Performance Evaluation of Gaussian Mixture Models for Inertial Sensor-based Gait Biometrics

Conference Paper · December 2018

1 author:

Margit Antal
Sapientia Hungarian University of Transylvania

Performance Evaluation of Gaussian Mixture Models for Inertial Sensor-based Gait Biometrics

Szilárd Nemes
MSc Student at Babes-Bolyai University
Faculty of Mathematics and Informatics
Cluj-Napoca, Romania
Email: nemesszili@gmail.com

Margit Antal
Sapientia University
Faculty of Technical and Human Sciences
Department of Mathematics and Informatics
Tirgu Mures, Romania
Email: manyi@ms.sapientia.ro

***Abstract*—Portable sensor-based gait biometrics has been actively researched in recent years. Generative models are rarely examined as part of gait recognition systems and there are several open questions with regards to machine learning approaches in this context. This paper shows that cross-day evaluations and the selection method of impostors influence system performance the most, and non-adapted Gaussian Mixture Models perform slightly better than adapted ones.**

I. INTRODUCTION

Human gait is defined as a person's way of walking. Murray et al. [1] established that this trait is unique to each individual, therefore it can serve as basis for user authentication. As summarized by Gafurov [2], there are three unobtrusive approaches for capturing gait patterns and related features: vision-based, floor-based and portable sensor-based. The latter solution is the latest to be explored in this context, first applied by the VTT Technical Research Center of Finland [3]. With the appearance of inertial sensors in mobile devices, research on gait recognition using data obtained from such sensors has also increased.

From an architectural point of view, a gait recognition system is a typical biometric system with the following components, as seen in figure 1: signal acquisition and preprocessing, segmentation, feature extraction and recognition. Two inertial sensors, the accelerometer and the gyroscope are used for data collection. These sensors can be used either as stand-alone, fixed to various parts of the body, or can be embedded in portable devices, such as smartphones.

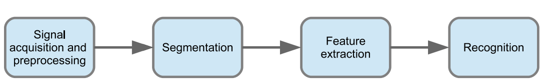

Fig. 1. Biometric system pipeline

Preprocessing usually consists of filtering and noise removal methods, as well as gait activity detection. There are two main types of segmentation methods: (i) frame-based, when the signal is split into fixed-length segments, and (ii) cycle-based, when the endpoints of gait cycles are detected. A gait cycle is defined as the time interval starting when one foot touches the ground and ending when the same foot touches the ground again [4].

Gait recognition systems may use the preprocessed data directly for comparisons (e.g. using Dynamic Time Warping). In this case, the explicit feature extraction step may even be missing. Usually time-domain, frequency-domain or both types of features are extracted from segments. Recently, deep learning methods have been used for feature extraction. Neverova [5] used a convolutional recurrent neural network (CRNN), while Gadaleta and Rossi [6] also opted for deep convolutional neural networks (CNN) for feature extraction.

The recognition component can be either template-based or machine-learning based. The machine-learning approach can be further divided into using discriminative and generative models. A distinctive feature of discriminative models is their focus on defining the decision boundary for separating the examined classes and they process exclusively labeled data. Generative models, on the other hand, use both types of data. Labeled data are used for the purpose of training probabilistic models for every class under examination, while unlabeled data are used to train the Universal Background Model (UBM), which is used to represent general behavior and, in an ideal case, should be subject independent. In case of generative models, the decision boundary is only indirectly present: processed data points are assigned to the most proximal model and their labels set accordingly.

The majority of papers that opt for the machine-learning approach use discriminative models (see the review paper by Sprager and Juric [4]). This review mentions only two papers that used Gaussian mixtures generative models (GMM) for the recognition component. Lu et al. [7] described a gait verification system for mobile phones without any assumption of body placement or device orientation. They demonstrated on a small dataset (47 subjects for UBM training, 12 subjects for evaluating supervised training) that adapting user models from a UBM has low computational overhead and it can run in real time on an Android smartphone. Zhong et al. [8] applied i-vector technology developed for speaker recognition [9] to the problem of gait recognition. This technology also uses UBM by GMM. i-vector technology was applied to a large dataset (Osaka Univ. gait dataset containing 744 subjects) and the best result of 7.1% Equal Error Rate (EER) was reported by fusing accelerometer and gyroscope data.

Recently, generative models were used by another two studies: [10], [5]. Muaaz and Mayrhofer [10] used time and frequency-based statistical features with adapted Gaussian mixtures (they adapted from UBM all three types of model parameters - means, covariance matrices and component weights). However, Neverova in her PhD thesis [5] used deep learning for feature extraction, then adapted Gaussian mixtures for recognition (only the mean parameters were adapted from the UBM).

In this paper, we propose to use this less widely used modeling approach for gait biometrics. We investigated the following research questions with regards to Gaussian Mixture Models for gait biometrics:

- Is cycle detection a compulsory component of a gait recognition system based on statistical features?
- How is performance influenced by using several step cycles?
- Does a UBM adapted GMM perform better than a classical GMM?
- How is authentication system performance affected by impostor samples selection?

Our research is reproducible: the ZJU-GaitAccel dataset is publicly available[1], and our results can be replicated with the code available on GitHub[2].

The rest of this paper is structured as follows: the details of our methods and materials used are presented in Section 2. The next section presents and discusses the achieved results. Section 4 presents the conclusions of our paper.

II. Materials and Methods

A. Dataset

The ZJU-GaitAcc dataset contains walking data captured by 5 accelerometers (Wii Remote, ADXL330) fastened at five different body locations [11]. The data obtained from the accelerometers were resampled with a frequency of 100 Hz. The dataset contains walking data of 175 subjects recorded in two sessions. 22 subjects participated in only one session (their recordings were grouped into a separate session, named `session0`), and 153 subjects participated in both sessions (`session1` and `session2`). There are 6 records per subject in one session. The time interval between the two sessions is at least one week and at most half a year. The subjects were tasked with walking on a 20m long floor. The recording lengths are 7-15s containing 7-14 full step cycles. One advantage of this dataset is that the authors also provided a manual annotation of the step cycles. In this paper, we used only the data from a single accelerometer, the one fastened to the right hip.

B. Feature Extraction

Let $a_x(t), a_y(t), a_z(t)$ denote the three dimensional time series obtained from the accelerometer, which was sampled with a frequency of 100 Hz. The raw data were segmented into cycles or fixed-length segments (in our case

[1]http://www.cs.zju.edu.cn/ gpan/database/gaitacc.html
[2]https://github.com/nemesszili/gaitgmm

TABLE I
Time-domain features. #nf – number of features.

Name	Explanation	#nf
$min_{ax}, min_{ay}, min_{az}, min_{am}$	minimum	4
$avg_{ax}, avg_{ay}, avg_{az}, avg_{am}$	average	4
$std_{ax}, std_{ay}, std_{az}, std_{am}$	standard deviation	4
$ad_{ax}, ad_{ay}, ad_{az}, ad_{am}$	mean absolute deviation around the mean	4
$zcr_{ax}, zcr_{ay}, zcr_{az}$	zero crossing rate	3
$bin_{ax}(i), i = 1 \dots 10$ $bin_{ay}(i), i = 1 \dots 10$ $bin_{az}(i), i = 1 \dots 10$ $bin_{am}(i), i = 1 \dots 10$	10-bin histograms	40

128 samples). Then, the magnitude was computed $a_m(t) = \sqrt{a_x(t)^2 + a_y(t)^2 + a_z(t)^2}$, therefore we obtained four time series: $a_x(t), a_y(t), a_z(t), a_m(t)$. We computed the statistical features from each segment shown in table I.

Figure 2 shows the differences between the two sessions of the first 4 subjects of the dataset. We selected to plot the most discriminative two features.

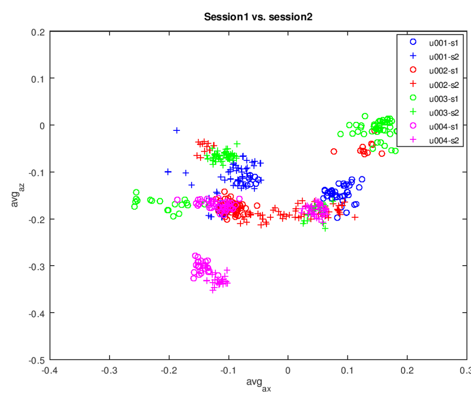

Fig. 2. `session1` vs. `session2` for 4 subjects of the dataset.

C. Adapted Gaussian Mixture Models

1) Model Adaptation: A GMM is composed from a fixed number (M) of multivariate Gaussian distributions with parameters $\Theta = \{w_i, \mu_i, \Sigma_i\}, i = 1, \dots, M$, where w_i are the mixture weights ($\sum_{i=1}^{M} w_i = 1$), and μ_i and Σ_i are the mean and covariance matrix parameters of each multivariate Gaussian distribution. Let $x \in R^D$, where D is the dimension of the feature vector. The probability density of the GMM defined by parameters Θ is defined as:

$$p(x|\Theta) = \sum_{i=1}^{M} w_i f(x|\mu_i, \Sigma_i) \quad (1)$$

$$\text{where} \quad f(x|\mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_i|}} e^{\frac{(x-\mu_i)' \Sigma_i^{-1} (x-\mu_i)}{2}}. \quad (2)$$

We created a GMM from a set of background users. This model is called Universal Background Model (UBM) and it was introduced for speaker verification [12]. The parameters of the UBM were estimated from the data of the background users by applying the expectation-maximization (EM) algorithm. Each user model was adapted from the UBM using maximum a posteriori (MAP) adaptation of the UBM [12]. MAP-adapted UBM models for gait biometrics were already used by Muaaz and Mayrhofer [10] and Neverova [5]. Muaaz and Mayrhofer adapted all model parameters, while Neverova adapted only the means. As only mean vector adaptation performed better (and it also means less computational overhead), we opted for this solution.

The mean vector adaptation was implemented as follows. Given a set of $X = \{x_1, x_2, \ldots, x_T\}, x_i \in R^D$ enrollment samples for a given user, we update the parameters of each mean as follows:

$$E_i(X) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t)x_t \tag{3}$$

where

$$n_i = \sum_{t=1}^{T} Pr(i|x_t), Pr(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} \tag{4}$$

Finally, the mean parameters of the client model were updated as follows:

$$\hat{\mu_i} = \alpha_i E_i(X) + (1 - \alpha_i)\mu_i, \text{where} \quad \alpha_i = \frac{n_i}{n_i + r}, \tag{5}$$

where r is a relevance factor, which is usually estimated empirically. The best results were obtained for $r = 0$, which means that the mean vectors were estimated entirely from the client data ($\alpha_i = 1$).

2) Scoring: Let X denote a set of test samples. The log-likelihood of client model samples, as well as UBM model samples were computed, and the difference is denoted as $score(X)$. Based on this score, the samples were either accepted as authentic or rejected:

$$score(X) = \log p(X|\Theta_{client}) - \log p(X|\Theta_{UBM}). \tag{6}$$

III. EXPERIMENTS

A. Evaluation Protocol

Performances are reported using common metrics in biometric systems, such as Equal Error Rate (EER), Area Under Curve (AUC) and Receiver Operating Curves (ROC) [13].

We used two types of segments: (i) step cycles - using the endpoints annotations of the dataset, (ii) fixed length segments - in this case we used 128 sample-length segments, which correspond to 1.28s of data, which is very close to the average step cycle length. Statistical features were extracted from the segments and then forwarded to the recognition component. The first and the last segments of the recordings were ignored in both type of segmentations.

The system is evaluated in two ways: (i) same-day, and (ii) cross-day. In the case of same-day evaluation, we used both the training and the test data from the same session; the first half of the data was used for training and the second half of the data for testing. However, in the case of cross-day evaluation, half of the data from `session1` was used for training and the data from `session2` were used for testing (positive scores). Data from `session0` were used for UBM training (the data from the first 11 users: $u001 - u011$) and for testing samples (negative scores) for user-specific GMM evaluations (the data from the last 11 users: $u012 - u022$). Evaluations of user models were performed by always using the same number of positive and negative samples. As the number of positive samples is fixed (half of the `session1` data for a given user), we always randomly selected (with uniform distribution) the same number of negative samples from `session0` (the data from users: $u012 - u022$). The baseline is defined as the evaluations performed on a single segment, which can either be one cycle or 128 samples. Apart from the measurements for answering the first research question, we used step cycle-based segments. The protocol for the baseline evaluation is depicted in table II.

TABLE II
BASELINE EVALUATION PROTOCOL.

	same-day	**cross-day**
Training	`session1` (50% - first half)	`session1` (50% - first half)
Testing positive	`session1` (50% - second half)	`session2` (all samples)
Testing negative	randomly selected `session0`: $u012 - u022$ numPosTest = numNegTest	randomly selected `session0`: $u012 - u022$ numPosTest = numNegTest

TABLE III
BASELINE RESULTS (ONE SEGMENT: 128-SAMPLES OR ONE CYCLE) FOR ONE DAY AND CROSS-DAY EVALUATIONS.

Day	**Segment**	**AUC**	**EER**
same-day	cycle	0.9412 (0.0710)	0.0871 (0.0812)
same-day	128 samples	0.9416 (0.0613)	0.0930 (0.0728)
cross-day	cycle	0.8149 (0.1990)	0.2195 (0.1913)
cross-day	128 samples	0.8183 (0.1864)	0.2225 (0.1802)

Feature extraction was performed by a Java application[3]. The rest of the evaluations were performed using Python (interpreter version 3.6) along with the following packages: scipy, sklearn, numpy, pandas and matplotlib.

B. Results

1) Question 1: Baseline results: Results for the baseline evaluation are given in table III. Metrics were computed for each of the 153 subjects of the dataset, then we report the average of the values, as well as the standard deviations in parenthesis. We see very similar performances when using the fixed length segments instead of step cycles based segments.

2) Question 2: The effect of using a sequence of step cycles: In order to increase the performance of the gait authentication system, we evaluated the system based on more consecutive gait cycles. The performances obtained for using a sequence

[3]https://github.com/nemesszili/gaitgmm/tree/javafeat

of segments for authentication is presented in table IV. Our result of 1.58% EER (5 step cycles, same-day) is comparable to the recent results of Gadaleta and Rossi [6], who obtained 0.15% EER by using 5 step cycles. It is important to note that the realistic (cross-day) evaluation produced larger errors, even when using a sequence of 10 step cycles.

TABLE IV
RESULTS FOR CYCLE-BASED EVALUATIONS USING A SEQUENCE OF STEP CYCLES.

	same-day		cross-day	
#cycles	**AUC**	**EER**	**AUC**	**EER**
1	0.9412	0.0871	0.8149	0.2195
2	0.9656	0.0607	0.8412	0.1908
3	0.9794	0.0417	0.8602	0.1675
4	0.9878	0.0247	0.8737	0.1521
5	0.9919	0.0158	0.8842	0.1359
6	0.9949	0.0091	0.8920	0.1237
7	0.9970	0.0038	0.8970	0.1167
8	0.9977	0.0032	0.8987	0.1126
9	0.9980	0.0025	0.8993	0.1101
10	0.9977	0.0029	0.9012	0.1074

In figure 3, we can see that increasing the number of step cycles results in higher AUC values for both the same-day and cross-day evaluations.

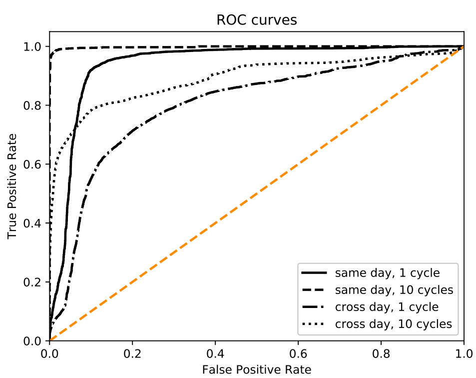

Fig. 3. ROC curves for different type of performance evaluations.

3) Question 3: GMM vs. adapted GMM: Although the literature recommends training adapted Gaussian Mixture Models, we investigated classical GMMs trained with Expectation Maximization algorithm. We used 8 mixture components in both classical and adapted GMM variants. Comparative results are depicted in table V.

4) Question 4: The effect of negative (impostor) test samples selection: According to Pisani et al. [14], there are two sources to obtain impostor (negative) samples: the registered users in the dataset (excluding the genuine one) or a separate dataset of unregistered users. In our case, we consider the users in `session0` as a separate dataset. Both cases were investigated: in the first case (baseline setting), we designated registered users from `session0` to be impostors (users $u012 - u022$), and in the second case, we selected these samples from currently unregistered users (in our case `session1`). In both cases, the samples were selected randomly and the system was evaluated using the same number of positive and negative test samples. In order to have the most realistic setting possible, recognition was based on the evaluation of 10 cycles in cross-day settings. The results are depicted in table VI.

TABLE V
COMPARISON OF GMM AND ADAPTED GMM METHODS.

		same-day		**cross-day**	
GMM type	**#cycles**	**AUC**	**EER**	**AUC**	**EER**
Classical	1	0.9725	0.0697	0.8338	0.2140
Adapted	1	0.9412	0.0871	0.8149	0.2195
Classical	5	0.9981	0.0061	0.9279	0.0919
Adapted	5	0.9919	0.0158	0.8842	0.1359
Classical	10	0.9996	0.0007	0.9504	0.0556
Adapted	10	0.9977	0.0029	0.9012	0.1074

TABLE VI
THE EFFECT OF NEGATIVE TEST SAMPLES SELECTION.

Negatives	**GMM type**	**AUC**	**EER**
Baseline	Classical	0.9504	0.0556
Other users	Classical	0.8932	0.1162
Baseline	Adapted	0.9012	0.1074
Other users	Adapted	0.8772	0.1347

C. Other studies on the ZJU-GaitAccel dataset

We note that our best results of 0.07% EER (same-day) and 5.56% EER (cross-day) were obtained by using the data from only one accelerometer (hip), classical GMM and 10 step cycles (see table V), while the authors of the dataset [11] obtained 8.9% EER using the same data (having recordings that contain 10 cycles on average). It is also worth mentioning that comparing the two results is somewhat difficult, due to the fact that Zhang et al. examined their dataset having both same-day and cross-day sessions mixed, while we clearly distinguish the two scenarios. Giorgi et al. [15] used deep learning for gait recognition on the same dataset using all five accelerometers data, however, they performed only the same-day type of evaluations. They reported 1.1% False Positive rate and 2.93% False Negative Rate when using 10 gait cycles. We should remark that the evaluation protocol is not clearly described in their paper and there are also mistakes related to the interpretation of the results.

Marsico and Mecca obtained 9.26% EER on the same dataset using multiple templates (recordings containing 10 cycles on average) and Dynamic Time Warping (DTW) for signals comparison [16]. Similarly to us, they used only data from the same accelerometer.

IV. CONCLUSION

In this paper we presented a generative approach to gait biometrics. Although our approach had already been studied by other research papers, we succeeded to contribute by clarifying some important aspects of this type of biometrics.

First of all, we experimentally demonstrated that the two types of segmentations (fixed-length and exact cycle endpoints based) resulted in very similar performances (see table III). We should note that cycle detection component is usually error prone, therefore using fixed length segments is a viable alternative approach.

Secondly, we performed measurements regarding the effect of using a sequence of consecutive step cycles for authentication. Our results are in line with recent results of the literature [6]: the performance of biometric verification improves significantly when using multiple step cycles for recognition (see table IV).

Thirdly, our measurements on this particular dataset have shown that there are small improvements in performance when using classical GMM instead of adapted ones (see table V). These improvements are even more pronounced in the case of cross-day evaluations. However, the source of impostor samples had a greater effect on the system performance (see table VI).

The most significant performance degradation was measured by the cross-day evaluations compared to same-day. This is due to the fact that gait is not a stable pattern in time and it is influenced by several factors such as shoes worn or surface.

REFERENCES

[1] M. Murray, "Gait as a total pattern of movement: including a bibliography on gait," in *Am. J. Phys. Med. Rehabil. 46 (1)*, 1967, pp. 290 – 333.

[2] D. Gafurov, "A survey of biometric gait recognition: Approaches, security and challenges," in *Proceedings of the Annual Norwegian Science Conference*, 2007, pp. 19 – 21.

[3] *Identifying people from gait pattern with accelerometers*, vol. 5779, 2005. [Online]. Available: https://doi.org/10.1117/12.603331

[11] Y. Zhang, G. Pan, K. Jia, M. Lu, Y. Wang, and Z. Wu, "Accelerometer-based gait recognition by sparse representation of signature points with clusters," *IEEE Transactions on Cybernetics*, vol. 45, no. 9, pp. 1864 – 1875, Sept 2015.

[4] S. Sprager and M. B. Juric, "Inertial sensor-based gait recognition: A review." *Sensors (Basel, Switzerland)*, vol. 15, no. 9, pp. 22 089 – 22 127, 2015.

[5] N. Neverova, "Deep learning for human motion analysis. (apprentissage automatique de représentations profondes pour l'analyse du mouvement humain)," Ph.D. dissertation, INSA de Lyon, France, 2016. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01470466

[6] M. Gadaleta and M. Rossi, "Idnet: Smartphone-based gait recognition with convolutional neural networks," *Pattern Recognition*, vol. 74, pp. 25 – 37, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317303485

[7] H. Lu, J. Huang, T. Saha, and L. Nachman, "Unobtrusive gait verification for mobile phones," in *Proceedings of the 2014 ACM International Symposium on Wearable Computers*, ser. ISWC '14. New York, NY, USA: ACM, 2014, pp. 91 – 98. [Online]. Available: http://doi.acm.org/10.1145/2634317.2642868

[8] Y. Zhong and Y. Deng, "Sensor orientation invariant mobile gait biometrics," in *IEEE International Joint Conference on Biometrics*, Sept 2014, pp. 1 – 8.

[9] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The kaldi speech recognition toolkit," in *In IEEE 2011 workshop*, 2011.

[10] M. Muaaz and R. Mayrhofer, "Accelerometer based gait recognition using adapted gaussian mixture models," in *Proceedings of the 14th International Conference on Advances in Mobile Computing and Multi Media*, ser. MoMM '16. New York, NY, USA: ACM, 2016, pp. 288 – 291. [Online]. Available: http://doi.acm.org/10.1145/3007120.3007164

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19 – 41, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200499903615

[13] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861 – 874, 2006.

[14] P. H. Pisani, N. Poh, A. C. de Carvalho, and A. C. Lorena, "Score normalization applied to adaptive biometric systems," *Computers and Security*, vol. 70, pp. 565 – 580, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404817301554

[15] G. Giorgi, F. Martinelli, A. Saracino, and M. Sheikhalishahi, "Try walking in my shoes, if you can: Accurate gait recognition through deep learning," in *Computer Safety, Reliability, and Security*, S. Tonetta, E. Schoitsch, and F. Bitsch, Eds. Cham: Springer International Publishing, 2017, pp. 384 – 395.

[16] M. D. Marsico and A. Mecca, "Biometric walk recognizer," *Multimedia Tools and Applications*, vol. 76, no. 4, pp. 4713 – 4745, Feb 2017.