Walk the Walk: Attacking Gait Biometrics by Imitation

Bendik B. Mjaaland[1],
Patrick Bours[2], and Danilo Gligoroski[3]

[1] Accenture Technology Consulting - Security, Norway,
bendik.mjaaland@accenture.com,
[2] Norwegian Information Security Laboratories (NISlab), Gjøvik University College, Gjøvik, Norway,
patrick.bours@hig.no,
[3] Norwegian University of Science and Technology, 7491 Trondheim, Norway,
danilog@item.ntnu.no

Abstract. Since advances in gait biometrics are rather new, the current volume of security testing on this feature is limited. We present a study on mimicking, or imitation, of the human gait. Mimicking is a very intuitive way of attacking a biometric system based on gait, and still this topic is almost nonexistent in the open literature. The bottom line question in our research is weather it is possible to learn to walk like someone else. If this turned out to be easy, it would have a severe effect of the potential of gait as an authentication mechanism in the future.
We have developed a software tool that uses wearable sensors to collect and analyze gait acceleration data. The research is further based on an experiment, involving extensive training of test subjects, and using various sources of feedback like video and statistical analysis. The attack scores are analyzed by regression, and the goal is to determine whether or not the participants are increasing their mimicking skills, or simply: if they are *learning*.
The experiment involved 50 participants enrolled into a gait authentication system. The error rates compete with state of the art gait technology, with an EER of 6.2%. The mimicking part of the experiment revealed that gait mimicking is a very difficult task, and that our physiological characteristics work against us when we try to change something as fundamental as the way we walk. The participants showed few indications of learning, and the results of most attackers even worsened over time, showing that training did nothing to help them succeed.
The research identified a natural boundary to the impostors' performance, a point of resistance so significant that it was given a name; a *plateau*. The location or value of this plateau predetermines the outcome of an attack; for success it has to lie below the acceptance threshold corresponding to the Equal Error Rate (EER).

Keywords Gait, biometrics, imitation, circumvention, fraud.

1 Introduction

Biometrics is the study of using intrinsic biological features to identify individuals. While the study of these features has had a long and successful history in forensics, the use of biometrics in automated recognition systems is a fairly recent accomplishment. Biometric technology is evolving at an unprecedented speed, and new ideas emerge accordingly. The deployment of large-scale biometric systems in both commercial (e.g. Disney World [11], airports [6]) and government (e.g. US-VISIT [18]) applications has served to increase the public awareness of this technology. This rapid growth in biometric system deployment has clearly highlighted the challenges associated with designing and integrating these systems.

The ability to automatically confirm the identity of an individual (i.e. authentication) is vital to information security, and biometrics is often used for this purpose. Among the recent advances in biometrics is the use of gait as an identifier. Research suggest that individuals can be identified by the way they walk, hopefully encouraging further development of gait biometrics. More research is indeed necessary, as the technology is far from mature; the performance is not generally competitive to other biometrics, and the overall volume of security testing is small.

Biometric features are divided into two categories: physiological and behavioral biometrics. Physiological biometrics are characteristics that you cannot alter easily, they are stable parts or properties of

your body. Examples are fingerprints, DNA, iris and retina. Behavioral characteristics can be altered and learned, such as gait, signature and keystroke dynamics. An interesting thing to note is that even though these two categories have similar vulnerabilities, they are subject to very different forms of attacks.

An area where biometric authentication is considered to have great potential, is authentication on cell phones. Cell phones are being used in high security applications such as mobile banking and commerce [8, 19]. Surveys of mobile phone users indicate that users do not follow the relevant security guidelines, for example they do not change their PIN codes regularly or use the same code for multiple services [4]. Furthermore, British crime survey for 2005/06 reported that estimated 800000 owners had experienced mobile phone theft [5]. There is an obvious need to strengthen the authentication of cell phones, and biometric features can indeed be used for this purpose. Imagine a stolen phone locking down asking for PIN-authentication, because it does not recognize the walk of the current holder.

To understand the threats against biometric systems, it is valuable to look at the different points of attack. Ratha et. al defined eight such points in [20], illustrated by Figure 1. These points are described as follows:

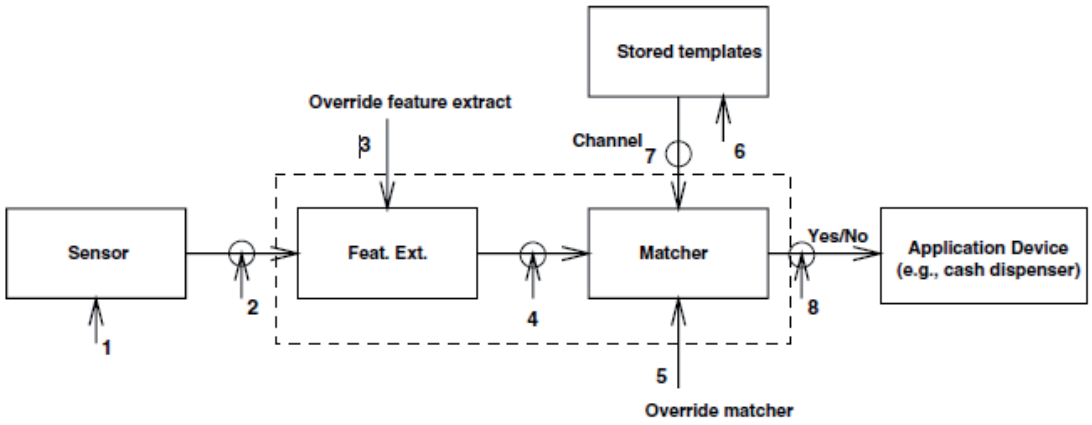

Fig. 1. Eight attack points in biometric authentication systems [20].

1. Presenting fake or imitated biometrics to the scanner or collector.
2. Re-submitting previous samples or altering transmitted samples.
3. Attacking the feature extractor so that it produces attacker dictated values.
4. Substituting extracted feature values with values dictated by the attacker.
5. Manipulating the matcher so that a desired score is produced.
6. Manipulating the template database.
7. Attacking the transmission channel between the database and the matcher.
8. Manipulating the decision produced by the system as a whole.

This research will concentrate on mimicking of gait, which corresponds to attack point one. The reader should keep in mind that this is only one of many approaches to attacking a biometric system, but it should also be noted that this is a very intuitive and easy way to attack gait authentication systems.

This paper is organized as follows. Section 2 introduces gait biometrics in its current state of art. Section 3 presents technology and methods chosen for this research, mainly focusing on the gait analysis and template processing. Section 4 describes the experiment conducted, and Section 6 presents the results and findings. Finally, Section 7 and 8 presents conclusions and future research options, respectively.

2 Gait Biometrics

2.1 State of the Art

The gait of a person is a periodic activity with each gait **cycle** covering two strides - the left foot forward and the right foot forward. It can also be split up into repetitive phases and tasks. Gait recognition has

intrigued researchers for some time, already in the 1970's there were experiments done on the human perception of gait [13]. The first effort towards automated gait recognition (using machine vision) was probably done by Niyogi and Adelson [17] in the early 1990's. Several methods are known today, and we can categorize all known gait capturing methods into three categories [9]: machine vision based (MV), floor sensor based (FS) and wearable sensor based (WS).

Machine vision was the main focus of gait biometrics in the earlier stages [13], and utilize the shape characteristic of the human gait. Most of the MV-based gait recognition systems are based on the human silhouette [9]. That is, the background is removed and the silhouette is analyzed for recognition. Many approaches to the analysis are possible. One is to compute the average silhouette over an entire gait cycle. Such methods face some challenges if the video background is not known and adjusted in advance of the video capturing.

Fig. 2. An MR100 wearable sensor attached to a belt, a video camera on a tripod in the back.

Floor sensors can be installed in floors or carpet-like objects, and are able to measure gait related features when walked upon. This will result in footstep profiles that can be based on positioning, or the timing between heel vs toe strikes [9].

Using wearable motion recording sensors to collect gait data is a rather newly explored field within gait biometrics. One of the earliest description of the idea can be found in Morris' [15] PhD thesis from Harvard University in 2004. Since then, the academic community at Gjøvik University College(GUC) has devoted much effort researching gait biometrics. Gafurov's PhD work covers a broad part of WS-based gait recognition [9], and several students have written their master's thesis on the same topic [3, 12, 16, 22]. Figure 2 shows lab equipment used in this research.

In [9] Gafurov et al. used ankle-attached sensors and achieved EERs of 5% and 9% by utilizing the *histogram similarity* and *cycle length method*, respectively. These methods were also applied when attaching sensors to the hip, and an EER of 18% was achieved [3]. This result was vastly improved by Holien by fine-tuning the cycle detection and other subtasks of Gafurov's algorithms; an EER of 2% was achieved for normal walking [12], outperforming any other research at the time of writing.

WS-based gait collection has the advantage of being a rather unobtrusive way of collecting biometric data. It also has the immense advantage over MV of avoiding external noise factors such as camera placement and background or lighting issues. Furthermore, MV and FS is an expensive solution in terms of camera and floor equipment, while the WS do not require any infrastructure in the surroundings, and it is mobile.

2.2 Security Testing of Gait Biometrics

Since gait biometrics is a rather new area within biometrics, only a small amount of security testing has been published in the open literature. Imitation is an intuitive and easily attempted way of attacking

gait biometrics, therefore it is vital that it receives focus in future research. Mimicking of gait has been performed as part of an experiment in two cases, both constituting minimal-effort mimicking attacks. This section will briefly present this research, and hopefully justify the need for further testing.

Gafurov's Spoofing Attempt. One mimicking experiment was conducted by Gafurov et al. [9, 10], and involved 100 test subjects. The group was divided into pairs, where everyone attempted to imitate the other. In other words, everyone played both the role of attacker and victim. In total, two rounds of mimicking were performed. In the first round, the targeted person walked twice in front of the attacker, and in the second the attacker was trying twice to mimick alone. By also including a friendly scenario, Gafurov created a baseline for comparison of the hostile scenario. The achieved EER for the friendly scenario was approximately 16%.

For the analysis of the hostile scenario, various distance metrics were applied. It turned out that the performance of the system when trying to mimic the gait was worse than the performance of the friendly scenario, though the difference was hardly significant. One way to explain the reduced performance is that focusing on how you walk might actually make your gait unnatural and uneven.

However, there are several reasons why this research picks up the topic and continues the spoofing attempts. Gafurov performed what he called a minimal-effort study. Four attempts to mimic a person will hardly provide a valid indicator on how training or learning affects results. The experiment designed for our research involves a lot more, and a very different kind of training. There will be more attempts, more sources of feedback and more time devoted to the training.

Stang's Spoofing Attempt. Another experiment on spoofing gait biometrics was designed and conducted by Stang [22]. A total of 13 participants contributed with 15 attempts of mimicking, in a very different setting from that of Gafurov. The trials were performed indoors, in a room where a projector displayed graphical information about the victim's gait. This was meant to be the main source of feedback for the impostors, along with a short, informal description of the gait they were targeting. This could for instance be "normal" or "slow" walk. The participants watched the graphical information while walking, as it was dynamically updated. Each attempt lasted five seconds, and the experiment lasted about 20-30 minutes in total. After each attempt a match score between 0 and 100 was displayed, based on correlation such that 100 is a perfect match.

Stang used a correlation coefficient to compare the gait data, and his report refers to a psychological study in an attempt to establish an interpretation of what is considered "high" and "low" correlation. He bases the thresholds for successful mimicking on this, 60 being the most strict value. There are obvious flaws in this approach, projecting human perception onto security strength is a highly speculative method. A 60% requirement is not very strict in biometrics, and this threshold should have been derived from a "friendly" scenario, where users are enrolled in order to find thresholds corresponding to the EER.

Another drawback in Stang's work is the way he attempts to identify learning. His approach consists of doing a linear approximation where he looks at the angle of the fitted curve, and draws conclusions based on his own impression of its steepness. The report does not present any valid statistical tests made on the data, and no confidence intervals or measures of fit are provided. Without hypothesis testing his conclusions are not justifiable, and the curve itself is based on very few data points, with a sample rate as low as 30, making the approximations extremely uncertain.

Finally, Stang's experimental environment is unfortunate. The mimicking is based on a five second walk, which is rather short when considering the high probability of an unnatural start and finish. Furthermore, this short time period does not leave any possibility for the attacker to adjust his manner of walking along the way.

2.3 Contribution

As previous security testing of gait biometrics have not gone further than to conduct minimal-effort mimicking attacks, the research on gait security must be significantly elaborated. In this research more effort will be put into the actual mimicking, participant training will be taken to a much higher level, and hopefully this will help advance knowledge in the field.

With an extensive training program spanning over several sessions, using sources of feedback like video taping, statistical data and personal coaching, participants have a much better basis to succeed. The main goal is to identify a learning curve for the test subjects, to determine how the training affects their results over time, and to learn how far this training can advance their skills.

3 Choice of Technology

3.1 Gait Analysis

Due to advantages mentioned in Section 2, WS-technology was selected for this research. The wearable sensors use accelerometers to collect acceleration data in three dimensions. Data from the X, Y and Z direction constitute fragments of gait acceleration, and Gafurov reports in [9] that using a combined signal has the advantage of making the scheme less sensitive to random noise.

The analysis is based on the Cycle Length Method (CLM) developed by Gafurov [9] at Gjøvik University College, but the method has been changed in several ways. The CLM method is essentially a framework on how to turn raw gait data into an averaged gait cycle of a fixed length. When the average cycle is found, this can be used as a user gait template.

A high-level overview of the processing of raw gait data is shown in Figure 3. The first stage consists of preprocessing, where raw gait data is filtered and interpolated. The next three stages estimate the length of the cycles, detect their exact locations and normalizes them to the same length, respectively. The set of cycles is then cleaned (i.e. by removing outliers), and finally the average gait cycle is computed using simple averaging techniques. This average can be compared to other averages by well-known distance metrics, and thus constitutes the template of a user. This will be further explained later in this section.

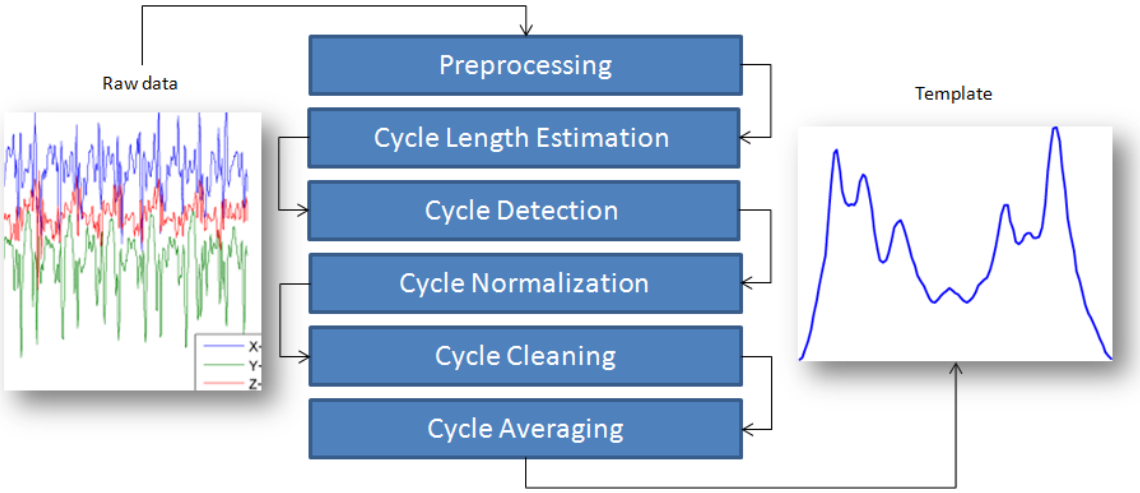

Fig. 3. Raw Gait Data Processing Overview.

3.2 Dynamic Time Warping

In biometrics it is common to produce a match score when comparing templates. This match score can be based on various distance metrics, measuring the separability between two sequences of data. In this research Dynamic Time Warping (DTW) is used, a choice based on the results of Holien's optimization in [12]. DTW can be used for various purposes in science; essentially the DTW measures the minimal amount of change needed in one sequence of data, in order to make it identical to another. When used to compare templates, DTW disposes the naturally occurring changes in walking speed, is able to compare signals of different lengths, and signals where the x-axis is shifted. More information on DTW can be found in [14], and in [12] where it stands out as the best distance metric for gait biometrics.

3.3 Hardware

For the gait collection Gjøvik University College provided two wearable sensors: the Motion Recording 100 (MR100), and the Freescale ZSTAR sensor. The MR100 (see Figure 2) had a great advantage over the Freescale - the sample rate. The MR100 takes approximately 100 samples per second, while the Freescale takes about 30 [22], and naturally - higher sample rate means more reliable measurements. For this reason, the MR100 was used during the entire course of this research.

There are numerous possibilities regarding sensor attachment, research by Gafurov, Ailisto and others utilized several different body parts such as hip (belt or pocket), lower leg and arm [1–3, 9]. Several portable devices such as cell phones and PDAs are in need of secure authentication mechanisms, and these are typically carried in the hip area. The fact that devices like these are so extensively used in the modern world, makes this a good reason to use hip attachment.

When analyzing hip movement, there are more attachment points to choose from. One is the pocket, but this location has some disadvantages. The main disadvantage here is that it is difficult to know the exact orientation of the device. It is also difficult to attach it in a stable manner inside pockets. Other challenges present themselves due to the design of the clothing; pockets vary in size, shape and location, and some jeans may not have pockets at all.

This research is based on belt attachment. The belt can be mounted to any person's hip regardless of what they are wearing. The device will always have the same orientation, and it will be visible at all times. EER values achieved for hip based gait authentication in general range from 2% to 13% [9], so the hip is a good choice also due to maturity.

Input: Raw 3-dimensional gait acceleration data G_u for the user u.
Output: A template t_u for the user u.
1. Interpolate G_u to obtain stable sample lengths.
2. Filter G_u using a Weighted Moving Average (WMA) filter.
3. Convert the raw values in G_u to G-forces.
4. From G_u, calculate the resultant acceleration R_u to obtain 1-dimensional data.
5. Estimate the length L of the gait cycles in R_u.
6. Using L, detect all cycles in R_u and store them in the set C_u.
7. Normalize all cycles in C_u to make their lengths identical.
8. Omit irregular cycles from C_u.
9. Return the average cycle t_u.

Table 1. Algorithm for creating templates. The raw gait data G_u for a user u is processed, resulting in a template for that user, t_u.

3.4 Creating Templates

As seen in the high-level algorithm in Table 1, and in the illustration in Figure 3, the template creation consists of several steps. Each step will be covered with some detail.

Step 1-4: Preprocessing. The preprocessing stage consists of interpolation, filtering, G-force conversion and resultant calculation. All these steps are necessary prior to the gait cycle processing.

1: The interpolation is necessary due to sample rate instability. On average the MR100 sampled about 100 times per second, but the actual rate was not constant. To ease further analysis of the gait signal, it was desirable to have a constant time axis with one sample every $\frac{1}{100}$ second. The raw sample rates were close to the desired values, so standard interpolation tools were put to use.

2: Raw acceleration data may contain extreme values due to many potential noise factors. There are many ways of removing noise from signal, Holien considered two possible filters in [12], namely the Moving Average and the Weighted Moving Average. The (Weighted) Moving Average filters rely on values neighboring the value in question. That is, if an extreme value lies in between several average values, it

will be drawn towards the average to an extent determined by the filter's characteristics. The Weighted Moving Average filters let the closest neighbors count the most, with a variable window size, while the non-weighted Moving Average filters do not. This research adopted Holien's filtering approach, the WMA with window size of five.

3: The MR100 sensor only provides raw acceleration data, so it is necessary to perform a conversion. The values provided are all in the range $[0, 1023]$, which corresponds to the range $[-6g, 6g]$ where g is the gravitational constant. In an ideal world we would be able to convert this using math alone, but the sensor values shift slightly during its lifetime. To solve this constants were found, experimentally, that shift the range of g-forces back to the intended normal.

4: Finally in the preprocessing, the resultant acceleration is needed. We are dealing with acceleration in three dimensions, and need to construct a combination of these in order to create a template for the user. Many combinations have been tried, some of them are very simple, even excluding specific dimensions [1], while others are more sophisticated [9]. In this research all dimensions were included:

$$r_t = \sqrt{x_t^2 + y_t^2 + z_t^2},\ t = 1, \ldots, n \tag{1}$$

where r_t, x_t and z_t are the magnitudes of resulting, vertical, horizontal and lateral acceleration at time t, respectively, and n is the number of recorded samples.

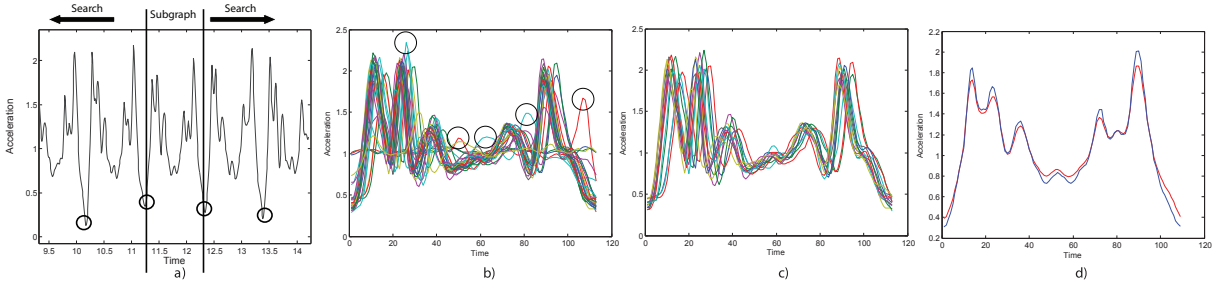

Fig. 4. Gait cycle processing. Cycle length estimation and cycle detection is illustrated in *a*). In *b*) and *c*) the detected cycles are superimposed on top of eachother, and in *b*) irregular cycles are identified and marked with circles. A cleaned cycle set can be seen in *c*), and the computed average (mean and median) in *d*).

Step 5: Cycle Length Estimation. In order to ease the process of detecting cycles, we want to obtain an estimate of how long each cycle is. Normal walk constitutes about 100 samples per cycle from the MR100, but this is not accurate enough. People will differ according to the length of their legs, their weight and walking speed. So prior to any cycle detection, we want to perform an automatic estimation of the cycle length.

In this research a correlation-based approach from [12] is used. A subgraph from the middle of the gait sequence is extracted and compared to other parts of the graph. By calculating the correlation between the subgraph and other parts of the graph, we get "peaks" of high correlation values where the subgraph fits well. The average distance between these peaks is the cycle length estimate. Other fine-tunings of this scheme were applied, but will not be described here.

Step 6: Cycle Detection. The "actual" cycle of a person could be defined to start when the right foot is lifted, and end when that foot is back in the same position. However, in gait biometrics this does not have to be respected, a cycle can be defined as any repetitive part of the gait data.

Using the cycle length estimate, approximate starting points of cycles can be found. From here, minima are used for the detection. Let N be the estimated cycle length and L the length of the entire gait sequence.

6.1. The minimum M within the middle section of the gait sequence is detected, and used as starting point. This minimum defines the start of a cycle, and will be used as a base point to find others.

6.2. A search is made forward in the gait signal by jumping N steps in this direction, and scanning the new point for another minimum, with a buffer of 10 samples in each direction. This is repeated until the end of sequence is reached. The minima are stored.

6.3. Step 6.2 is repeated backwards from M.

6.4. The resulting cycle vector is produced, containing the sample locations of all the discovered minima. Hence, the exact location of each cycle is now known.

Step 7: Cycle Normalization. In order to perform the averaging of gait cycles, we need all the cycles to be of the same length. Several noise factors can cause the number of samples to fall short or exceed the average. Hence, the deviating cycles has to be stretched or shrunk in order to fit the desired range of samples. If the cycle is longer or shorter than the average, interpolation is used to perform this adjustment. If the cycle is equal to the average, it can be used as it is.

Step 8: Cycle Cleaning. To optimize the results it is desirable to exclude outliers before calculating the average cycle. Such cycle cleaning can be performed in many ways; in this research DTW is used as a statistical distance metric to find irregular cycles. Every cycle is compared to each other, and each cycle's average distance to every other cycle is calculated. Then, the grand distance average is computed taking the mean of these values. Finally, every cycle's average is compared to the grand average - if the value is too far from the grand average the whole cycle is omitted from the cycle set. The process is repeated for a series of "sliding" thresholds, decrementing in size.

Step 9: Cycle Averaging. With all cycles normalized to the same length, calculating the cycle average is technically straightforward. Let each sample in the cycle belong to a group or set, and average these values. Another way to picture this is as column averaging in a matrix filled with one cycle in each row.

Input: Two gait templates t_1 and t_2.
Output: Distance score D.
1. Identify the highest acceleration value, or peak P in t_1.
2. In the set Q, store the n highest acceleration peaks found in t_2, where n is an integer.
3. For all n peaks in Q:
3.1. Circularly shift t_2 such that P in is aligned with the current peak.
3.2. Calculate and store the DTW distance score between t_1 and the shifted t_2.
4. Return the lowest DTW score D.

Table 2. Algorithm for comparing templates. Two templates t_1 and t_2 are compared using DTW. A distance score D is returned as output.

3.5 Comparing Templates

Template comparison forms the basis of verification in the biometric system. The procedure consists mainly of DTW distance comparison, but an enhancement used in this research is time shifting, illustrated in Figure 5. Due to natural fluctuations in the gait data, templates can differ in time, and become "out of sync". In essence, the system may interpret cycles to have different start and end points, even if the user is the same. Hence, a circular shift is performed to obtain the best possible match. The algorithm for template comparison is shown in Table 2.

4 Experiment Description

The experiment is divided into three scenarios. The first is the **friendly scenario**, consisting of regular gait data from a test group of 50 participants. The subjects did not receive much instructions or training,

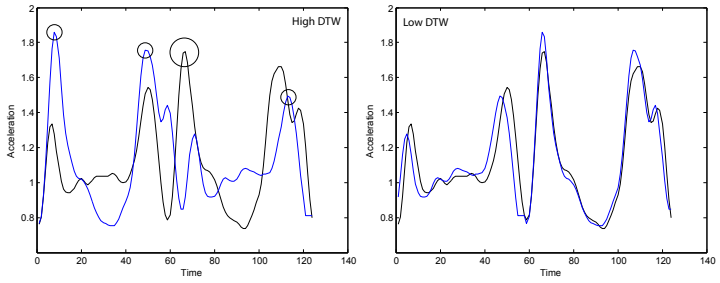

Fig. 5. Time shifting. Left: Two similar templates with different cycle interpretations. Right: The templates have been circularly shifted to achieve a better match, using maxima to align.

but simply walked a fixed distance while wearing a sensor. This allows the calculation of error rates and provides a basis for comparison with other parts of the experiment.

After completion of the first scenario, one victim and seven attackers were selected from the group of participants. The victim was chosen mainly for two reasons. First, his gait exhibited distinctive visible characteristics, which was a great advantage for the impostors. Second, the victim's gait was also reasonably stable (i.e. low intra-class variation). High stability meant that the victim could provide live gait demonstrations without any significant change in the manner of walking. This was particularly important since the demonstrations could not take place all on the same day, nor could the video shootings.

The attackers were chosen such that their initial distances to the victim (i.e. the average distance between the victim's and the attacker's normal gait) consisted of both high and low values. Also, a reasonably stable gait was a requirement. Simply put, the attackers represented "normal" people, with no specific advantages or disadvantages on average.

The **short-term hostile scenario** is the second part of the experiment, where a group of six attackers attempted to imitate the victim. After each imitation attempt, the attacker was able to review his or her own gait on a video recording, and further improve the mimicking by comparing it to a video of the victim. Statistical results were also available. One session took about an hour, and five sessions were held for each attacker.

Finally, the **long-term hostile scenario** had only one attacker. He was selected mainly because he was the only attacker who had enough time for this scenario. This attacker trained to mimic the victim in the same way as the other attackers did, but in this case there were more sessions, over a longer period of time. After conducting the previous scenarios of the experiment, this part of the research had a more solid foundation of experience, and was planned to endure for six weeks.

All experiments were conducted indoors, to eliminate some noise factors (i.e. climate conditions). The same floor material was used in every setting. For the friendly scenario, 50 test subjects from Gjøvik University Collegewere used, 14 of these female. The test subjects were all between 19 and 66 years old with a mean value of 23.0 years. The attackers and the victim for the other two scenarios were all selected from the friendly test group, and the attackers are referred to by numbering: A01, A03, A04, A18, A21, A38 and A41.

Each of the 50 participants in the first part of the experiment walked 10 times, where each walk resulted in a dataset containing approximately 20-25 cycles. These datasets were then converted into average cycles using the described CLM method, resulting in a total of 500 average cycles that could be used in the analysis of the friendly scenario. In the short-term hostile scenario, all 6 attackers participated in 5 sessions, where each session was split in 3 and in each of the 3 parts the attacker walked 4-6 times. The reason a session was split into 3 parts was that the attacker could evaluate the performance in between and try to improve using that feedback. In this way each attacker provided at least 60 attempts on mimicking the victim. The long term scenario was setup in principle in the same way, but more sessions were done. In this scenario there was just one attacker who provided over 160 mimicking attempts.

5 Data Analysis

5.1 Statistical Tools

The main objective of the research was to identify learning - in other words a systematic change in the performance of the attackers over time. A regression analysis was conducted for the purpose of finding a learning curve. It would not be interesting to find an expression fitting the observations perfectly (e.g. using splines), but rather something that can identify a trend, and tell us whether the mimicking results are improving or worsening as a result of the training.

During this research, no work related to regression of gait biometrics was found in the open literature. The nature of learning may justify the use of a regression model based on an exponential curve. This curve can potentially describe quick learning in the beginning, followed by diminishing progress. The chosen model is given by:

$$Y(X) = \beta_1 + \beta_2 e^{\frac{\beta_3}{X}}, \tag{2}$$

where β_1, β_2 and β_3 are the regression parameters, and $Y(X)$ is the estimate of observation X. This model removes natural fluctuations in performance, and also converges. The latter property is desirable because it corresponds to the fact that a learning process normally has a limit to its effect.

Further, a second regression was conducted, this time on the residuals from the model in Equation (2). This is helpful because systematic change in the residual approximations would imply that the original model was a bad choice. A simple linear model was chosen for this purpose:

$$R(X) = \lambda_1 + \lambda_2 X, \tag{3}$$

where λ_1 and λ_2 are constants, the regression parameters, and $R(X)$ is the residual estimate of observation X. A hypothesis test was conducted to see whether λ_2 was significant or not, with the null $H_0 : \lambda_2 = 0$. With a failure to reject H_0, Equation (2) is validated in terms of residual regression.

Finally, the certainty of the regression parameters β_1, β_2 and β_3 were determined by a 95% confidence interval. This gave a window of a certain range, in which we can be 95% certain we will find the subject parameter [21].

5.2 The Plateau

As the experiment was conducted, attackers felt that despite that they sometimes improved performance, they found it very hard to improve beyond a certain point. This observation was considered very significant, and therefore given a name; a *plateau*.

A plateau can be defined as "a state or period of little or no change following a period of activity or progress" [7], so on a learning curve it would correspond to the curve flattening out. Hence, observations concentrate around some value on the Y-axis. If exactly one plateau exists for each individual, then the success of an attacker is **predetermined** - the plateau has to lie below or near the acceptance threshold for an impostor to ever be able to succeed. How near it has to be depends on the variance in the data, as deviations potentially can be successful attacks.

The data in this experiment is not sufficient to make final conclusions on how the participants would be affected by an even more extensive training program. The uncertainty of the future is one of the reasons why the name "plateau" was chosen. If a temporary plateau is reached, and the performance later increases due to an extended training period, the term still makes sense. In this case one can imagine several plateaus belonging to the same performance plot. This situation should generate interesting questions, such as how to break through the different plateaus.

Intuitively plateaus can be identified by looking at points of resistance, average values and converging curves. Coefficients from fitted "trend lines" can also be put to use for this purpose. Still, the most scientific way to find the plateau would be to look for a mathematical *limit*. The limit of a function tells us exactly where it is heading when x goes to infinity.

What separates the plateau from a mathematical limit should be addressed. While the limit is a purely mathematical concept that may or may not properly illustrate learning as we know it, the plateau opens for more human-like function behavior. The main difference is that a if a function exhibits a limit,

only one such limit can exist. If only one plateau exists for an attacker, then the plateau and limit are identical. However, in the above it was suggested that an observed set of data points could exhibit several plateaus, and it would be valuable to study how each one could be "broken" if this was the case.

If several plateaus exist for an attacker, and the attacker is improving over time, then the lowest one will equal the limit. It is important to note that when statistical distance metrics are used, lower score means less separability, or difference, and higher mimicking performance. Hence, if the attacker is worsening his performance, the highest plateau equals the limit. In general we can refer to this as the final plateau.

Final (and single) plateaus are identified by taking the mathematical limit of the learning curve:

$$\rho_{nm} = \lim_{x \to \infty} Y_{nm}(X), \tag{4}$$

where where ρ_{nm} is the plateau of participant nm (e.g. 01 for A01), and Y is the learning curve of that participant.

6 Results

6.1 System Performance

The Decision Error Tradeoff (DET) curve shows system performance in terms of different thresholds, and the relationship between False Match Rate (FMR) and False Non-Match Rate (FNMR) in the system [12]. The curve is constructed by performing pairwise comparisons between all 500 templates enrolled in the system. The left part of Figure 6 shows the DET curve for the mimicking experiment. The Equal Error Rate (EER) is 6.2%, corresponding to an acceptance threshold of $T = 8.6449$. Since performance optimization was not the objective of this reserch, these results were indeed good enough to proceed. Also, by not lowering the EER further the attackers had an easier task, which is a scientific advantage in this case.

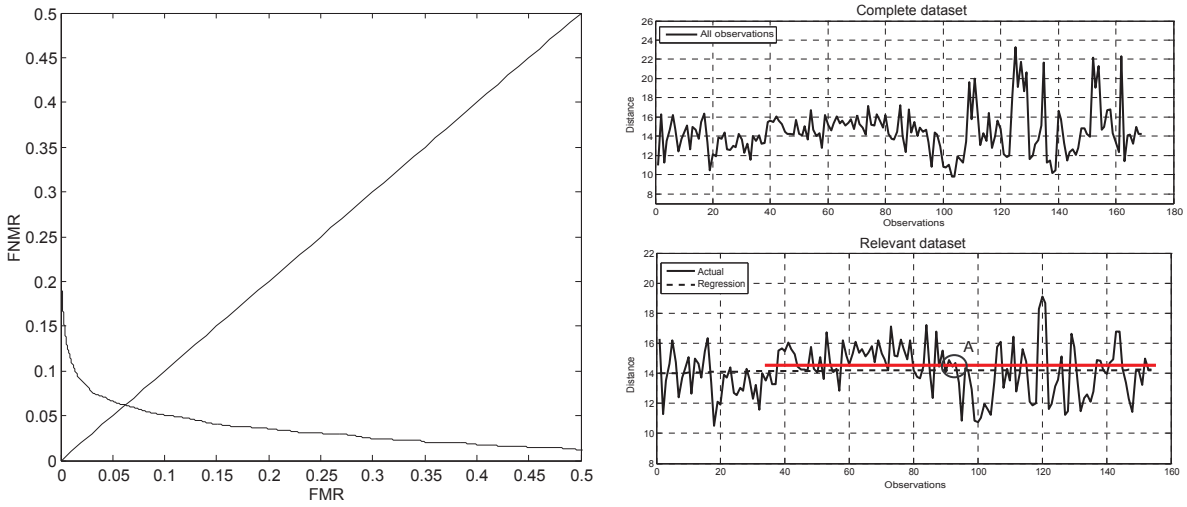

Fig. 6. Left: DET curve from the friendly scenario, EER = 6.2%. Right: Long-term attacker A01 regression analysis, with a plateau superimposed in the bottom right plot.

6.2 Mimicking Performance

As mimicking results are measured using DTW, downward sloping regression curves indicate improving results. Two identical gait samples yield a mimicking score of zero. The threshold of acceptance in this

research is at $T = 8.6449$, so for the attacker to succeed he or she would have to exhibit a learning curve converging to a point below this value. For a quick overview; four out of seven attackers worsened their performance during training (A01, A04, A21 and A38), in other words - they were better off when they did not attempt to mimic at all. None of them were near the acceptance threshold, even by the most optimistic forecasts. Their plateaus were identified at $14.2550, 14.8051, 13.1676$ and 13.3554, respectively. An extract of individual attack performances will be presented in this section.

A01 makes the first example, provided in the right part of Figure 6. The dotted regression line is given by $Y_{01}(X) = 13.9962 + 0.2588e^{\frac{-19.8894}{X}}$, representing the learning curve. In this case the curve is rising, indicating worsening performance. Using confidence intervals for the parameters β_1 and β_2, we can calculate a window of 95% certainty where the attacker is heading over time. The regression line of A01 converges to $\rho_{01} = 14.2550$ (a plateau), significantly higher than the acceptance threshold. The 95% confidence interval of this particular attacker is $[11.7599; 16.7501]$, so not even the most optimistic forecast yields a sufficiently low result for this attacker.

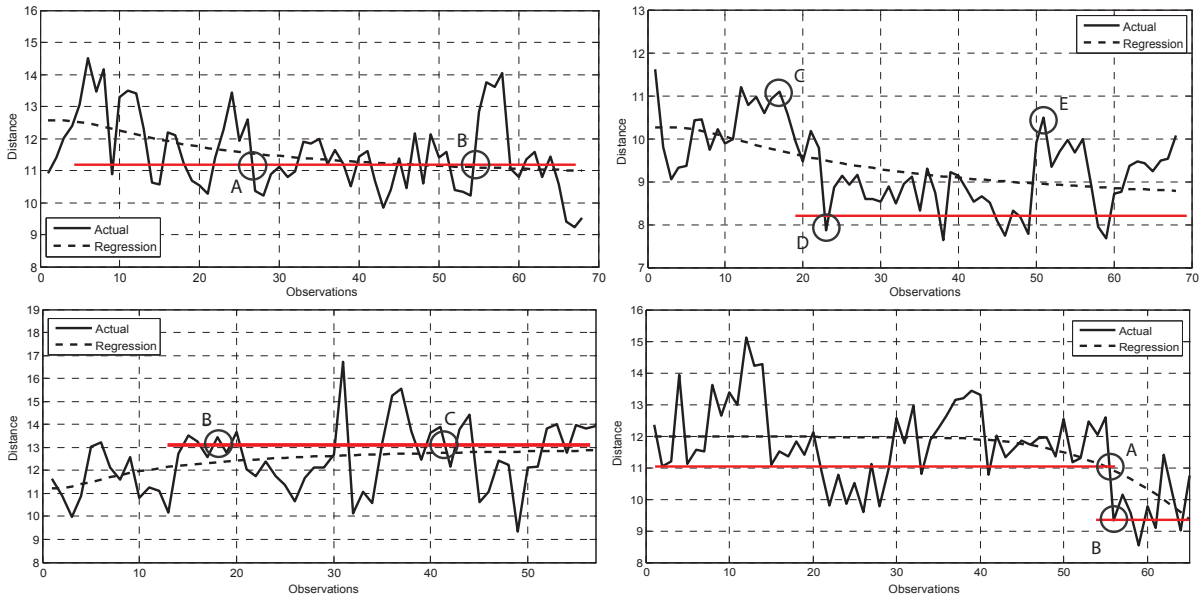

Fig. 7. Regression of mimicking performance, with plateaus sumperimposed. Top left to right: A03 and A18. Bottom left to right: A21 and A41.

A03 is one of the two attackers that exhibit a downward sloping learning curve, indicating improving performance. A03's performance was fluctuating around his plateau from the very beginning. Looking at his performance plot in the upper left of Figure 7 it can be seen that it is mostly his variance that changes. Very large fluctuations can be seen in the beginning, while at point A the results stabilize. A03 learned to focus on the particular characteristics of his gait that gave the best results, concentrating his result values around the plateau. During the last 10 - 15 attempts, starting at point B in A03's performance plot, he decided to introduce more changes. This resulted in strong fluctuations in both directions, but never produced stable improvement. A03 reported difficulty in combining gait characteristics, and felt that he ended up walking in an unnatural way.

A18 is the second and last attacker with an improving learning curve. For another numeric example; A18's regression curve is defined as $Y_{18}(X) = 10.2662 - 2.0549e^{\frac{-22.3333}{X}}$. His learning curve converges to

$\rho_{18} = 8.2113$, which makes him the only successful impostor in the experiment. A03's curve converges to a limit above the acceptance threshold, $\rho_{03} = 10.5176$.

Although A18 did improve his performance, this improvement is mainly occurring between point C and D, seen in the upper right plot of Figure 7. After point D, the decreasing values seem to meet resistance, a performance boundary. The mathematical limit of the regression curve confirms diminishing improvement. A single plateau is found at the limit, 8.2113, with a 95% confidence interval of $[6.6288 < \rho_{18} < 9.7939]$.

When A18 realized that he had problems improving further, he made some radical changes to his gait. This made his walk mechanical and uneven, and eliminated some previously adopted gait characteristics. In the plot this can be seen, starting at point E with a significant increase, followed by high fluctuations and instability for the rest of the training. Such observations were made also for other attackers - significant changes of the gait seemed to neutralize previous improvement and acquired skill.

A01 found an unnatural walk that sometimes gave better results compared to what his plateau suggested. It turned out, even if that particular walk was a better way of mimicking, he could not stabilize it. This can be seen in his plot at the bottom left in Figure 6, starting at point A. The variance increased and the attacker lost control of the previously improved gait characteristics, and the plateau kept forcing A01's performance back.

A21 was the attacker with the least control of his gait, shown at the bottom left of Figure 7. Most of the time he experienced problems when trying to change characteristics of his walking, and the reader may verify this looking at the high variation in his results. His worsening performance did sometimes stabilize around the plateau, for instance around point B and C in his performance plot.

There were significant differences between the participants in terms of *how* they reached their plateau. The three attackers A03 and A18 changed the most during the training, producing steeply sloped learning curves. Obviously this does not necessarily mean they increased their skills, just that their original score was far away from the plateau. Other attackers, like the long-term attacker A01, initially got results very close to their plateaus, and thus found it a lot harder to improve beyond that point.

The results show that it is possible to slightly improve the performance of gait mimicking using training and feedback. By experimenting with small and large changes of gait characteristics, two attackers did move somewhat closer to the victim. However, the performance increase shown is very limited. It was clear that the attackers met their natural boundaries and had huge problems moving on from that point. This indicates that even if you can train to adopt certain characteristics of your victim, the outcome of your attempt is predetermined by your plateau. If it lies too high, you will never be able to mimic that person. And most attackers actually experienced worsening performance, converging towards a plateau far above the acceptance threshold.

6.3 Breaking Through

A41 exhibits an ill-defined regression curve with no practical interpretation. The attacker achieves no improvement on average, but the last attempts are somewhat better than the first ones. This improvement does not endure long enough for the regression model to capture any meaningful effect, hence the regression curve makes a sudden decline and converges to negative values with no real meaning. It is useful to look at the possible existence of several plateaus simultaneously, challenging the assumption of a predetermined attack outcome. The regression error can in this case be seen as an indication of a plateau break. Looking at the plot for A41 in the bottom right part of Figure 7, two plateaus are superimposed, illustrating such a break in point A. We cannot calculate the second plateau directly due to the regression result, hence the lines are not mathematically derived and are provided for illustration only.

The reader should note that the multi-plateau observations here are uncertain. Not enough data is available to support any claims on several plateaus, and the results have very high fluctuations in the areas of possible plateau breaks. A more plausible explanation would be that the "real" plateau is the second, or final plateau, and that the first is caused by noise from poor mimicking. Furthermore, the research also indicated that the assumed new plateau was very hard to reach, and that the degree of learning was inadequate to pose a real threat. Even with many plateaus, reaching the second requires extreme effort, and any proceeding plateaus are likely to be harder to reach than the one before.

7 Conclusion

We tested the security strength of gait biometrics against imitation. A software tool for gait recognition was successfully designed and implemented, and the choices of methods were rooted in the current technological state of the art. The software was based on analyzing acceleration during walk, relative to the users hip in three dimensions.

An experiment consisting of three scenarios was conducted, where the ultimate goal was to train participants to be able to mimic the gait of a preselected victim. The research intended to determine whether or not the mimicking skills of the participants improved over time. An EER of 6.2% was achieved with 50 enrolled participants, which is adequate considered the maturity level of gait biometrics. The rest of the experiment was "hostile", with seven attackers training to imitate the gait of the same victim.

The participants did not show a significant improvement overall, and put in short, learning was not present. A regression analysis was conducted in order to establish this fact - most learning curves were sloping upwards, indicating worsening performance. Further this paper showed that the attackers hit a natural boundary that prevented them from improving their performance beyond a certain point. The effect of this phenomenon was striking, and given the name *plateau*.

It was found that training had little or no effect on the plateau, and it seems to be a physiologically predetermined boundary to every individual's mimicking performance. If only one such plateau exists, then it is mathematically the same as a limit, in essence - the value to which the learning curve converges. Natural fluctuations will be present, but the average results will approach the plateau over time.

The main finding of this research is that the attackers, despite exhibiting varying skills, hardly learned anything at all. If one successfully adopted gait characteristic improved an attacker's performance, we were likely to see other characteristics worsen in a chain-like effect. As with all biometrics, circumvention makes the system less secure. Almost all biometrics can up to some level be circumvented. We have shown in this paper that mimicking gait is a highly non-trivial task.

8 Future Research

It should be noted that the findings in this paper cannot necessarily be generalized to apply to other analysis methods. The results here applies to the combination of methods and configurations presented in this report, and one cannot be sure that the results would look the same, say, in the frequency domain. However, a lot of the difficulties in gait mimicking are likely to be physiologically rooted, and thus it is reasonable to assume that the indicators are relevant in other contexts. Further research should involve similar experiments, using various configurations, more training and different methods.

Other aspects of mimicking could also be analyzed - like threats through cooperation. Cooperation in gait mimicking essentially means that two people try to walk like each other, and then maybe "meet in the middle". Hence, one person could enroll walking somewhat like a different person and, if successful, they could both authenticate with the same template.

It would be beneficial to try to identify so-called sheep and wolf characteristics within gait biometrics. Some people may be easier targets for imitation, "sheep", and some people may be better at impersonating than others, "wolves". Further, such (dis)advantages could be genetically determined, and these issues together can form entire new lines of research within gait mimicking.

On the field of gait biometrics in general there is a lot of work to do. The performance of gait recognition systems are not generally competitive to other biometrics at the time of writing, so the invention of new methods, and further development of the existing methods is necessary.

References

1. Heikki J. Ailisto, Mikko Lindholm, Jani Mantyjarvi, Elena Vildjiounaite, and Satu-Marja Makela. Identifying people from gait pattern with accelerometers. *In Biometric Technology for Human Identification II. Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, 2005.
2. Heikki J. Ailisto, Mikko Lindholm, Jani Mantyjarvi, Elena Vildjiounaite, and Satu-Marja Makela. Identifying users of portable devices from gait pattern with accelerometers. *In Proceedings of SPIE*, 5779, 2005.

3. Tor Erik Buvarp. Hip movement based authentication - how will imitation affect the results? Master's thesis, Gjøvik University College - Department of Computer Science and Media Technology, 2006.
4. N.L. Clarke and S.M. Furnell. Authentication of users on mobile telephones a survey of attitudes and practices. *Computers & Security*, 2005.
5. N.L. Clarke and S.M. Furnell. Mobile phone theft, plastic card and identity fraud: Findings from the 2005/06 british crime survey. *http://www.homeoffice.gov.uk/rds/pdfs07/hosb1007.pdf*, Last visit: 15.04.2008.
6. R. Clarke. Biometrics in airports how to, and how not to, stop mahommed atta and friends. Available online at http://www.anu.edu.au/people/Roger.Clarke/DV/BioAirports.html, 2003.
7. Oxford Dictionaries. *Compact Oxford English Dictionary of Current English*. 3rd edition, 2005.
8. B. Dukic and M. Katic. m-order - payment model via sms within the m-banking. *In 27th International Conference on Information Technology Interfaces*, 2005.
9. Davrondzhon Gafurov. *Performance and Security Analysis of Gait-based User Authentication*. PhD thesis, University of Oslo, 2008.
10. Davrondzhon Gafurov, Einar Snekkenes, and Patrick Bours. Spoof attacks on gait authentication system. *Special Issue on Human Detection and Recognition*, 2007.
11. Karen Harmel and Laura Spadanuta. Disney world scans fingerprint details of park visitors. The Boston Globe, September 3rd, 2006.
12. Kjetil Holien. Gait recognition under non-standard circumstances. Master's thesis, Gjøvik University College - Department of Computer Science and Media Technology, 2008.
13. Anil K. Jain, Patrick Flynn, and Arun A. Ross. *Handbook of Biometrics*, volume 556. Springer US, 2008.
14. Eamonn J. Keogh and Michael J. Pazzani. Derivative dynamic timewarping. *In in First SIAM International Conference on Data Mining, (Chicago, IL, 2001)*, 2001.
15. Stacy J. Morris. A shoe-integrated sensor system for wireless gait analysis and real-time therapeutic feedback. *PhD Thesis, Harvard University - MIT Division of Health Sciences and Technology*, 2004.
16. Torkjel SØndrol. Using the human gait for authentication. Master's thesis, Gjøvik University College - Department of Computer Science and Media Technology, 2005.
17. S.A. Nixon and E.H. Adelson. Analylzing gait with spatiotemporal surfaces. *In proceedings of IEEE Workshop on Non-Rigid Motion*, 1994.
18. U.S. Department of State. Safety and security of u.s. borders/biometrics. State official online information, 2008.
19. Key Pousttchi and Martin Schurig. Assessment of todays mobile banking applications from the view of customer requirements. *In 37th Annual Hawaii International Conference on System Sciences (HICSS04)*, 2004.
20. Nalini K. Ratha, Jonathan H. Connell, and Ruud M. Bolle. An analysis of minutiae matching strength. *IBM Thomas J. Watson Research Center*, 2001.
21. Michael Smithson. *Confidence Intervals, in the Series of Quantitative Applications in the Social Sciences*. SAGE Publications Ltd, 2003.
22. Øyvind Stang. Gait analysis: Is it easy to learn to walk like someone else? Master's thesis, Gjøvik University College - Department of Computer Science and Media Technology, 2007.