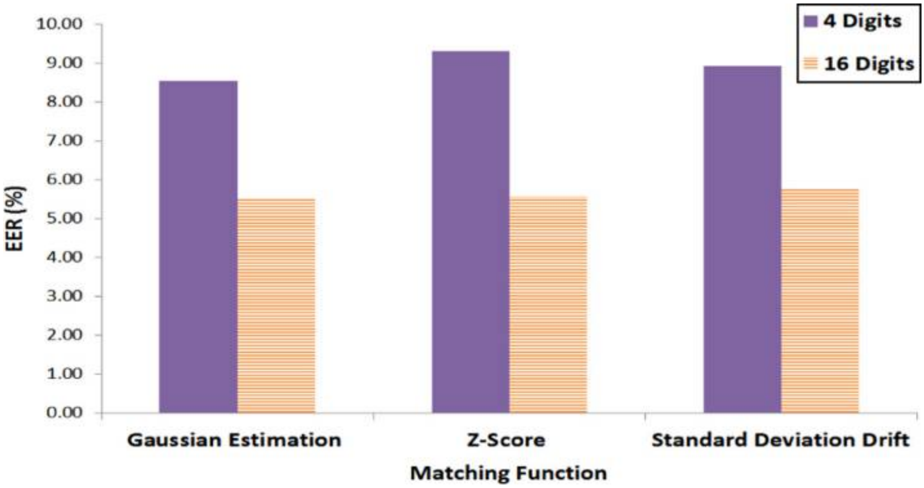

Figure 8. The effects of the input string lengths on the EER values

input strings, the better the accuracy performance one could achieve from a PIN-based biometrics authentication system.

6.3.5 Subject size. We have also investigated the effects of subject sizes on the EER values. This experiment is based on three matching functions: GE, ZS and SD. Two subject sets are used: one set has 50 subjects and the other has 150 subjects. For each subject set, we used two input strings, 4-digit and 16-digit. The experimental results are shown in Table III. Ideally, an input string tested on a larger subject size should achieve a lower or at least a comparable error value than when tested on a smaller subject size. If this case holds, it indicates that the proposed method is scalable at different subject sizes. As shown by the results in the table, when the input string length is 16 digits, the differences in the EER values produced by using different matching functions remain fairly constant when the subject size goes up from 50 to 150. This indicates that the 16-digit input string can provide us with more consistent experimental results. From the table, we can also observe that when the subject size of 50 is used, the average EER value decreases slightly from 7.93 to 6.81, as the input string length increases from 4 digits to 16 digits. However, when the subject size is 150, this average value decreases considerably from 8.92 to 5.59, as the input string length increases from 4 digits to 16 digits. This inconsistency may be due to the fact that an input string with a shorter length is less able to discriminate different touch patterns (as discussed in the previous section).

	4-Digit		16-Digit	
Classifier	50 subjects	150 subjects	50 subjects	150 subjects
GE	7.71	8.55	6.27	5.49
ZS	8.51	9.30	6.70	5.54
SD	7.57	8.92	7.45	5.74
Average EER	7.93	8.92	6.81	5.59

Table III. EER values vs subject sizes

6.3.6 Classifier performance. The EER values produced by all the three functions are shown in Table IV. Two input string lengths, i.e. 4 digits and 16 digits, are used and the subject size chosen is 150. From the results, it can be seen that, generally, the fluctuations in the results produced by the different classifiers are smaller when the input string length is 16 digits in comparison with the 4-digit input string length. For example, when the input string length is 16 digits, the differences in the EER values are less than 0.25 per cent, whereas when the input string length is 4 digits, the differences are at least 0.37 per cent. Among the three functions, the GE produces the lowest EER value, i.e. 8.55 per cent when the input string is 4 digits in length and 5.49 per cent when the input string is 16 digits in length. This says that, even if the input string is known to the impersonator, nine out of ten impersonation attempts can be successfully identified. As the input string length increases, the success rate in identifying impersonation attempts also increases. These results are encouraging. They indicate the potential of using touch dynamics in conjunction with knowledge-based authentication to strengthen the security level of user authentication in mobile devices or service accesses via mobile devices. In addition, touch dynamics biometrics is cost-effective, as it does not require the use of additional hardware, and usable, as it is already part of the mobile device interface. So, it can be an attractive building block for other authentication solutions not only in a physical, but also a virtual environment.

6.3.7 With and without touch dynamics. One potential application area of touch dynamics biometrics is to integrate it into an existing authentication system to extend the system into a so-called multi-factor authentication system. Assuming a two-factor authentication system is used: one factor is PIN-based authentication and the other is touch dynamics authentication. Then an impersonator, to successfully sneak through the authentication verification process, would have to produce an acceptable touch pattern, in addition to successfully passing the PIN verification process. As shown in Table V, in the case where the PIN is exposed, the chances for the impersonator to be successfully authenticated are drastically reduced from 100 per cent (if only a PIN is used) to 12.21 per cent (if both a 4-digit PIN and the touch dynamics are used) or to 9.43 per cent (if both a 16-digit PIN and the touch dynamics are used). To put things into perspective, assuming that there are ten impersonation attempts, the system will only fail to detect once. Conversely, given the same case, the PIN-based authentication will fail to detect any of the impersonation attempts. However, the price to pay for this

Table IV. Performance between classifiers on different input lengths

		4-Digit			16-Digit	
Classifier	FAR	FRR	EER	FAR	FRR	EER
GE	12.21	4.89	8.55	9.43	1.56	5.49
ZS	15.27	3.33	9.30	8.64	2.44	5.54
SD	8.95	8.89	8.92	10.36	1.11	5.74

Table V. The comparison between FAR and FRR with the presence of touch dynamics

Feature	FAR	FRR
PIN	100	0
PIN (4-digit) + touch dynamics	12.21	4.89
PIN (16-digit) + touch dynamics	9.43	1.56

enhanced impersonation detection capability (i.e. enhanced security level) is that by using the touch dynamics-based authentication method, there is an increase of 4.89 per cent (4-digit) or a 1.56 per cent (16-digit) chance to reject a legitimate subject incorrectly. In other words, 1 out of 25 legitimate login attempts may be incorrectly rejected. With the PIN-based authentication method, on the other hand, a login attempt will not be rejected as long as the PIN entered is correct. In summary, with the use of the touch dynamics-based multi-factor authentication method, there is a trade-off between security and usability. Our future effort is to investigate how to enhance or maintain the security level of this method, while minimising the usability cost.

7. Related work

In this section, we provide a literature review of the related works with regards to evaluating the accuracy performance of touch dynamics biometrics and touch dynamics biometric data collections. We also highlight open issues and opportunities in the topic area.

7.1 Performance investigation

Our related work review only focuses on touch dynamics biometrics using PIN-based input strings. The experimental work reported in the paper by Sen and Muralidharan (2014) was carried out to test the viability of identifying subjects based on touch dynamics biometrics using numerical input strings. In this experiment, only ten subjects were recruited and each subject was asked to input a predefine PIN ("1593") on a HTC Nexus-One smartphone. To investigate if an impersonator could imitate another subject's touch pattern in the event if the subject's PIN is known to the impersonator, the author designed a visualisation tool to facilitate a separate set of attackers to imitate a legitimate subject's input pattern. Even by deliberately exposing the PIN, the timing and pressure feature information via the visualisation tool to the attackers, the method was still able to achieve a FAR value of 16 per cent. Though some interesting results were obtained from this experiment, the number of subjects used was too small to draw any conclusive remark. This is also the case for the work conducted by several other works (Amin *et al.*, 2015; Buriro *et al.*, 2015; Li *et al.*, 2015).

The accuracy performance of touch dynamics applied on 4-digit PIN was also investigated by de Mendizabal-Vazquez *et al.* (2014). They extracted data with regard to timing, finger touch size and pressure by using device build-in touchscreen sensors, and in addition, they also extracted linear and angular acceleration as feature vectors using accelerometer and gyroscope sensors. As the size of data collected from these two sensors is large, they applied a pre-processing technique to reduce the size of the data. As a result, the computational resource required for data classification is reduced. However, the data set in this experiment was collected in a quite constrained setting, where subjects had to hold the mobile phone in a fixed position. By using a Euclidean distance-based classifier, they obtained a performance of 20 per cent EER on a 4-digit PIN input. The performance comparisons of individual features were not given. Also, the use of the additional sensors (i.e. accelerometers and gyroscopes) to extract linear and angular acceleration feature data means that this technique may not be applicable to the mobile devices that are not equipped with these sensors.

Zheng *et al.* (2014) also conducted their experiment on several 4- and 8-digit PINs. The authors used a statistical one-class learning classifier and obtained average EER values of 3.65, 6.96 and 7.34 per cent using three different 4-digit PIN numbers, “3244”, “1111” and “5555”, respectively. This shows that a higher repetition of digits can reduce accuracy performance. The experiment also compared the accuracy performances of two different 8-digit PINs (i.e. “12597384” and “12598416”). Surprisingly, the EER value of one of the 8-digit PINs (“12598416”) was 4.45 per cent, marginally worse than the 4-digit PIN (“3244”). This contradicts with our experimental observation that a longer input string can produce a better accuracy performance. However, similar to our experimental observation, the accuracy performance provided by a combination of multiple feature data types outperforms that provided by a single feature data type.

The experiment reported by Tasia *et al.* (2014) was carried out on a larger number of subjects than the experiments described above, and it used input PINs ranging from 4 to 8 digits in length. An EER of 8.4 per cent was achieved by using a simple statistical classifier. In this paper, the authors have also studied the time taken to perform the classification and verification of different PIN lengths and feature combinations; both consumed an average of 12 ms each. This experimental result is useful when considering the deployment of touch dynamics biometrics on power-limited mobile devices.

The work reported by Chang *et al.* (2015) was rather unique. The author proposed a method to allow subjects to change their PINs without rebuilding the classification model. The subjects were asked to input ten different randomly selected 10-digits PINs. To reduce work burdens on the subjects, each subject was only required to provide five samples for each PIN. Based on all the samples collected, the authors produced a table of all possible feature data type values for each digit. They achieved EER values of 23, 21 and 18 per cent on three different PINs with the string lengths of 6, 8 and 10, respectively. This set of results is consistent with our experimental observation that, when the 16-digit PIN is used, a lower EER value can be achieved than using a 4-digit PIN. However, as the subjects involved in this experiment were at the age 17-20 years, it is not clear whether the experimental results are applicable to other age groups.

Most of the experiments reported in literature extracted timing data from the two shortest possible feature lengths (i.e. uni-graph and di-graph). An exception of these is the work carried out by Trojahn *et al.* (2013). In this work, the authors compared the accuracy performances of timing features produced from three different feature lengths (i.e. the uni-graph, di-graph and tri-graph) of a 17-digit PIN. The experimental result suggested that the uni-graph produces a better accuracy performance than the tri-graph. This may be due to the fact that the timing feature data expressed by a shorter feature length contains a higher level of granularity than a longer feature length. The authors also remarked that, by combining the results from the uni-graph and di-graph, they could achieve a lower error rate. This is similar to the observations we made in our experiment that combining multiple feature data types produces a better accuracy performance than using a single feature data type. The similar observation has also been reported in other works, such as Tasia *et al.* (2014) and Zheng *et al.* (2014).

By far, the best accuracy performance reported in literature was achieved by Wu and Chen (2015). They achieved an average EER value of 0.56 per cent. The work trained a classification model by using the support vector machine and both legitimate and illegitimate subject samples were used. In the experiment, each subject provided four different 8-digits PINs. Two main discoveries were made in their experiments. Firstly, a PIN with more repetitiveness (e.g. “11111111”) resulted in a lower accuracy performance than a PIN with less repetitiveness (e.g. “16843752”). Secondly, the accuracy performance can be increased by increasing the size of legitimate subject samples, by using a combination of different feature data types and by pre-processing feature data in terms of data normalisation and outlier removal. In their experiment, they imposed a condition that the experimental subjects were selected from those who were very familiar with touchscreen smartphones. This condition might have played a role in achieving the high accuracy performance. It is not clear if the same level of accuracy performance could still be achieved if this condition is removed. Also, in their experiment, a two-class classifier was used to build the classification model. In other words, to build the classification model, samples from both legitimate and illegitimate subjects are required. However, in real life, as mobile devices are very much personal devices, illegitimate subject samples may not be always available.

There were also experiments (Dhage *et al.*, 2015; Giuffrida *et al.*, 2014; Huang *et al.*, 2012; Kambourakis *et al.*, 2014) carried out on character-based passwords. As the scope of our work is on numerical PIN inputs, we do not discuss character-based experiments any further. A summary of the comparisons between our work and the related works is given in Table VI.

7.2 Public data sets

As touchscreen devices only came along not long ago, there are still limited benchmark data sets publically available; so far, we are only able to find three such data sets, but only one of the data sets uses PIN-based input and none of these is collected on a widescreen digital tablet.

The data collection process conducted by El-Abed *et al.* (2014) involved 51 subjects. The input is a fixed password “rhu.university” entered on a virtual keyboard of a window touchscreen phone (Nokia Lumia 920). Each subject attended three sessions with an average of five days apart. The first of the three sessions was used as a practice

Paper	Length	Subjects	Input	Device	FAR	FRR	EER
Trojahn *et al.* (2013))	17	152	Fixed	Samsung Galaxy Nexus	4.19	4.59	–
Sen and Muralidharan (2014)	4	10	Fixed	HTC Nexus-One	14.1	14.1	15.2
Zheng *et al.* (2014)	4	80	Fixed	Samsung Galaxy Nexus	–	–	3.65
	8				–	–	4.45
de Mendizabal-Vazquez *et al.* (2014)	4	80	Fixed	–	–	–	20
Tasia *et al.* (2014)	4-10	100	Random	Motorola Milestone	8.4	8.32	8.4
Wu and Chen (2015)	8	100	Fixed	–	–	–	0.56
Buriro *et al.* (2015)	4	12	Fixed	Google Nexus 5	1	1	–
Li *et al.* (2015)	4	15	Fixed	HTC Droid DNA	–	–	4.2
Amin *et al.* (2015)	7	12	Fixed	Samsung Galaxy Note N7000	13.9	0.53	–
Chang *et al.* (2015)	6	100	Random	HTC Desire Z	–	–	23
	8				–	–	21
	10				–	–	18
This paper	4	150	Fixed	Samsung Galaxy Tab 10.1	12.21	4.89	8.55
	16				9.43	1.56	5.49

Table VI. Comparison to existing touch dynamics work carried out on PIN input

session, so the actual data collection started from the second session. A total of 15 samples were collected from each subject in the two sessions. Only the timing feature was captured from this data set, whereas our data set also captures the finger touch size and pressure feature.

Another related effort on collecting and sharing their data sets publically was made by Antal and Szabó (2014). This work differs from ours in a number of ways. Firstly, the number of subjects involved is more than three times smaller than ours. Also the entire subject population recruited in this work was students, which is different from our case where the population consists of the members of a university as well as general public. In addition, different from our case where all the data were collected via the use of a single device, the data collection in this work was done via the use of two types of devices. In all, 37 subjects provided their inputs on a Nexus 7 device, while the remaining five via the use of a LG Optimus L7 P700 smartphone. The paper did not explain if the use of the two device types would have any performance implications. To allow the sample data to be used in EER estimations, the input string was predefined (".tie5Roanl"), which is also the case in our data collection. Also, in this work, the touch events captured include not only the input string but also *shift key* (toggle between lower and uppercase characters) and *keyboard switch key* (toggle between characters and numerical keys). These secondary key events may capture valuable and distinctive information about a subject. Inspired by this idea, in addition to capturing touch events on digit input, we have also recorded the *Enter key* event (pressed upon the completion of a PIN input) in our data set. Also in this related work, most of the subjects provided their passwords for 30 times each on two isolated sessions in a period of two weeks (the time duration between the two sessions is not mentioned in the paper). As some invalid inputs were removed, so the resulting data set only contains 51 input samples per subject (instead of 60 from both sessions).

Tasia *et al.* (2014) have also reported a collection of a data set based on numerical inputs. In this data collection process, the device used was an early-generation smartphone with a physical keypad running on Android 2.0.1 (Éclair) API level 6, which was released in December 2009. By contrast, we adopted a more recent high-resolution digital tablet with a later version of a mobile operating system. The subjects were only required to provide two samples per session and five sessions were used with an interval of at least one week apart for each session to eliminate intra-session typing variations. Different from our case where data were collected in a non-restrictive environment, their data were collected in a classroom, a rather confined environment. The age distribution of the subjects involved is biased towards young people, where 85% of the total subjects are at the age of 25 or younger. This is different from our case, where our data set was collected from a diversified population and with different age groups and backgrounds. Also in this data set, subjects were allowed to freely choose a PIN, and most of the chosen PINs have a length of 4 to 8 digits. However, the actual PIN selections by each subject were not recorded in the shared data set. In this data set, raw finger touch sizes and pressure data have been recorded, and the timing feature was only recorded in a post-processed format (the duration and the latency). In other words, raw timing values were not recorded. This missing information may hinder the usability of the data set in a wider context. Finally, different from the usual practice, test samples collected for the FAR test were collected separately from those for FRR test. Ten subjects are randomly chosen to act as impersonators. These impersonators were given the PINs of every other subject and were asked to impersonate the subjects by providing five samples for each subject. In this way, the samples used for the FRR tests cannot be reused for impersonation test samples, thus significantly reducing the size of the samples available for the FAR tests.

The creation and collection of live data is a time- and resource-consuming process, and this may be the reason for the lack of open data sets (Giot *et al.*, 2009). However, the research, design and performance evaluation of touch dynamic biometrics systems require the availability of such benchmark data sets. To overcome these restrictions, we have collected touch dynamics data from 150 subjects. The data set is made available to download at https://goo.gl/sNACU8.

An overview of presently available public data sets is summarised in Table VII, where T, S and P indicate timing, finger touch size and pressure values, respectively.

7.3 Open issues and opportunities

This section outlines open issues and potential research opportunities we have identified in doing our research:

- *Optimal input length*: In general, the longer the input length, the better the accuracy performance we may achieve. However, the use of a longer input length increases the difficulty of remembering it by a subject, thus reducing usability. Further research work is necessary to study the trade-off between the accuracy performance and the usability.
- *Feature data enrichment*: Exiting efforts on increasing the accuracy performance of touch dynamics biometrics have largely been focusing on using different classification techniques. An alternative way to increase the accuracy performance is by increasing the quality or variety of feature data. This can be done by deriving different types of feature from raw sensor data.
- *Effective classification technique*: As the primary input for a classification technique in this problem context is subjects' touch pattern samples, we may also categorise samples based on the types of the samples used and determine the type of classifier that we should use. For example, a one-class classifier (e.g. distance measure) is modelled or trained by only using legitimate samples, and a two-class classifier (e.g. neural network) is modelled or trained by using both legitimate and illegitimate samples. A mobile device is a highly personal device (rarely shared between multiple users), obtaining illegitimate samples in practice is not easy. So in the mobile device context, the use of a one-class classifier may be a more viable

Dataset	Subject	Population	Sample	Input	Feature	Setting	Platform
El-Abed *et al.* (2014)	51	Restricted	15	"rhu.university"	T	Confined	Phone
Antal and Szabó (2014)	42	Restricted	51	".tie5Roanl"	T, S, P	Confined	Phone
Tasia *et al.* (2014)	100	Restricted	5	6 to 8 digits	T, S, P	Confined	Phone
This paper	150	Diversified	10	"5560", "1379666624680852"	T, S, P	Flexible	Tablet

Table VII.
Comparison of public data sets

option. Further research work is necessary to study the accuracy performance difference between a one-class and two-class classifier.

- *Adaptive learning*: Human touch patterns are subject to change over time. To accommodate this change, a pattern adaptation method should be implemented in the underlying authentication system to update existing reference templates or classifier to reflect the most recent touch pattern of a subject. More work is necessary to evaluate the effectiveness of the adaptation approach.
- *Standardise evaluation criteria*: There is an inconsistency in the performance evaluation metrics (as shown in Table VI) used by different researchers, and this has made it difficult to compare the accuracy performances produced using different methodologies and/or by different researchers. To facilitate effective comparisons of the related works, all the three standard accuracy performance evaluation metrics (i.e. FAR, FRR and EER) should be used.

8. Conclusion

Touch dynamics-based authentication may provide us with a number of benefits, such as it is an inherent feature of a majority of the mobile devices already in use and it is readily deployable as an additional authentication factor to strengthen e-authentication assurance levels. This paper has investigated the feasibility and benefits of adopting a touch dynamics-based authentication method by integrating it with the PIN-based authentication method. To evaluate the effectiveness of this integrated approach, a proper data set is required. With this motivation, we have reported how a comprehensive data set is collected. The data set can also serve further research on various issues in this context, such as further investigation and comparison of different classification methods or the potential use of touch dynamics for authentication purposes. We have then illustrated the extraction of different feature data from the data set and how the captured feature data could be used for authentication to identify a user. We have also applied three light-weight matching functions to the data set to study their accuracy performances. The matching functions used in this experiment were chosen on the ground that they are computationally less expensive than the other functions, so the resulting authentication system could consume less power and introduce less delay in authenticating a user. We have also investigated how the accuracy performance may be influenced by variations in factors such as the timing resolution of timing feature data, combinations of different feature data types, input string lengths and subject sizes. Experimental results show that, with the use of the two-factor authentication method, even if an impersonator knows the input string (i.e. PIN) of a legitimate subject, nine out of ten impersonation attempts can be successfully identified. We also showed that the accuracy performance can be increased by combining different feature data types. The results we have obtained so far demonstrate that touch dynamics biometrics can be an effective solution to strengthen the security level offered to mobile devices.

References

Amin, R., Gaber, T. and ElTaweel, G. (2015), "Implicit authentication system for smartphones users based on touch data", in Abraham, A., Jiang, X.H., Snášel, V. and Pan, J.-S. (Eds), *Intelligent Data Analysis and Applications*, Springer International Publishing, pp. 251-262.

Antal, M. and Szabó, L.Z. (2014), "Keystroke dynamics on android platform", *Proceedings of the 8th International Conference Interdisciplinarity in Engineering, Romania*, pp. 131-136.

Bleha, S., Slivinsky, C. and Hussien, B. (1990), "Computer-access security systems using keystroke dynamics", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12 No. 12, pp. 1217-1222.

Bo, C., Zhang, L., Jung, T., Han, J., Li, X.-Y. and Wang, Y. (2014), "Continuous user identification via touch and movement behavioral biometrics", presented at the 2014 IEEE International Performance Computing and Communications Conference (IPCCC), pp. 1-8.

Buriro, A., Crispo, B., Frari, F.D. and Wrona, K. (2015), "Touchstroke: smartphone user authentication based on touch-typing biometrics", in Murino, V., Puppo, E., Sona, D., Cristani, M. and Sansone, C. (Eds), *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, Springer International Publishing, pp. 27-34.

Buschek, D., De Luca, A. and Alt, F. (2015), "Improving accuracy, applicability and usability of keystroke biometrics on mobile touchscreen devices", *To Appear: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Campisi, P., Maiorana, E., Lo Bosco, M. and Neri, A. (2009), "User authentication using keystroke dynamics for cellular phones", *Iet Signal Processing*, Vol. 3 No. 4, pp. 333-341.

Chang, T.-Y., Tsai, C.-J., Tsai, W.-J., Peng, C.-C. and Wu, H.-S. (2015), "A changeable personal identification number-based keystroke dynamics authentication system on smart phones", *Security and Communication Networks*.

Cho, S., Han, C., Han, D.H. and Kim, H.-I. (2000), "Web-based keystroke dynamics identity verification using neural network", *Journal of Organizational Computing and Electronic Commerce*, Vol. 10 No. 4, pp. 295-307.

Chong, M.K., Marsden, G. and Gellersen, H. (2010), "GesturePIN: using discrete gestures for associating mobile devices", *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services, ACM, New York, NY*, pp. 261-264.

Cisco (2015), "Cisco visual networking index: global mobile data traffic forecast update 2014-2019", *Cisco*, 3 February, available at: www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html (accessed 14 April 2012).

Clarke, N.L. and Furnell, S.M. (2007), "Authenticating mobile phone users using keystroke analysis", *International Journal of Information Security*, Vol. 6 No. 1, pp. 1-14.

Crawford, H., Renaud, K. and Storer, T. (2013), "A framework for continuous, transparent mobile device authentication", *Computers & Security*, Vol. 39, pp. 127-136.

De Luca, A., Hang, A., Brudy, F., Lindner, C. and Hussmann, H. (2012), "Touch me once and i know it's you! Implicit authentication based on touch screen patterns", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY*, pp. 987-996.

de Mendizabal-Vazquez, I., de Santos-Sierra, D., Guerra-Casanova, J. and Sanchez-Avila, C. (2014), "Supervised classification methods applied to keystroke dynamics through mobile devices", presented at the 2014 International Carnahan Conference on Security Technology (ICCST), pp. 1-6.

Dhage, S., Kundra, P., Kanchan, A. and Kap, P. (2015), "Mobile authentication using keystroke dynamics", presented at the 2015 International Conference on Communication, Information Computing Technology (ICCICT), pp. 1-5.

Draffin, B., Zhu, J. and Zhang, J. (2014), "KeySens: passive user authentication through micro-behavior modeling of soft keyboard interaction", in Memmi, G. and Blanke, U. (Eds), *Mobile Computing, Applications, and Services*, Springer International Publishing, pp. 184-201.

El-Abed, M., Dafer, M. and El Khayat, R. (2014), "RHU keystroke: a mobile-based benchmark for keystroke dynamics systems", presented at the 2014 International Carnahan Conference on Security Technology (ICCST), pp. 1-4.

Gaines, R.S., Lisowski, W., Press, S.J. and Shapiro, N. (1980), "Authentication by keystroke timing: some preliminary results", No. R-2526-NSF, Rand Corporation, Santa Monica, CA.

Gascon, H., Uellenbeck, S., Wolf, C. and Rieck, K. (2014), "Continuous authentication on mobile devices by analysis of typing motion behavior", *Lecture Notes in Informatics*, Vol. 228, pp. 1-12.

Giot, R., El-Abed, M. and Rosenberger, C. (2009), "GREYC keystroke: a benchmark for keystroke dynamics biometric systems", *IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems BTAS '09*, pp. 1-6.

Giuffrida, C., Majdanik, K., Conti, M. and Bos, H. (2014), "I sensed it was you: authenticating mobile users with sensor-enhanced keystroke dynamics", in Dietrich, S. (Ed.), *Detection of Intrusions and Malware, and Vulnerability Assessment*, Springer International Publishing, pp. 92-111.

Huang, X., Lund, G. and Sapeluk, A. (2012), "Development of a typing behaviour recognition mechanism on android", presented at the 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 1342-1347.

Kambourakis, G., Damopoulos, D., Papamartzivanos, D. and Pavlidakis, E. (2014), "Introducing Touchstroke: keystroke-based authentication system for smartphones", *Security and Communication Networks*, available at: http://doi.org/10.1002/sec.1061

Kim, D., Dunphy, P., Briggs, P., Hook, J., Nicholson, J.W., Nicholson, J. and Olivier, P. (2010), "Multi-touch authentication on tabletops", *Proceedings of the 28th International Conference on Human Factors in Computing Systems, Atlanta, GA, ACM, New York, NY*, pp. 1093-1102.

Kolly, S.M., Wattenhofer, R. and Welten, S. (2012), "A personal touch: recognizing users based on touch screen behavior", *Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones, ACM, New York, NY*, pp. 1:1-1:5.

Li, Y., Yang, J., Xie, M., Carlson, D., Jang, H.G. and Bian, J. (2015), "Comparison of PIN- and pattern-based behavioral biometric authentication on mobile devices", presented at the MILCOM 2015 - 2015 IEEE Military Communications Conference, pp. 1317-1322.

Loy, C.C., Lai, W.K. and Lim, C.P. (2007), "Keystroke patterns classification using the ARTMAP-FD neural network", *IIHMSP 2007 Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Vol. 1, pp. 61-64.

Maiorana, E., Campisi, P., González-Carballo, N. and Neri, A. (2011), "Keystroke dynamics authentication for mobile phones", *Proceedings of the 2011 ACM Symposium on Applied Computing, ACM, New York, NY*, pp. 21-26.

Ngugi, B., Kahn, B.K. and Tremaine, M. (2011a), "Typing biometrics: impact of human learning on performance quality", *Journal of Data and Information Quality*, Vol. 2 No. 2, pp. 11:1-11:21.

Ngugi, B., Tremaine, M. and Tarasewich, P. (2011b), "Biometric keypads: improving accuracy through optimal PIN selection", *Decision Support Systems*, Vol. 50 No. 4, pp. 769-776.

Niu, Y. and Chen, H. (2012), "Gesture authentication with touch input for mobile devices", in Prasad, R., Farkas, K., Schmidt, A.U., Lioy, A., Russello, G. and Luccio, F.L. (Eds), *Security and Privacy in Mobile Information and Communication Systems*, Springer Berlin Heidelberg, pp. 13-24.

Obaidat, M.S. (1995), "A verification methodology for computer systems users", *Proceedings of the 1995 ACM Symposium on Applied Computing, ACM, Nashville, TN*, pp. 258-262.

Robinson, J.A., Liang, V.M., Chambers, J.A.M. and MacKenzie, C.L. (1998), "Computer user verification using login string keystroke dynamics", *IEEE Transactions on Systems Man and Cybernetics Part a-Systems and Humans*, Vol. 28 No. 2, pp. 236-241.

Saravanan, P., Clarke, S., Chau, D.H. (Polo) and Zha, H. (2014), "LatentGesture: active user authentication through background touch analysis", *Proceedings of the Second International Symposium of Chinese CHI*, *ACM, New York, NY*, pp. 110-113.

Sen, S. and Muralidharan, K. (2014), "Putting 'pressure' on mobile authentication", presented at the 2014 Seventh International Conference on Mobile Computing and Ubiquitous Networking (ICMU), pp. 56-61.

Serwadda, A., Phoha, V.V. and Wang, Z. (2013), "Which verifiers work? A benchmark evaluation of touch-based authentication algorithms", presented at the 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1-8.

Stewart, J.C., Monaco, J.V., Cha, S.-H. and Tappert, C.C. (2011), "An investigation of keystroke and stylometry traits for authenticating online test takers", *2011 International Joint Conference on Biometrics (IJCB)*, pp. 1-7.

Tasia, C.-J., Chang, T.-Y., Cheng, P.-C. and Lin, J.-H. (2014), "Two novel biometric features in keystroke dynamics authentication systems for touch screen devices", *Security and Communication Networks*, Vol. 7 No. 4, pp. 750-758.

Teh, P.S., Teoh, A.B.J. and Yue, S. (2013), "A survey of keystroke dynamics biometrics", *The Scientific World Journal*.

Trojahn, M., Arndt, F. and Ortmeier, F. (2013), "Authentication with keystroke dynamics on touchscreen keypads – effect of different N-graph combinations", presented at the MOBILITY 2013, The Third International Conference on Mobile Services, *Resources, and Users*, pp. 114-119.

Wu, J. and Chen, Z. (2015), "An implicit identity authentication system considering changes of gesture based on keystroke behaviors", *International Journal of Distributed Sensor Networks*.

Xu, H., Zhou, Y. and Lyu, M.R. (2014), "Towards continuous and passive authentication via touch biometrics: an experimental study on smartphones", *Symposium On Usable Privacy and Security (SOUPS 2014*), USENIX Association, Menlo Park, CA, pp. 187-198.

Zheng, N., Bai, K., Huang, H. and Wang, H. (2014), "You are how you touch: user verification on smartphones via tapping behaviors", presented at the 2014 IEEE 22nd International Conference on Network Protocols (ICNP), pp. 221-232.

Corresponding author

Pin Shen Teh can be contacted at: pinshen.teh@manchester.ac.uk