International Conference on Cyber Security and Computer Science (ICONCS'18),
Oct 18-20, 2018 Safranbolu, Turkey

Gait Recognition via Machine Learning

A. KEÇECİ[1], A. YILDIRAK[2], K. ÖZYAZICI[3], G. AYLUÇTARHAN[4], O. AĞBULUT[5] and İ. ZİNCİR[6]

[1] Yaşar University, İzmir/Turkey, aybuke.kececi@yasar.edu.tr
[2] Yaşar University, İzmir/Turkey, armagan.yildirak@yasar.edu.tr
[3] Yaşar University, İzmir/Turkey, kaan.ozyazici@yasar.edu.tr
[4] Yaşar University, İzmir/Turkey, gülsen.ayluctarhan@yasar.edu.tr
[5] Yaşar University, İzmir/Turkey, onur.agbulut@yasar.edu.tr
[6] Yaşar University, İzmir/Turkey, ibrahim.zincir@yasar.edu.tr

***Abstract* - The basis of biometric authentication is that each person's physical and behavioral characteristics can be accurately defined. Many authentication techniques were developed for years. Human gait recognition is one of these techniques. This article was studied on HugaDB database which is a human gait data collection for analysis and activity recognition (2017, Chereshnev and Kertesz-Farkas). Combined activity data of different people were collected in HugaDB database (2017, Chereshnev and Kertesz-Farkas). The activities are walking, running, sitting and standing (2017, Chereshnev and Kertesz-Farkas). The data were collected with devices such as wearable accelerometer and gyroscope (2017, Chereshnev and Kertesz-Farkas). Only the walking dataset of the HugaDB was used artificial neural network-based method for real-time gait analysis with the minimal number of Inertial Measurement Units (2018, Sun et al). In this paper, each person is considered as a different class because there are multiple users' gait data in the database and some machine learning algorithms have been applied to walking, running, standing and sitting data. The best algorithms are chosen from the algorithms applied to the HugaDB data and the results are shared.**

***Keywords* – machine learning, security, gait recognition, human detection**

I. INTRODUCTION

IDENTIFICATION is one of the most important aspects in security. Biometrics is one of the techniques that can be used to identify an individual. For example, fingerprint recognition is used for identifying people from each other by using their fingerprints. In addition to fingerprint recognition, biometrics can multiply with ear, vein, license plate, retina and gait recognition. Gait recognition, identifying a person's body movement, is also a technique of biometrics.

In gait analysis, a person's movement describes personal way of walking and that means it could be used for identifying a person. Gait recognition is a biometric technique that is used for identifying biological and behavioral specification. Gait recognition technology methods divide into two; first one is holistic-based method and the second one is model-based method. Holistic-based approach relies on extracting statistical features of motion-based while model-based method identifies body parts to create a 3D gait model.

In this paper, gait recognition technology was preferred as an identification system. This system is one of the distinctive attributes. Open source database HugaDB was used which consists of multi classes such as; running, walking, standing and sitting. Data were collected from a body sensor network consisting of six wearable inertial sensors located on the right and left thighs, shins and feet. In addition to that, two EMG sensors were also used on the quadriceps to measure muscle activity. At the end, 2,111,962 samples performed up to 10 hours were collected from 18 participants.

In the following sections, first we discussed the relevant work on gait recognition, then machine learning techniques that are used were explained. After this the dataset that was used was explained. Fourth, how this dataset was implemented are deliberated and then the results obtained were assessed. Finally, conclusion and future work were presented.

II. RELEVANT WORK

Sharma and Bansal proposed a human recognition system with using backpropagation neural network classifier. Firstly, the system presented in the study extract foreground object and they used gaussian mixture model for background subtraction. They used combination of width and MPEG-7 region-based shape descriptors which is a new attribute. Because these definers are more capable to express of separated parts of the silhouette. The person is defined by the measured distance between the separated parts. It is compared with the data stored in the database after extracting feature from the walking silhouette. Backpropagation neural network is used for learning and recognition. [1]

Zhang et al. produced framework of gait recognition using a Siamese neural network which can engage metric learning from a specific distance to direct the similarity metric to be small for pairs of gait from the same person, and large for pairs from different persons, to systematically extricate for tough and selective gait details for identification. They used Gait Energy Image (GEI) because of data limitation problem and utilized the K-Nearest Neighbor (KNN) for recognize the same person in stage. They worked on the OULP-C1V1-A dataset from the OU-ISIR LP gait benchmark. In this study, Siamese neural network (SiaNet.FC) compared with some methods

such as HWLD, GEI, FDF, CNN.FCI and the best result was achieved with SiaNet.FC. The results are 96.02% in Rank-1 Identification Rate and 98.31% in Rank-5 Identification Rate. [2]

Gaba and Ahuja identify people with gait analysis. In their paper, first the movement of an individual is detected, second, background subtraction is applied in order to eliminate the unnecessary information, third, feature extraction using the Hanavan's model is implemented for extracting the distinct parameters and for final step, BPNN+LDA and BPNN+MDA techniques are utilized to execute recognition. The best accuracy achieved by the system is 98.8%. [3]

Liang et al. presented a gait recognition system that uses the golden ratio segmentation method because they believe that the clothing affects the percentage of the gait recognition systems. They used CASIA-B dataset in their experiments. Their proposed system's results brought about 94.76 % in different clothing circumstances and 91.53% with bags. [4]

Wu et al. studied on human identification with gait by similarity learning and deep convolutional neural networks (CNNs). The method used for cross-view gait recognition was examined with more than one dataset. For CASIA-B, their results shows that the accuracy of identification is 94.1%. For OU-ISIR gait dataset, the results are above 98% under identical view conditions and 91% for cross-view scenarios. The final dataset USF, achieves 96.7% accuracy rate. [5]

Castro et al. studied on gait-based people identification using convolutional neural network. They used TUM-GAID as sample dataset. In order to achieve the most successful results, they chose a 'one-vs-all' linear SVM which has 98% success rate while the NN approach on PCA compressed descriptors yielded an approximate result which is 97.9%. [6]

In order to search for and analyze the differences in more than one feature on the datasets, as can be seen in previous researches machine learning algorithms were applied.

III. Machine Learning

With the increased of the usage of technology, data collection has been easier in different disciplines including medicine, business, education, security and so on. Automatic visual surveillance is of paramount importance due to security problems in recent years. Cameras as security tools provide large data sources for human recognition. Gait recognition is among the most appropriate biometric methods. Moreover, the development of open source and commercial machine learning and data mining tools enabled experts to employ machine learning and data mining algorithms to support decisions on these data collected in different fields.

Machine learning is a system that studies the structure and function of algorithms that can learn as a structural function and make estimation through data. Such algorithms work by constructing a model to perform data-based estimates and decisions from sample inputs rather than strictly following static program instructions. Also, it is a method paradigm that makes inferences from existing data using mathematical and statistical methods and makes predictions about the unknown with these inferences. There are two types of learning techniques; supervised learning and unsupervised learning.

Supervised Learning: Supervised Learning is the learning process from tagged observations. Labels teach the algorithm how to label observations. For example, within the "make money" statement in mail, it should be called spam.

Unsupervised Learning: It is the learning process from unlabeled observations. The algorithm is expected to make self-discoveries and discover invisible patterns.

In security, machine learning is used by biometrics. Biometrics is a biological data that measured. The main characteristic study is to authenticate a person. Type of biometric is not important because steps of the process are same. These steps, capture, process and comparison. In this section, we discuss some of the most well-known machine learning algorithms discussed in the related work.

A. Multiclass Classification

In machine learning, multiclass or multinomial classification is the problem of classifying instances into one of three or more classes. In the multiclass classification, each training point belongs to one of N different classes. The goal is to construct a function which, given a new data point, will correctly predict the class to which the new point belongs.

B. Binary Classification

Binary Classification is a form of supervised machine learning where we classify the elements (examples) of a given data set into two groups on the basis of a classification rule.

C. Machine Learning Algorithms Implemented

In this section, we summarize the machine learning algorithms we used in this research.

1) RIPPER

RIPPER is one of the basic and most popular algorithms. Classes are examined in increasing size and an initial set of rules for the class is brought about using cumulative reduced error the algorithm proceeds by treating all the samples of a particular perception in the training data as a class and finding a set of rules that cover all the members of that class. Consequently, it proceeds to the next class and does the same, repeating this until all classes have been covered.

2) Multilayer Perceptron

ANNs are typically organized in layers. Layers are made up of a number of interconnected nodes (neurons) which contain an activation function. Patterns are presented to the ANN via the input layer, which communicates to one or more hidden layers where the actual processing is done via a system of weighted connections. The hidden layers then link to an output layer. [7]

A Multi-Layer Perceptron (MLP) consists of one input layer, one or more LTU layers called the hidden layer, and an output layer. Other layers, except the output layer, contain the bias neuron and are fully connected to other layers.

3) Decision Tree

Decision trees have a predefined target variable. They offer a strategy from top to bottom. A decision tree is a structure that is used to divide a data set containing a large number of

records into smaller sets by applying a series of decision rules. In other words, it is a structure used by dividing large amounts of records into very small records by applying simple decision-making steps. [8]

4) Random Forest

It is aimed to increase the classification value by using more than one decision tree during the classification process. Random forest is a classification model that tries to make more accurate classification by using more than one decision tree.

5) IB1

IB1 uses the nearest neighbor classifier. It uses the standardized Euclidean distance to find the closest sample to the desired sample and makes the same class as the sample. If more than one sample has the same (smallest) distance to the test sample, the first found is used. The Euclidean distance is calculated by giving weights according to the distance from the sample in the learning set to the distance to the desired sample.

6) Bootstrap Aggregating (Bagging)

Bootstrap aggregating is called bagging, statistical classification and machine learning algorithms designed to improve the stability and accuracy used in the machine learning community is a meta-algorithm. It also reduces variance and helps prevent over-insertion. Although it is applied to the decision tree method it can be used by any means. Bagging, approximation model is a special case.

7) Classification via Regression

The linear regression approach is used for classification in this classifier. When classifying, each generated regression model is configured for each value of the class.

8) Random Tree

The Random Tree operator works exactly like the Decision Tree operator with one exception: for each split only, a random subset of attributes is available. It is recommended that you study the documentation of the Decision Tree operator for basic understanding of decision trees.

9) Naïve Bayes

The Naïve Bayes classification aims to determine the class of presented data to the system by a series of calculations defined according to the probability principles. The Naïve Bayes classification provides data that is taught to the system at a certain rate. The data submitted for teaching must have a class / category. With the probabilistic operations performed on the taught data, the new test data presented to the system is operated according to the previously obtained probability values and it is tried to determine which category of test data is given. The more number of data taught, the more accurate it is to determine the actual category of test data.

10) BayesNet

BayesNet have directed acyclic graph (DAG) which is a graphical model structure. It provides learning using various search algorithms and quantity measures. [10]

IV. Data Preparation

A. Data Preprocessing

Today's real-world data are generally prone to be large, distributed, and contain heterogeneous data sources, with noisy data, or with forgotten data or inconsistent data. Low quality data also leads to low quality mining results. Missing, excessive, repeated data may have been entered during data entry or transfer. In not well-organized related databases, the same records can be entered under different variable names. Data preprocessing is a very important starting point of data mining. Data collected in applications may be inadequate, inconsistent, or noisy.

The reasons for these are erroneous data collection tools, data entry problems, misinterpretations of users during data entry, data transmission errors, technological limitations, inconsistency in data naming or structure.

B. Data Cleaning and Transformation

Data cleaning, completion of missing data, correction of noise in order to diagnose outliers and eliminating inconsistencies in the data. There are different ways to fill in missing values for any variable. For any sample belonging to the same class, the mean of the variable can be used. For example, the average income value for customers in the same credit risk category can be used instead of missing values. Or the most appropriate value can be used based on existing data. Techniques such as regression or decision tree may be used to determine the most suitable value mentioned herein. Another problem that needs to be used for data cleaning is noisy data. Noise is the variance or random error in the measured variable. Techniques such as histograms, clustering analysis and regression can be used to diagnose noisy data.

The data is uneditable in databases where the original formats differ from another one for a variety of reasons. Data transformation is fit in the appropriate formats for data mining. Often, conversion types are used, which are called correction, merging, generalization, and normalization. One or more of these transformation types can be used when the data is converted to the appropriate format for the data mining. Data transformations aimed at transforming to formats suitable for data mining are usually done in the following five different ways. Correction ensures noisy parsing and reduction of data. Techniques such as partitioning, clustering and regression are used. Consolidation involves the summarization or merging of data. Generalization is the process of transforming low-level variables or raw data into higher-level variables. Normalization is one of the most frequently used data conversion operations.

Finally, with these steps, the subjects that have missing files in the dataset were not used in the evaluation, the unrelated columns were removed, and the files were set according to the type of file which is .arff extension to be used to make it easier to access the dataset.

V. Experiments and Evaluations

The database is used for gait analysis is HugaDB [11]. It is created by results of different experiments. Participants put on inertial sensors (accelerometer and gyroscope) and performed activities like running, walking, standing and so on. Activities

are performed and recorded at different times. At the end, samples are collected, and database is created. Originally, there are 637 files of data collected from 18 participants in the database. There are many activities performed by the participants, but not all the files are used. Some activities are not performed by all the participants or one participant performed one activity multiple times. And, there is also a "various" category in the files. The files named "various" have data of different activities. Each of the various files have different combinations of activities. Running, standing, walking and sitting are chosen to be studied in the project.

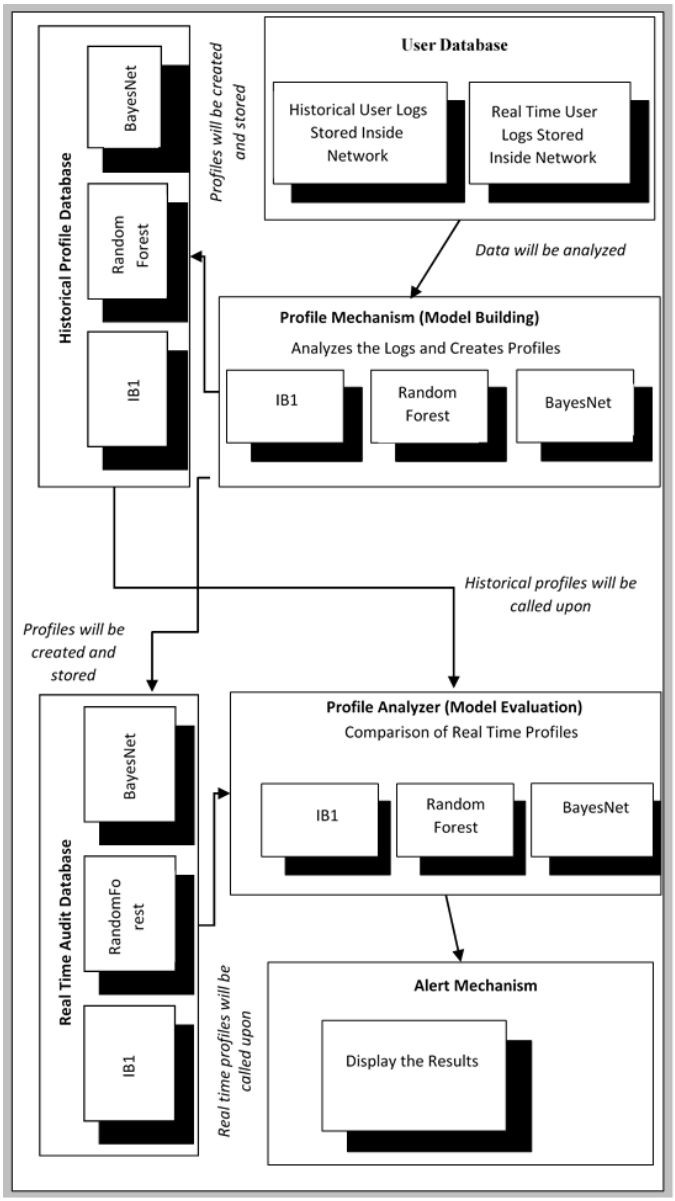

Figure 1: System Components

All files have the same data format. They are composed of 39 columns that the first 36 columns are inertial sensors, next 2 columns are EMG sensors and the last one is activity id. In the project, data remained the same. However, txt files transformed into arff files to be used in WEKA. For each activity, data of 18 participants are combined. There are 17 sitting, 16 standing, 17 walking and 4 running files. All of them are used for training.

When data preparation and editing steps are finished, 10 data mining algorithms were chosen. The WEKA framework was used to run the selected algorithms which are working with 10-fold Cross-Validation. WEKA takes N size labeled data, then it produces 10 equal sized sets. Each set is split into two groups. %90 size of labeled data is used for training and %10 size of labeled data are used for testing. The WEKA produces a classifier with an algorithm form %90 size of labeled data and applies that to the %10 size of labeled testing data for the first set. It does the same thing for the second set to 10 equal size sets and produces more classifiers. After that, it evaluates the performance of %10 size of labeled data classifiers, which are composed of 10 equal sizes which are %90 training and %10 tests. All results from 10 selected algorithms are compared and the algorithms that gave the best results on the prepared datasets were determined. In the algorithms used in this study, the best result was obtained by Random Forest classifier with above 99% total accuracy and 0.99 ROC. The results of the study are shown in the following tables.

In this study, 10 algorithms were chosen, and 3 of them were selected; IB1, Random Forest, Bayesian net. These algorithms were chosen considering True Positive, True Negative, ROC and Precision rates.

- True Positive (TP): A true positive is an outcome where the model correctly predicts the positive class.
- True Negative (TN): A true negative is an outcome where the model correctly predicts the negative class.
- ROC: An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.
- Precision (P): Precision is proportion of actually correct positive identifications, as in (1).

$$P = \frac{TP}{TP + TN} \quad (1)$$

Table 1: Results of Running

	TP	TN	ROC	P
Bagging	0.9916	0.9957	0.9996	0.9893
BayesNet	0.9977	0.9985	0.9998	0.9963
ClassificationViaRegression	0.9963	0.9937	0.9985	0.9843
IB1	0.9972	0.9993	0.9982	0.9981
J48	0.9893	0.9939	0.9920	0.9847
JRip	0.9846	0.9837	0.9853	0.9600
MultilayerPerceptron	0.9986	0.9996	1	0.9991
NaiveBayes	0.9986	0.9996	1	0.9991
RandomForest	1	0.9996	1	0.9991
RandomTree	0.9743	0.9898	0.9821	0.9744

Table 2: Results of Sitting

	TP	TN	ROC	P
Bagging	0.9967	0.99994	0.99999	0.9808
BayesNet	1	0.99988	1	0.9624
ClassificationViaRegression	1	0.99996	1	0.9871

IB1	1	0.99999	0.99999	0.9967
J48	1	0.99998	0.99999	0.9935
JRip	0.9870	0.99996	0.9955	0.9870
MultilayerPerceptron	1	1	1	1
NaiveBayes	1	1	1	1
RandomForest	1	1	1	1
RandomTree	0.9772	0.99998	0.9886	0.9934

Table 3: Results of Standing

	TP	**TN**	**ROC**	**P**
Bagging	0.9967	0.99994	0.99999	0.9808
BayesNet	1	0.99988	1	0.9624
ClassificationViaRegression	1	0.99996	1	0.9871
IB1	1	0.99999	0.99999	0.9967
J48	1	0.99998	0.99999	0.9935
JRip	0.9870	0.99996	0.9955	0.9870
MultilayerPerceptron	1	1	1	1
NaiveBayes	1	1	1	1
RandomForest	1	1	1	1
RandomTree	0.9772	0.99998	0.9886	0.9934

Table 4: Results of Walking

	TP	**TN**	**ROC**	**P**
Bagging	0.9786	0.9998	0.9993	0.9827
BayesNet	0.9839	0.9999	1.0000	0.9872
ClassificationViaRegression	0.9716	0.9999	0.9964	0.9906
IB1	0.9969	0.9999	0.9985	0.9981
J48	0.9708	0.9998	0.9883	0.9781
JRip	0.9733	0.9998	0.9910	0.9774
MultilayerPerceptron	0.6497	0.9992	0.8298	0.8862
NaiveBayes	0.9572	0.9993	0.9994	0.9325
RandomForest	0.9969	0.9999	1	0.9994
RandomTree	0.9416	0.9995	0.9705	0.9468

VI. Conclusion and Future Work

Human gait is a distinctive feature of a person that is determined by, among other things, an individual's weight, limb length, footwear, and posture combined with characteristic motion. Gait can be used as a biometric measure to recognize known persons and classify unknown subjects.

It should be noted that, because terrorists are relatively rare, identifying one in the crowd is still a huge problem. But, gait recognition technology shows some promise and this could help to spot people behaving suspiciously in sensitive areas, like airports, embassies, or military facilities.

This study has worked on HugaDB which is an open source database. This database compounds different human activities just as running, sitting, walking and standing. Data were collected from a body sensor network consisting of six wearable inertial sensors located on the right and left thighs, shins, and feet. In total, 2,111,962 samples were collected from all the 18 participants, and they provided a total of 10 hours of data. Our total accuracy with Random Forest classifier is above 99%.

The future work aims to create a better system by using our own data set to achieve higher accuracy and improve the speed of implementation.

References

[1] O. Sharma and S. K. Bansal, "International Journal of Innovative Technology and Exploring Engineering," International Journal of Innovative Technology and Exploring Engineering, vol. 3, no. 1, pp. 217–220, 2013.

[2] C. Zhang, W. Liu, H. Ma, and H. Fu, "Siamese neural network based gait recognition for human identification," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

[3] I. Gaba and S. P. Ahuja, "Gait analysis for identification by using BPNN with LDA and MDA techniques," 2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE), 2014.

[4] Y. Liang, C.-T. Li, Y. Guan, and Y. Hu, "Gait recognition based on the golden ratio," EURASIP Journal on Image and Video Processing, vol. 2016, no. 1, 2016.

[5] *Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 2, pp. 209–226, Jan. 2017.*

[6] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. P. D. L. Blanca, "Automatic Learning of Gait Signatures for People Identification," Advances in Computational Intelligence Lecture Notes in Computer Science, pp. 257–270, 2017.

[7] "Multiclass Classification," Amazon. [Online]. Available: https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-classification.html.

[8] P. Gupta, "Decision Trees in Machine Learning – Towards Data Science," Towards Data Science, 17-May-2017. [Online]. Available: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052.

[9] FutureLearn, "Classification by regression - Data Mining with WEKA," FutureLearn. [Online]. Available: https://www.futurelearn.com/courses/data-mining-with-WEKA/0/steps/25397.

[10] Bayesian Network Classifiers in WEKA for Version 3-5-7. (2008). Available: https://www.cs.waikato.ac.nz/~remco/WEKA.bn.pdf.

[11] R. Chereshnev and A. Kertész-Farkas, "HuGaDB: Human Gait Database for Activity Recognition from Wearable Inertial Sensor Networks," Lecture Notes in Computer Science Analysis of Images, Social Networks and Texts, pp. 131–141, 2017.