Touch me once and I know it's you! Implicit Authentication based on Touch Screen Patterns

Alexander De Luca, Alina Hang, Frederik Brudy, Christian Lindner, Heinrich Hussmann
Media Informatics Group, University of Munich
Amalienstr. 17, 80333 Munich, Germany
{alexander.de.luca, alina.hang, hussmann}@ifi.lmu.de, {brudy,lindnerch}@cip.ifi.lmu.de

ABSTRACT
Password patterns, as used on current Android phones, and other shape-based authentication schemes are highly usable and memorable. In terms of security, they are rather weak since the shapes are easy to steal and reproduce. In this work, we introduce an implicit authentication approach that enhances password patterns with an additional security layer, transparent to the user. In short, users are not only authenticated by the shape they input but also by the way they perform the input. We conducted two consecutive studies, a lab and a long-term study, using Android applications to collect and log data from user input on a touch screen of standard commercial smartphones. Analyses using dynamic time warping (DTW) provided first proof that it is actually possible to distinguish different users and use this information to increase security of the input while keeping the convenience for the user high.

Author Keywords
Security; implicit authentication; password pattern

ACM Classification Keywords
H5.2 [Information Interfaces and Presentation]: User Interfaces – Input devices and strategies, evaluation;

General Terms
Experimentation, Human Factors, Measurement.

INTRODUCTION

With the introduction of the Android operating system for mobile phones, an alternative to PIN-authentication on mobile devices was introduced and widely deployed for the first time. The password pattern, similar to shape-based authentication approaches like Draw-a-secret [18] or PassShapes [36], enables user authentication by drawing a shape on the screen. The shape consists of an arbitrary number of strokes (or lines) between nine dots as shown in figure 1.

CHI'12, May 5–10, 2012, Austin, Texas, USA.

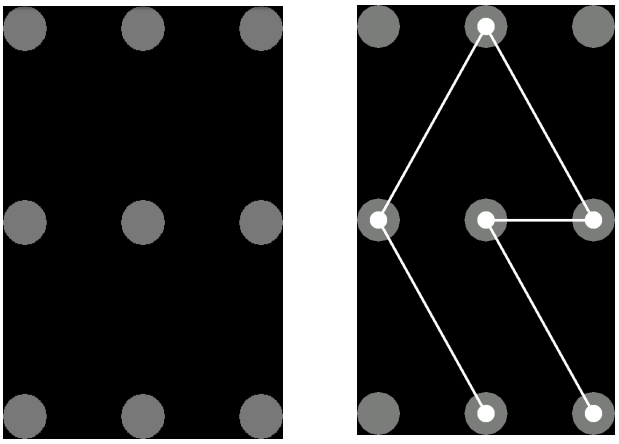


Figure 1: Standard layout of the password pattern authentication system. The blank screen (left) and an exemplary shape (right).

In a study by Clarke et al. [8], 41% of their respondents expressed concerns with respect to PINs and alphanumeric passwords, supporting the need for alternative authentication techniques. In comparison to these approaches, shape-based authentication better supports the way the brain remembers and stores information. The shape can be remembered as an image, therefore exploiting the pictorial superiority effect [25,33]. Additionally, since the pattern is drawn manually in exactly the same way every time and repeated regularly, the user's motor memory [13,30] further improves the memorability. This effect was shown to be effective [12,36], even when the shapes are performed by the user's gaze [11].

Despite its manifold advantages, this approach has major drawbacks, the most important one being security. Drawn passwords are very easy to spy on [11,36], which makes shoulder surfing, a common attack in public settings [27], a serious threat. Other attacks include the infamous smudge attack [1], in which finger traces left on the screen are used to extract the password. Due to its weak security properties, this authentication approach does not fully meet the requirement of adequately protecting the user's data stored on the device. Nowadays, not only private but also valuable business information is stored on the user's handheld [20]. Therefore, resistance to attacks is a major concern when designing respective authentication systems.

In this work, we extended the password pattern approach with an implicit authentication layer to improve its security. Whenever a user authenticates, the system not only checks if the shape was correct but also how it was entered, to determine if the person should gain access. To achieve this, we used touch screen data of current smartphones (pressure, coordinates, size, speed, time etc.) to distinguish between the rightful user and an attacker.

We performed two consecutive user studies to verify the viability of this approach. In the first study, we examined simple unlocks (e.g. a horizontal stroke) while the second study used password patterns both times enhanced with implicit authentication. The results show that it is possible to distinguish different users, thus increasing security during the authentication process. At the same time, the complexity of the approach is hidden since the users only interact with the familiar password pattern system. Thus, the main contribution of this work is to provide first proof that this implicit authentication approach actually works.

BIOMETRIC AUTHENTICATION

Apart from "something you know" authentication schemes (e.g. [11,12,21]), biometrics is an often-used alternative [9]. According to Wood [35] there are two types of biometric authentication approaches, physiological and behavioral biometrics. Physiological biometrics relies on "something the users are". In [32], Sonkamble et al. present an overview of different possible features, including the users' fingerprint, face, hand geometry, voice or iris as physiological biometrics. In [2,3], multiple biometric features are combined to implement a person identification system. Biometric authentication systems on mobile devices were, for instance, used by Rokita et al. [28], using face and hand features. Marcel et al. [24] implemented a mobile authentication system based on simultaneous face and voice recognition, using built-in sensors of the mobile device. The advantage of physiological approaches is that they work instantaneous. In general, however, they require additional hardware (e.g. fingerprint scanners). In addition, there are also user concerns related to the storage of physiological features [26].

Behavioral biometrics, on the other hand, is more commonly used for continuous authentication. As the term behavioral implies, these approaches are based on the users' behavioral cues and authentication may happen implicitly. Exemplary cues are the user's gait [10,16], location information [14] or keystroke patterns [5,22]. Shi et al. [17,31] also proposed the use of behavioral biometrics as replacement for password-based authentication or as second level of authentication. Their authentication system is based on multiple cues such as location information or communication. These features are combined with cloud computing to reduce the energy consumption on the mobile device [6]. In [23] acceleration signals are used for user identification, whereas [4,7] used typing patterns to authenticate users on mobile devices using static classifiers and neural network classifiers, respectively. Tanviruzzaman et al. [34] developed a mobile system called ePet, which uses the user's gait and location information as behavioral cues. ePet continuously checks against anomalous user behavior and denies further access on the mobile device in case anomalies are detected. Continuous authentication on mobile devices was also introduced by Yazji et al. [37]. Their system observes activities on the mobile file system as well as its network access. Due to the permanent re-authentication, their system has a latency of five minutes, with an accuracy of 90 %. Continuous authentication is always bound to latency, leaving the system unsecured for a certain amount of time, when no additional means of authentication are taken.

The approach presented in this work employs behavioral biometrics, the way a user performs the password pattern, but immediately authenticates the user. Using common touch screen data makes the need for any additional hardware obsolete. Furthermore, we combine behavioral biometrics with the input of graphical passwords.

THREAT MODEL

We assume an attacker that is already in possession of the user's password pattern (the shape). That is, the first security barrier has already been breached. How the attacker got this information is of no concern for this work. In addition, the attacker managed to retrieve the mobile device (e.g. using pickpocketing) and wants to gain access to valuable information on it. For this, as for other commercial systems, the attacker has three tries until the device will be blocked.

The approach presented in this work relies on implicit authentication and has been designed to provide security against such an attack. Thus, even after losing the mobile device and the authentication credential, the proposed system should still provide the required security.

UNLOCK USER STUDY

The main idea of the pilot user study was to collect as much data as possible using simple unlocks as known from smartphones like the iPhone or Android devices. Our objective was to gain first insights into the possibilities of identifying and distinguishing users based on the data collected with a capacitive touch screen.

We developed an Android application that was used for data collection. Figure 2 shows four different unlock screens that we implemented. Two of them (horizontal and vertical) were based on unlocks from existing devices. The remaining two were newly developed for this study to add unlocks that would produce more data. The application logged all data available from the touch screen: pressure (how hard the finger presses), size (area of the finger touching the screen), X-coordinate, Y-coordinate and time. The only exception to this was the vertical unlock with two fingers, which had two sets of XY-coordinates, pressure and size, one for each finger. Depending on the workload of the device, one event with all of the previously mentioned data was collected every nine to twelve milliseconds. No other sensor data was logged.

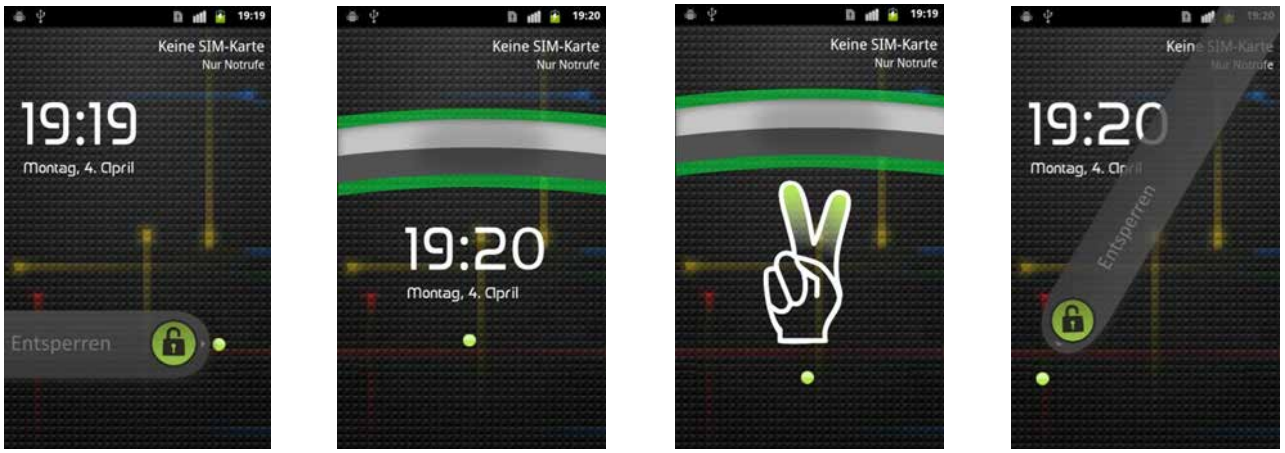


Figure 2: The four different unlock screens as used in the first user study. From left to right: Horizontal, vertical, vertical with two fingers and diagonal unlock.

User Study Design

As study design, a repeated measures within participants factorial design was used. The independent variable was *unlockScreen* with four levels (horizontal, vertical, vertical with two fingers, diagonal).

The task was to unlock the device 160 times with each of the four different unlock screens over a period of two days. The order of the unlock screens was counterbalanced to minimize learning effects.

Procedure

Before the experiment started, the study was explained in detail to each participant. After that, they were asked to fill out an initial questionnaire mainly collecting demographics. During this process, each participant was assigned an ID to (a) allow for anonymous data collection and (b) determine the order in which the unlock screens would be presented to the user.

Two Android Nexus One mobile devices were put on the table in front of the participant. One device had the study application installed while the second was used for a distraction task. During the test, each participant was asked to unlock the screens 80 times with each setup. At first, each user got the application device in test mode and could play around with the current unlock screens until they felt familiar with it. They were also instructed to perform the task in the same way for each unlock using the same finger(s). Every 20 unlocks, the application stopped and the participants were asked to perform a distraction task. For this purpose, they had to input a text message on the second device. When resuming the unlock task, they were reminded by the experimenter which finger(s) they used in case they did not remember.

Within two days after the experiment, the participants were asked to come to the lab again and performed the exact same experiment a second time. This was done to collect more realistic (and less biased) data by minimizing habituation effects and to see how the performance would change over time. Each unlock was additionally used as an attack for the other users. The second day ended with a final questionnaire, collecting opinions about the different unlock screens. As incentives, 8$ gift vouchers were given to the participants.

Participants

We recruited 48 participants for the study with an average age of 25 years. The youngest participant was 18, the oldest 59 years old. 22 were female, 26 male. The majority of the participants were students (73%), the remaining ones came from different professions. Each participant owned at least one mobile phone (65% smartphones) at the time the study was conducted. 17% of the participants were left-handed.

Having 48 participants allowed for perfectly counterbalancing the four unlock screens (4! = 24). That is, each permutation was performed by exactly two users.

Data Analysis

For the analysis of the data, dynamic time warping (DTW) was used. This algorithm originates from speech recognition [29] and allows for comparing two sets (time sequences) of data with each other. The algorithm looks for similarities between the sets and calculates the costs to match one onto the other. The result is a warp distance that can be used to determine how similar a set is to the reference set. A warp distance of 0 (zero) indicates absolute identical sets. The bigger the distance, the more different the sets are. Thus, it is highly appropriate for the purpose of this work, for which we used the DTW implementation for R by Toni Giorgino [15].

In this work, a sequence consists of a time series of touch screen data (all combinations of X-coordinate(s), Y-coordinate(s), pressure, size, time). Again, the exception was the two fingers vertical unlock, which had two sets of XY-coordinates, two pressure and two size values.

It has to be noted that the choice of using DTW is not obvious. However, from related work, we knew that if this approach can work, there is a high chance that DTW will show it. We do not claim that DTW is better or worse than other approaches (e.g. machine learning).

Reference Sets

To identify a user, a reference set is required. This set represents the baseline for comparison and acts as the fingerprint of the user.

For each unlock screen, the reference set was created by taking the first 20 unlocks (each one a single unlock) for each user. Each of them was compared to the 19 other unlocks using DTW. Then, the average warp distance for the respective unlock was calculated. This is a common approach found in related work [19]. In the end, the unlock with the lowest average warp distance was chosen as the reference set. Taking the first 20 unlocks is based on the fact that the rest was required to measure false positive and false negative rates.

In a second step, the reference set was again compared to the remaining 19 unlocks. The 19 warp distances were then used to calculate the mean, median, minimum, maximum and standard deviation. Those values are used to define the upper border or the threshold for the interval (starting from 0) in which an unlock is considered as valid. This approach is depicted in figure 3. The green line represents the upper border of the interval. A possible value for this border is mean + standard deviation. The main assumption behind this is that additional unlocks performed by a user are likely to be within these intervals while the ones of other users (attackers) should lie above.

Logins and Attacks

The unlocks that were not used for creating the reference set were compared to the reference set using DTW. To check the system's resistance to attacks, the unlocks of all other participants were compared to the reference set. Thus, the success of the system was measured along the following parameters: True positives (TP): correctly accepted users. True negatives (TN): correctly rejected attackers. False positives (FP): wrongly accepted attackers. False negatives (FN): wrongly rejected users.

To compare the unlocks, the previously mentioned thresholds for valid unlocks were used in different variations. Furthermore, overheads (raising the threshold value) of 5%, 10%, 15%, 20%, 30% and 50% were added to the intervals and analyzed as well. The comparison itself was repeated several times using all possible combinations of the collected data (pressure, size, coordinates, etc.). This was done to find out which combination would perform best for the respective unlock screen. In short, the analysis was performed using different thresholds and different parameter combinations.

Results

The results are based on 30,720 unlocks (640 per participant). Due to the big amount of data that was analyzed using different combinations and intervals, the analysis was performed on a grid engine.

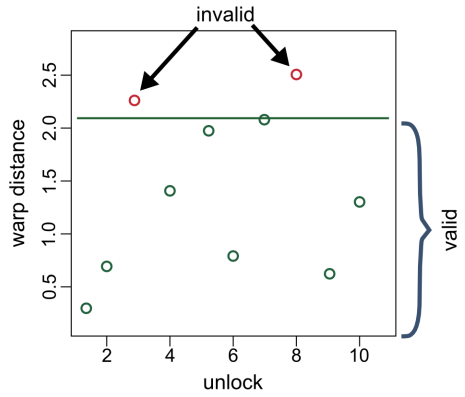


Figure 3: Identifying valid unlocks. The warp distance to the reference set is calculated. If the result is within the interval (lower than the threshold), the user is valid.

Accuracy Measurement

To calculate the accuracy on a percentage level, we used the following formula [38]:

number of correct assessments /number of all assessments.

$$Accuracy = \frac{\sum TN + \sum TP}{\sum TN + \sum TP + \sum FN + \sum FP}$$

It should be noted that unbalanced amounts of true positive and true negative rates easily influences the accuracy. In our work, there are much more attacks than valid unlocks, which means that false positives have a much higher influence on the accuracy than false negatives. Therefore, this number can only be seen as an indicator and we also need to look at the values separately rather than the accuracy only.

Logins and Attacks

It turned out that for all screens, similar upper borders of valid warp distances were the best performing (based on their accuracy). This was the value in the middle between the mean and the maximum plus the standard deviation (T=(mean+max)/2+STD). The following results for all screens are therefore based on this value. Mean, maximum and standard deviation are based on the warp distances between the reference set and the remaining 19 unlocks of the first 20 unlocks.

Table 1 shows the best results for the different unlock screens separated by the two days of the study. Desirable values are high true positive and true negative rates with low false positive and false negative rates. At first glance, a very interesting trend can be identified. While the true negative rate stays constant over the two days, the true positive rate decreases for all unlock screens. In the worst case (horizontal unlock), it decreases by 12%. However, the accuracy stays more or less constant, confirms the strong influence of higher numbers of attacks on the accuracy as mentioned before. This explains why the overall accuracy

	True Positives	True Negatives	False Positives	False Negatives	Accuracy	2-Day Accuracy
Horizontal day 1 (pressure)	2,640 (92%)	93,235 (52%)	87,245 (48%)	240 (8%)	52%	52%
Horizontal day 2 (pressure)	3,091 (80%)	93,726 (52%)	86,754 (48%)	749 (20%)	52%	
Vertical day 1 (time + XY)	2,822 (98%)	89,925 (50%)	90,555 (50%)	58 (2%)	50%	50%
Vertical day 2 (time + XY)	3,610 (94%)	89,387 (50%)	91,093 (50%)	230 (6%)	50%	
2 Finger Vertical day 1 (time)	2,689 (93%)	66,246 (37%)	114,234 (63%)	191 (7%)	38%	37%
2 Finger Vertical day 2 (time)	3,324 (87%)	65,350 (37%)	115,130 (63%)	516 (13%)	37%	
Diagonal day 1 (size + pressure)	2,666 (93%)	102,191 (57%)	78,289 (43%)	214 (7%)	57%	57%
Diagonal day 2 (size + pressure)	3,402 (89%)	102,223 (57%)	78,257 (43%)	438 (11%)	57%	

Table 1: Number of false positives, false negatives, true positives and true negatives as well as the accuracy for all unlock screens separated by the two study days. This table shows only the best results out of the different combinations of intervals and touch screen data.

for the diagonal unlock is the highest, because it had the best true negative rates in the experiment. A decreasing true positive rate with a constant true negative rate means that attackers stay as different from the user as before while the user's unlock differentiates more, the more time passed.

Overall, the vertical unlock with two fingers performed worst due to a very low true negative rate (37%). This means that 63% of all attacks were successful. This result is confirmed by the low overall accuracy of 37%. Again, this was the best result for the vertical unlock with two fingers, which was achieved using the event time.

Even though its accuracy is not the highest in the analysis, the vertical unlock with one finger performed best among the four unlock screens with a very high true positive rate (98% on the first and 94% on the second day). The optimal result was achieved using a combination of event time and the XY-coordinates for the analysis. This means that in the worst case, 6% of valid unlock attempts failed. At the same time, 50% of attacks were successful. Having a closer look at the data revealed more interesting findings that support the vertical unlock with one finger as being the best approach in the study. Of all the valid users, there was not a single user who could never be correctly identified. At the same time, 96% of users were always correctly identified.

Another interesting trend is worth being mentioned. For all unlock screens, around 50% of the attackers were responsible for more than 70% of successful attacks.

Discussion

The unlock approach provided a convenient way to quickly gather big amounts of data for analysis. Based on the results, we could investigate whether it is possible to differentiate users based on the way they perform unlocks. Especially the high true positive rates are encouraging.

Having shown how well users could be identified and thus how well the approach performs in terms of usability, there is a major drawback of this approach. In the best case, the true negative rate was 57%. This means that a little bit more than four out of ten attacks would have been successful. From a security point of view, this is not a very satisfying result.

Even though there is room for improvements for the unlock approach, the most promising way to go seemed to be using a method that allows for collecting significant more data per data set. Therefore, we decided to take the lessons learned from the pilot study and use them to enhance the security of the password pattern approach.

The lessons learned strongly influenced the design and analysis of the subsequent password pattern user study. The most important lessons learned are:

Factor time: The results of the unlock study showed that the usability of the system went down on the second day. This can be drawn back to the significant break between the two study parts. Users did not remember exactly how they per-

formed the pattern, which influenced the results. This influence is big enough to justify the claim that the data collection period of the next study should be done using a long-term design to gather more realistic data as the system has to work in everyday use.

Lab setting: The unlock study was performed in a controlled lab environment. This might have positively influenced the results. A more realistic study design is therefore preferable for the password pattern study.

Informed participants: We instructed the participants to always perform the unlock in exactly the same way. In an optimal setting, the system should work and provide security without this knowledge (working implicitly). Thus, uninformed users seem to be the more realistic choice for the follow-up study.

PASSWORD PATTERN USER STUDY

One of the main weaknesses of the pilot study was that data was collected using only two sessions. Within these two meetings, users were very likely biased to performing the unlock the same way. This effect is increased by the fact that they were told to perform the unlock as similar as possible. Overall, this leads to a strong positive bias that is not desired to test real world applicability.

Another problem was that a simple unlock only allowed for collecting a small amount of touch screen data. As seen in the pilot study, the unlock screens that created longer time series (not necessarily more data as shown by the bad results of the two fingers approach) had a tendency to lead to better results.

These problems were addressed in the password pattern study. The password pattern approach allowed for the collection of much longer time series. Additionally, we decided not to use a lab study but a long-term real world study instead to gather more realistic data. The pattern password has the additional advantage that many users are already familiar with it.

For the study, we developed an Android application that could easily be deployed to a bigger group of users. The application had two modes. On first start, the application was in training mode, allowing the users to train their password pattern until they felt familiar with it. After that mode was ended, there was no way to return to it. In study mode, the application allowed exactly one authentication per calendar day and closed automatically after one correct or three failed authentication attempts.

The standard layout, known from android phones as shown in figure 1 was used for the prototype. Several other layouts were tested but after informal studies, the decision was made to pick the layout the users are already used to.

This study copes with the problems of the pilot study resulting in more realistic data and overall longer time series (more sensor data) per authentication attempt. This enabled us to check whether changes over days (as observed during the pilot study) are acceptable if more complex gestures (password patterns) are used.

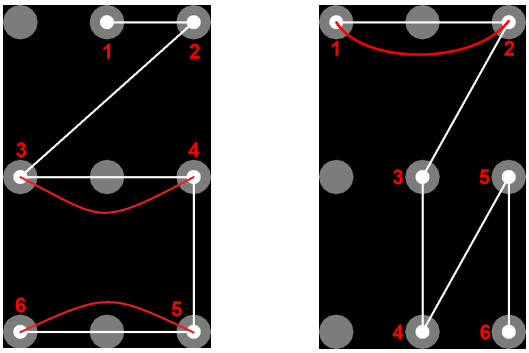


Figure 4: Password descriptions for the participants. The red lines indicate that a point has to be skipped.

User Study Design

The approach was evaluated using a repeated measures within participants longitudinal design. The task was to authenticate once a day using the test application. Overall, the participants were asked to perform the authentication task 21 times, resulting in a three weeks study.

The first day started with a training (the first mode of the test application). This data was not used for the analysis. Collecting the training data over a week would not have been feasible and the expected dropout rate would have been too high.

Procedure

On the first day of the study, the participants received an e-mail with detailed instructions on how to install the application and how to perform the training task. The Android application was provided via a download link and not via the Android market. This decision was made to keep the application private to the participants and to reduce delays in deployment. The application was installed on the participants' Android phones. After the training task, the 21 input days started. The daily input approach was introduced to further minimize the learning and habituation effects encountered in the pilot study.

Each participant was given a unique password pattern (sent with the e-mail) that was randomly assigned to the user based on an anonymous participant ID. The patterns consisted of five strokes, which make up a password space of 32,768 possible patterns (around 3.2 times higher than the password space of a four-digit PIN). There were three different categories of patterns: easy, medium, hard. Easy pattern consisted of simple strokes only. Medium patterns had one stroke for which a point had to be skipped while hard patterns had at least two skips. Figure 4 shows two patterns (medium and hard) as provided to the participants. As opposed to our expectations, the difficulty did not influence the results in any way and thus will not be mentioned again later in this paper.

In order to ensure that the users would not forget the input, an e-mail reminder was sent every day around noon. In case a user still forgot the input, an extension by one day was granted. That is, in the best case, the study took 21 days. Overall, 17 extensions for ten participants were granted. The maxi-

mum extension of the study was 26 days for one participant who forgot the input five times. Reasons for forgetting the input included weekend trips or days out.

When each participant had performed their 21 inputs, they were asked to come to a meeting and to bring their mobile device with them. During this event, the logged data was copied. After this, a second application was installed on the device of the user. The user was asked to input the correct password patterns of all other users three times with this new app. The objective was to simulate attacks on all participants. The respective log files were copied as well.

Till the end of the study, the participants were not instructed to perform the authentication in exactly the same way. This was done to avoid bias and to get insights on standard behavior, influenced by the users' daily routine. After the attack experiment, the participants were debriefed. The final meeting ended with a questionnaire covering usability and security questions. We also asked them to reflect on their behavior (e.g. if they tried to perform the password pattern in the same way each time), which was required to analyze the data.

Participants

The long term study started with 38 participants. The only prerequisite was the possession of an Android mobile phone (Android 2.1 or higher). 34 participants completed the study. The dropout rate was 11% (four users). In addition to those four users, three users had to be removed from the data sets. In two cases, the experimenter accidently gave away the purpose and description of the study before the participants had finished their tasks. The third person had to be removed since it turned out that he had someone else performing his tasks.

This means that overall, the study was correctly finished by 31 participants and thus the results are based on their data. The average age of these participants was 27 years. The youngest was 19 and the oldest 36 years. 19 participants were female, 12 were male. The incentive was the chance to win a popular gaming console at the end of the study.

Data Analysis

As for the pilot study, we used dynamic time warping (DTW) for the analysis. Data sets consisted of a series of touch screen events (XY-coordinates, pressure, size, time, speed). As opposed to the pilot study, the parameter speed was introduced (the time passed between two different coordinates).

Reference Sets

In this study, the reference set to identify a user was created by taking the first five valid authentication attempts and comparing them to each other. The overall approach is analog to the reference set creation of the pilot study with the exception that different possible reference sets were tested: smallest average warp distance, smallest median warp distance, smallest min and smallest max. In short, the warp distance to the other four authentications were calculated and the one with the lowest value (median, mean, min, max) was selected as the reference set. Then the mean, median, minimum, maximum and standard deviation to the other four was calculated and used for comparison with the remaining authentications. Depending on the number of identified possible reference sets, this comparison was done up to four times per user.

Since the number of attacks and own inputs is more even than for the unlock study, the accuracy measurement as introduced in the pilot study is a much more meaningful indicator of the quality in this study. Still it is important to look at the single numbers in detail.

Logins and Attacks

Since the data was collected using the mobile devices of the users, we had to deal with a big variety of hardware setups, including different screen resolution and quality. The different hardware would have influenced the results if we had compared every participant with every other participant. In order to avoid this, only users that owned the same type of devices were compared to each other. The biggest group within those was using the Nexus One with overall 18 users. In the end, there were five users with unique hardware setups. Thus, for those users, no valid attacks existed and they were removed from calculating the (overall) accuracy.

As for the unlock study, the true positive, true negative, false positive and false negative rates together with the accuracy were taken as a measure of performance. Again, all possible combinations of touch screen parameters (X-coordinate, Y-coordinate, size, pressure, time, speed) were taken into account. Also different variations of mean, median, standard deviation, minimum, maximum together with different overheads were tested.

Results

The following results are based on the data of the 31 valid participants. Overall there were 645 valid authentication attempts (including the data that was used to create the reference set) and 2790 attacks. We removed six inputs since the password pattern was wrongly input. Keep in mind that five users did not have valid attacks and therefore, accuracy is calculated using the data of 26 participants.

For the analysis of true positives and false negatives, the first five valid authentication attempts that were used to create the reference set were not taken into account. Based on the fact that users could fail to authenticate (by using a wrong shape), the maximum number of true positive plus false negatives per person was 16. This required significantly less comparisons than in the pilot study. Highly decreasing complexity of the analysis, the calculations were performed using a standard personal computer.

Logins and Attacks

After the analysis, the reference set based on the smallest median showed the best results in a combination with using the maximum warp distance as the threshold for valid inputs. The parameter combination that performed best consisted of pressure, size and speed.

True Positives	False Negatives	True Negatives	False Positives	Accuracy
398	92	852	231	77%
False Rejection Rate: 19%		False Acceptance Rate: 21%		

Table 2: Results for the reference set based on the smallest median in combination with the maximum as an upper border. Parameters are pressure, size and speed.

True Positives	False Negatives	True Negatives	False Positives	Accuracy
15	1	49	2	96%
12	4	51	0	94%
16	0	45	6	91%

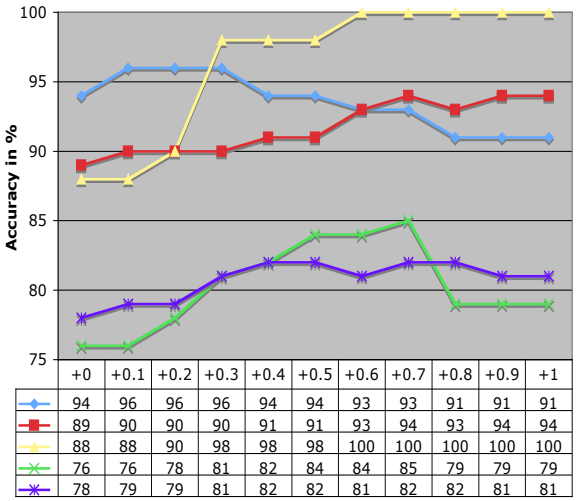
Table 3: Top 3 accuracies of participants in the password pattern study.

Table 2 shows the results for these combinations. Overall, the accuracy is 77% with a 19% false rejection rate and 21% false acceptance rate. The focus of the analysis was to go for high true positive rates to keep the system convenient and satisfying for the users. As mentioned before, in the final questionnaire, the participants were asked whether they tried to perform the input in the same way each time. The users that stated to apply this approach had a higher average accuracy (81%) than the users that did not think about this during input (72%). Interestingly, the difficulty of the password pattern does not influence the accuracy.

When compared to the pilot study, the overall accuracy increased by more than 20%. Even though the data was collected with significant breaks in between two consecutive inputs and under much more realistic circumstances, the password pattern allows for much more accurate measurements of the users' identity. Still, at first glance, 77% accuracy seems unsatisfying.

However, when looking at the results in more detail, the picture becomes positive. For instance, out of the 26 participants, for whom valid attacks existed, six reached an accuracy of 90% or higher. The top user reached an accuracy of 96% with one false negative and two false positives (out of 51). Table 3 shows the top 3 participants and their results.

In addition, one specific group of users drew our attention. Its users had extremely low false acceptance rates (mostly zero false positives) but at the same time their false rejection rate was rather high or unsatisfying (in the worst case there was only one true positive). The second best user (see table 3) was part of this group. By setting the threshold too low, (almost) all attackers were excluded but many valid attempts failed as well. Looking at these users' reference sets showed that their thresholds were quite low compared to other users. Again, taking into account the hypothesis that users are more similar to themselves than to attackers in the way they perform the password patterns, we performed a small experiment to see whether this theory holds. We iteratively increased the threshold for these users. That is, higher warp distances between two authentication attempts were accepted as valid. We conducted this for the previously mentioned five users. The results are shown in figure 5. For each participant, the overall accuracy could be increased significantly. In one case, it was improved from 88% to 100%. This was due to the fact that false acceptance rates stayed constant while true positives increased.



	+0	+0.1	+0.2	+0.3	+0.4	+0.5	+0.6	+0.7	+0.8	+0.9	+1
	94	96	96	96	94	94	93	93	91	91	91
	89	90	90	90	91	91	93	94	93	94	94
	88	88	90	98	98	98	100	100	100	100	100
	76	76	78	81	82	84	84	85	79	79	79
	78	79	79	81	82	82	81	82	82	81	81

Figure 5: Threshold increased for five users with low true positive rates. The graphs show that with increased threshold (warp distance +0 till +1), the accuracy significantly improves. Peaks show up at different overheads.

Discussion

The improvements on the approach and the study lead to more realistic and less biased data triggering much better results. At first glance this seems odd but was achieved by using password patterns instead of a simple unlock. Enhancing them with implicit behavioral authentication allows for creating a convenient authentication system with good security properties under the worst circumstances (attacker in possession of the mobile device and the password pattern).

The results of the second study support our claim that password patterns create data that is distinct enough to distinguish between different users. Overall, it can be stated that using touch screen data to identify users works to a certain degree. This is supported by the fact that increasing the threshold for valid authentication attempts improves overall accuracy (as shown in figure 5). The comparison to the pilot study also indicates that this approach works the better the more sensor data is available.

An interesting observation was that informed users performed better than uninformed ones. That is, users that tried to perform the authentication in the same way each time (considering the finger to use, speed etc.) achieved, on average, higher accuracy values than the ones that did it randomly. This means that the security (and performance) can be influenced by the user. The more consistently the authentication is done, the smaller the threshold, making it harder for an

attacker. However, uninformed users achieved good results as well, which indicates that the approach can work in normal everyday use without specific precautions. This is very promising from a usability point of view. It means that users can rely on the highly usable and memorable password patterns at highly improved security.

LIMITATIONS AND IMPROVEMENTS

As opposed to the pilot study, the password pattern study had much less inputs, as it is hard to collect realistic data over a longer period of time. With time frames longer than three weeks, the dropout rates of an experiment of this kind can quickly increase. Still, more data would help to further support our findings. The lack of attackers for some of the participants is undesirable. In the best case, in future work, attacks should be performed using the participant's (the victim's) own device. However, convincing users to give out their mobile phone, which usually holds private data, seems hard. Thus, a solution could be to hand out mobile phones for the purpose of the experiment. On the other hand, this would negatively influence the realistic setting.

As the experiment with increased threshold showed, there is still a lot of open space to improve the approach. For this purpose, both, the reference set creation and the comparison algorithm can be optimized. For instance, the results of the second study suggest that a dynamic reference set or a reference consisting of several unlocks/authentication attempts, which changes over time as the user makes use of the system, can positively influence accuracy. The idea of a dynamic reference set, and thus a dynamic threshold is further supported by the fact that the five different users shown in figure 5 reached their best results at different overheads (added to the threshold). Weighted parameters are a promising option in changing the algorithm as well. As is, all parameters have the same importance and weight when compared with DTW.

Finally, there is a weakness of the two studies with respect to the analysis of their security. In both cases, the attacks happened without the attacker ever seeing the actual input of the victim. That is, we cannot say whether it is possible to "shoulder surf" the way a user does the input and copy it.

CONCLUSION AND FUTURE WORK

In this paper, we presented an implicit approach to improve authentication on current mobile devices. The basic idea was to exploit touch screen data of common smartphones (without adding additional hardware) to identify users based on the way they perform an action. For this, we chose to evaluate unlock screens as well as password patterns that come with Android phones. The basic assumption was that password patterns are convenient and usable but at the same time highly insecure. By adding implicit authentication, an invisible layer of security is added to the input, which makes the system resilient to attacks under the worst circumstances (stolen mobile phone and password pattern).

It should be noted again that the approach, presented in this paper, provides immediate authentication as opposed to most work in the field of behavioral biometrics. Once the password pattern is input, the system decides instantly whether the user is authorized. The results of two studies provide first proof that it is possible to distinguish users and to improve the security of password patterns (and even screen unlocks). The results also show that the more data points a data set consists of, the easier it is to make this distinction. This means that by increasing the password length, positive effects on accuracies might be observed. However, this would come at the costs of decreased usability and memorability.

The two main open points for future work are: (a) We are currently implementing a prototype based on the presented approach that does the calculation on the mobile device to perform another long-term study based on this application. This way, dynamic reference sets can be tested in real time. Additionally, it will enable us to perform shoulder surfing tests to further evaluate and judge the security of the approach. (b) The accuracy of the system has to be increased. While in this work, the focus was on showing that implicit authentication works, in future work, less naïve approaches have to be compared to improve the accuracy of the system. For instance, DTW should be compared to, for instance, machine learning approaches.

ACCESS TO THE DATA

Interested in getting access to the (anonymized) data of the two studies? Just contact the first author of this paper.

ACKNOWLEDGMENTS

This work was funded by a Google Research Award.

REFERENCES

1. Aviv, A. J., Gibson, K., Mossop, E., Blaze, M., Smith, J. M. Smudge attacks on smartphone touch screens. *In USENIX 4th Workshop WOOT 2010*.
2. Bigun, J., Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzales-Rodriguez, J. Combining biometric evidence for person authentication. *Advanced Studies in Biometrics*. Springer (2005), 1-18.
3. Brunelli, R., Falavigna, D. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(10). (1995), 955-966.
4. Buchoux, A., Clarke, N.L. Deployment of keystroke analysis on a smartphone. In *Proceedings AIMS 2008*.
5. Card, S., Moran, T., Newell, A. Computer text-editing: An information-processing analysis of a routine cognitive skill. *Cognitive Psychology, 12(1).* (1980), 32–74.
6. Chow, R., Jakobsson, M., Masuoka, R., Molina, J., Niu, Y., Shi, E., Song, Z. Authentication in the clouds: a framework and its application to mobile users. *In Proceedings Workshop CCSW 2010*. ACM Press (2010), 1-6.
7. Clarke, N.L., Furnell, S.M. Authenticating mobile phone users using keystroke analysis. *International Journal of Information Security, 6(1).* Springer (2007), 1-14.

8. Clarke, N.L., Furnell, S.M., Rodwell, P.M., Reynolds P.L. Acceptance of subscriber authentication methods for mobile telephony devices. *Computers & Security, 21 (3).* (2002), 220-228.
9. Coventry, L., De Angeli, A., Johnson, G. Usability and biometric verification at the ATM interface. In *Proceedings CHI 2003*. ACM Press (2003), 153-160.
10. Cutting, J., Kozlowski, L. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society, 9(5).* (1977), 353–356.
11. De Luca, A., Denzel, M. and Hussmann, H. Look into my eyes! Can you guess my password? In *Proceedings SOUPS 2009*. ACM Press (2009), 7:1-7:12.
12. Dunphy, P., Yan, J. Do background images improve "draw a secret" graphical passwords? In *Proceedings CCS 2007*. ACM Press (2007), 36-47.
13. Fleishman, E., Parker, J. Factors in the retention and relearning of perceptual-motor skill. *Journal of Experimental Psychology, 64.* (1962), 215–226.
14. Francis, L., Mayes, K., Hancke, G., Markantonakis, K. A location based security framework for authenticating mobile phones. *In Proceedings Workshop M-MPAC 2010*. ACM Press (2010), 5:1-5:8.
15. Giorgino, T. Computing and visualizing dynamic time warping alignments in R: the DTW package. *Journal of Statistical Software, 31(7)*. (2009), 1-24.
16. Gafurov, D., Helkala, K., Søndrol, T. Biometric gait authentication using accelerometer sensor. *Journal of Computers, 1 (7).* Academy Publisher (2006), 51-59.
17. Jakobsson, M., Shi, E., Golle, P., Chow, R. Implicit authentication for mobile devices. In *Proceedings* HotSec 2009. USENIX Association, 9-9.
18. Jermyn, I., Mayer, A., Monrose, F., Reiter, M. K., Rubin, A. D. The design and analysis of graphical passwords. In *Proceedings SSYM 1999*. USENIX Association.
19. Kar, B., Dutta, P. K., Basu, T. K., Vielhauer, C., Dittmann, J. DTW based verification scheme of biometric signatures. In *Proceedings ICIT 2006*.
20. Karlson, A., Brush, A.J., Schechter, S. Can i borrow your phone? Understanding concerns when sharing mobile phones. In *Proceedings CHI 2009*. ACM Press (2009), 1647-1650.
21. Kim, D., Dunphy, P., Briggs, P., Hook, J., Nicholson, J., Nicholson, J., Olivier, P. Multi-touch authentication on tabletops. *In Proceedings CHI 2010*. ACM Press (2010), 1093-1102.
22. Legget, J., Williams, G., Usnick, M. Dynamic identity verification via keystroke characteristics. *International Journal of Man-Machine Studies, 35 (6)*. Academic Press Ltd (1991), 859-870.
23. Mantyjarvi, J., Lindholm, M., Vildjiounaite, E., Makela, S. M., Ailisto, H.A. Identifying users of portable devices from gait pattern with accelerometers. In *Proceedings ICASSP 2005*.
24. Marcel, S., Cool, C., Atanasoaei, C., Tarsetti, F., Pesán, J., Matejka, P., Cernocky, J., Helistekangas, M., Turtinen, M. MOBIO: mobile biometric face and speaker authentication, In *Proceedings CVPR 2010*.
25. Nelson, D. L., Reed, V. S., Walling, J. R. Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory 2* (5). (1976), 523-528.
26. Pons, A.P., Polak, P. Understanding user perspectives on biometric technology. *Commun. ACM,* 51 (9). ACM Press (2008), 115-118.
27. Rogers, J. Please enter your four-digit pin. Financial Services Technology, U.S. Edition Issue 4 (Mar. 2007).
28. Rokita, J. Krzyzak, A., Suen, C.Y. Cell phones personal authentication systems using multimodal biometrics. In *Proceedings ICIAR 2008*. Springer (2008), 1013-1022.
29. Sakoe, H., Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing, 26(1)*. (1978), 43-49.
30. Shadmer, R., Brashers-Krug, T. Functional stages in the formation of human long-term motor memory. *The Journal of Neuroscience, 17(1)*. (1997), 409-419.
31. Shi, E., Niu, Y., Jakobsson, M., Chow, R. Implicit authentication through learning user behavior. In *Proceedings ISC 2010*. Springer (2011), 99-113.
32. Sonkamble, S., Thool, R., Sonkamble, B. Survey of biometric recognition systems and their applications. *Journal of Theoretical and Applied Information Technology, 11(1)*. (2010), 45-51.
33. Standing, L. Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology, 25(2).* (1973), 207-22.
34. Tamviruzzaman, M., Ahamed, S. I., Hasan, C. S., O'brien, C. ePet: When cellular phone learns to recognize its owner. In *Proceedings Workshop SafeConfig 2009*. ACM Press (2009), 13-18.
35. Wood, H.M. The use of passwords for controlled access to remote computer systems and services. In *Proceedings AFIPS 1977*. ACM Press(1977), 27-33.
36. Weiss, R., De Luca, A. PassShapes: utilizing stroke based authentication to increase password memorability. In *Proceedings NordiCHI 2008.* ACM Press (2008), 383-392.
37. Yazji, S., Chen, X. Dick, R.P., Scheuermann P. Implicit user re-authentication for mobile devices. In Proceedings *UIC 2009.* Springer (2009), 325-339.
38. Zhu, W., Zeng, N., Wang, N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. In *Proceedings Nesug 2010*.