Generation of Synthetic Data for Behavioral Gait Biometrics

Aleksander Sawicki
Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland *a.sawicki@pb.edu.pl*

Khalid Saeed
Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland
Department of Computer Science and Electronics Universidad de la Costa, Barranquilla, Colombia *k.saeed@pb.edu.pl*

Wojciech Walendziuk
Faculty of Electrical Engineering, Bialystok University of Technology, Bialystok, Poland *w.walendziuk@pb.edu.pl*

Abstract

The research involved creating synthetic samples to enrich the training set and improve classification performance. Data generation was a key element of the biometrics gait system based on wearable sensors. The aim of the study was to investigate which parameters of the Long short-term memory–Mixture Density Networks (LSTM–MDN) models would provide the greatest increase in recognition metrics. Validation was conducted for normalized and non-normalized data for a large 100-person dataset. In the first case, the use of synthetic data from VAE-type generative models increased the F1-score from 0.754 to 0.776, while for proposed architectures increased metrics to 0.789. For normalized data, VAE-based models worsened recognition performance. Whereas the proposed model increased the F1-score from a baseline of 0.928 to 0.966. The conducted experiments indicate that generating synthetic data based on MDN models is more profitable in the cases of distribution shift between training and testing set.

Keywords: Gait biometrics, Generative Models, VAE, MDN, accelerometer

1. Introduction

Recently, solutions implementing the issue of generating synthetic samples have become more and more popular. Commercial text-to-image solutions, such as Midjournay, can be considered familiar even to people who are not familiar with the details of generative models. In this study, a generative model was implemented to multiply the training set with synthetic samples. We investigated the effect of the number of artificially generated samples on the performance of the CNN classifier. Generating synthetic samples in such a context is extremely sophisticated due to the need to generalize the input data while maintaining certain specific features which are characteristic to a subject. This article presents basic research on the use of generative models, with particular emphasis on LSTM-MDN models in the context of multiplying training data in a gait biometrics system.

As part of the work carried out, a system for identifying people based on gait was developed, operating on the basis of wearable sensors such as accelerometer and gyroscope. Particularly noteworthy is the fact that very small sets were used to train the models (about 30 training samples per participant), and the biometrics experiment itself

was validated in cross-day validation (where training and validation of the biometrics system is carried out on two separate days). In this configuration, there is a distribution shift between the training and test sets. Finally, it should be noted that enriching training sets with artificially generated samples has been successfully adapted to HAR (Human Activity Recognition) applications [1, 2].

Behavioral biometrics, despite the numerous problems associated with high sample variability, has several important advantages. First of all, it is very difficult to intentionally forge a sample and gain unauthorized access. In addition, this type of biometrics does not require active interaction with additional devices such as fingerprint factors. Typically, verification of the participant can be carried out without their active participation. In our opinion, these features mean that this type of behavioral biometrics should be further developed and it explains our interest in this field of science.

2. Literature Review

This work is a continuation of series of publications on the use of LSTM-MDN models to enrich the training set with synthetic samples and the impact of this approach on the effectiveness of the gait biometrics system in cross-day validation scenarios. At the beginning, it should be noted that models of this type have been successfully used to generate handwriting samples [3] or generate accelerometer signals for HAR purposes [4]. Models of this type may provide some alternative to generative models based on autoencoders in the form of timeVAE [5] or RHVAE approaches [6].

Publication [7] presents pilot results on data production using generative models for the author's data corpus. The study was conducted using MLP and probabilistic module (the number of modeled distributions was $M = 1$). The usefulness of the created samples was compared with the data created by autoencoders such VAE type model. The experiments were conducted with the use of non normalized data. In contrast, in [8], the LSTM-MDN model was implemented and used as a generative model. In this case, data generation was again performed using a single normal distribution with non normalized data. The experiments were carried out using three corpora for multi-day validation, including the first author's corpus and two sets created using mobile phones. These corpora were taken under different environmental conditions, the authors laboratory corpus (100 subjects), SIGNET semi-polygonized corpus (28 subjects) and Boston daily life scenario (29 subjects). Both papers [7,8] ignored the influence of the number of modeled normal distributions, the influence of the data normalization aspect and the number of synthetic samples created,

In the work [9], We developed a biometric gait system using motion sensors embedded in a mobile phone. Involved dataset had a unique feature in availability of three motion tracking session. In the performed study training set was built with the samples collected during two tracking sessions. The work verified the feasibility of using data generation for normal distributions M in range $\langle 1,4 \rangle$ and different numbers of generated samples. The study was conducted only for non-normalized data, and the biggest drawback was the use of a small data corpus of only 13 individuals.

The present work is a natural continuation of earlier studies [7, 8, 9], with no significant shortcomings mentioned in them. It openly compares the results for non normalised and normalised data. Results are presented for a varying number of synthetically generated samples as well as the number of modeled data distributions. And, most importantly, the experiments were conducted on a large, 100-member corpus of data, where the learning and test sets were collected on two separate days.

3. Methodology of the Research

3.1. Dataset

The study used an original corpus of movements collected by the authors as part of their previous work. The dataset included gait recordings of people who participated in two motion tracking sessions on two independent days. Such a data acquisition session enabled cross-day validation (training includes data collected on one day, with validation of a sample from day two). This approach is closest to the everyday life scenario. The experiment involved the participation of 100 individuals affiliated with the university

academic community, including students and academic staff members. Demographic data: age in years (M = 32.18, SD = 8.73); height in meters (M = 1.73, SD = 0.11). Exclusion criteria: healthy; no past or current injuries. Data collection took place in a laboratory conditions. Participants performed 20 walking trials on a hard ceramic surface over a distance of 3 meters. Two devices were involved in the data acquisition process: Perception Neuron 32 [11] motion capture system and Kinect v 2.0 depth camera [12]. First device was composed of a dedicated suit with 17 Inertial Measurements Unit (IMU). Each included a three-axis accelerometer and a gyroscope. It should be clarified that processing the data from the depth camera signal is not the subject of this work.

Despite the availability of numerous IMU sensor set, it was decided to use data from a single sensor located in the upper part of the right thigh. This location is similar to the location of the motion sensor built into a mobile phone and located in the right pants pocket. Our pilot studies showed that the use of all available sensors provides very good identification scores. However, such a system would be completely unimplementable in real-world applications.

3.2. Data preprocessing

Preliminary preprocessing includes aspects such as: *segmentation of gait cycles*, *interpolation to a fixed length*, *conversion between coordinate systems*, *frequency filtering*, and optional *data normalization*. The data segmentation process aims at isolating the so-called gait cycles, i.e. the period between the moment when the right leg hits the ground. Even though the experiment used a single sensor located in the right thigh area, the use of accelerometer signals allows for satisfactory detection of impacts. The segmentation algorithm is described in detail in [12], and allowed us to extract 3376 and 3321 gait samples for the first and second day, respectively. The next step is to interpolate the data to a fixed length of 128 frames. A constant sample duration is a requirement of the classifier. Sensors such as accelerometer and gyroscope measure quantities in the local reference frame of the sensor, which means that the indications depend on the mounting method. Due to the fact that the experiment participants moved along a constant straight path and the knowledge of the orientation of the sensor and the location of the experiment [13], it was possible to use signal transformation for the so-called global frame of reference. Performing this processing minimizes the impact of the sensors inclination/montage on their measurements values.

The measurement data was then subjected to low-pass frequency filtering with a 3rd order Butterworth filter, with a cut-off frequency of 6 Hz. This approach is common in the literature [14]. The final but optional preprocessing step was min-max normalization in range ⟨-1,+1⟩. This form of data processing is intended to enhance the process of classifier training. Figure 1 shows an example of the data collected for a single study subject. Classification process involved data from a triaxial accelerometer and gyroscope, therefore the figure has six sub-chart. The green color shows the data collected during the first day, and the red color shows the data collected during the second day.

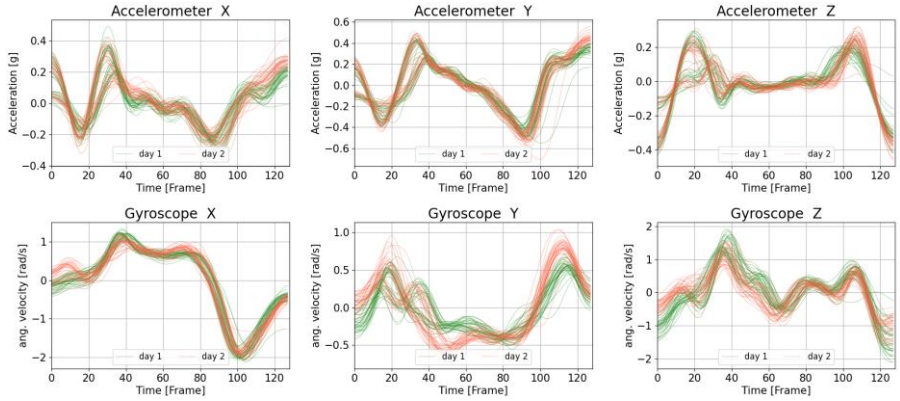

Fig. 1. Gait samples of the selected participant constituting after preprocessing.

3.3. Synthetic data generation

In the training process 100 instances of generative models (number of subjects in the dataset) were created for each type of generative model. We trained each of these models using the measurement data of a selected experiment participant. In the inference process, each instance was able to generate the desired number of samples in the form of a 6 x 128 array. In the proposed solution, the model learned the gait characteristics of the selected participant. The work verified the impact of the number of generated synthetic samples on the effectiveness of the biometrics system. It was decided to verify the number of {15, 30, 60, 120, 240} samples, which constituted {30.8, 47.1, 64.0, 78.0, 87.7}% of the content of the training sets, respectively. This was performed independently of the used generative model type. For the timeVAE and RHVAE models, off-the-shelf implementations were used. For the LSTM-MDN models, a hand-prepared implementation in Python was used. The open source solution in this case was largely concerned with the generation of 1D or 2D data, which, in the case of 6-dimensional data, limited its applicability.

The present work was a continuation of our research into the possibility of using LSTM-MDN networks to generate synthetic samples. In particular, a large aspect was concerned to the exploration of the number of distribution models and the number of synthetic data produced on the performacnce of the biometric system. The research was carried out separately for non-normalised and normalised input data .

It should be noted that the LSTM component was used as a regressor dedicated to time series processing, whereas the use of a probabilistic component provided a non-deterministic inference and data generation process. At each step t of the given time series, the Mixture Density Network component has the ability to model M distributions.

The input data of the target decision module of the biometrics system consisted of a fixed-length time series of 128 samples. Therefore, in this approach, the input of the LSTM-MDN network was provided with the frame number in a given gait sample. The output of the architecture were parameters defining the independent normal distribution of the three-axis accelerometer and gyroscope. The output vector y_t includes the weights π_j, the vector of mean values μ_j, and the standard deviation vector σ_j of each of the M modelled normal distributions (1) [3].

$$y_t = \{ \pi_t^j, \mu_t^j, \sigma_t^j \}_{j=1}^{M} \tag{1}$$

The LSTM-MDN model will model $13 \times M$ output values for each moment of time t. Due to the simultaneous modelling of 6 channels, $6 \times M$ standard deviations, $6 \times M$ mean values, and M component weights will be available for each of the M-modelled distributions. In the process of optimizing network weights, the ADAM algorithm is used to minimize the cost function:

$$L(x) = \sum_{t=1}^{T} - \log\left(\sum_j \pi_t^j / (2\pi\sigma_1\sigma_2\sigma_3\sigma_4\sigma_5\sigma_6) \exp(-Z/2) \right), \tag{2}$$

where the Z parameter is described by equation (3):

$$Z = \sum_{ax=1}^{6} (x_{ax} - \mu_{ax})^2 / \sigma_{ax}^2 \tag{3}$$

Figure 2 shows an example of modelled gyroscope Y-axis readings (also visible in Figure 1) by two instances of the LSTM-MDN network differing in the number of modelled normal distributions M. The graphic consists of two main sections, for each of them a modelled mean value with standard deviation. Individual distributions differ in drawing color. Below there is a smaller graph of the modelled component weights.

In the graphic shown, it can be seen that the network has correctly learned to model normal distributions. However, by observing the probability plots, some redundancy in the model can also be seen. There are parts of the data whose occurrence is impossible to

observer. For example, for the parameter $M = 4$, in the final part of the gait cycle from about step 110 to step 128, the second (orange) and third (green) normal distributions are virtually impossible to select. The probability of drawing them is zero.

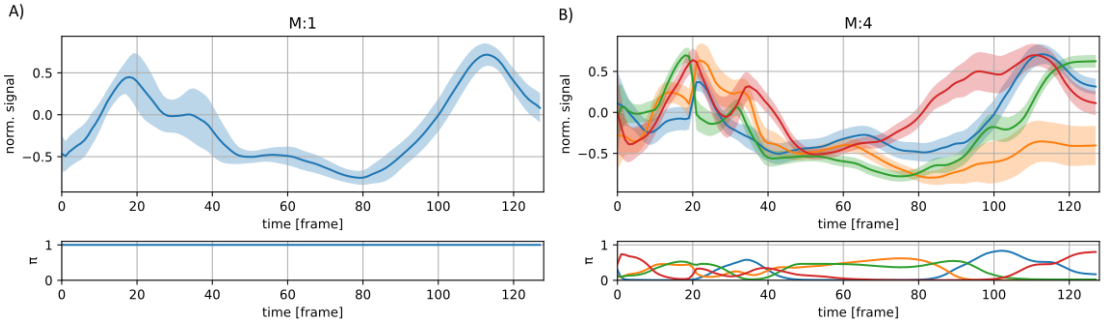

Fig. 2. Modeling of measurement data by the LSTM-MDN network for a) one B) four distributions.

The LSTM-MDN model is not able to directly generate new synthetic samples by inference. This is because it models the parameters of distributions. The following pseudo-code extract describes the data generation process based on Gaussian sampling using the generated parameters of the normal distributions First of all, the generation process requires specifying for which participant the data are to be generated (P), how many samples will be produced (AUG_NUM), how many normal distributions (M) are to be modelled, the type of standard deviation gain (STD_GAIN), and finally normalisation (NORM) parameter. It is worth noting that at each step *t*, a random selection of the currently selected normal distribution (k) is preformed. The actual sampling then takes place with its involvement. The optionally generated data is subjected to an inverse transformation (the case without data normalisation) and mandatory frequency filtering.

```
Gait sample generation procedure:
Select: P an ID of selected participant
Select: AUG_NUM a number of generated samples
Select: M a number of modeled normal distribution
Select: STD_GAIN a standard deviation gain
Select: NORM a flag that indicate if normalise data

Begin
participant_data_block=[]
for sample:=0 to AUG_NUM do:
      sample_data_block=[]
      for t:=0 to 128 do:
       mu,std,weight=request_from_pretrained_model(P,M)
       std=std*STD_GAIN
       for j in M do:
            k =draw_distribution(weight)
            data_block=gaussian_sampling(mu[k],std[k])
            if not NORM then:
               data_block==inverse_transform(data_block,min_max_scaler,P)
               sample_data_block.append(data_block)

      for selected_ax to 6 step 1 do:
         selected_ax_block=sample_data_block[selected_ax]
         filtered_ax_block=filter(selected_ax_block)
         sample_data_block.update(selected_ax,filtered_ax_block)
      participant_data_block.append(sample_data_block)
end
```

3.4. Data classification procedure

The decision-making module was based on an artificial convolutional neural network receiving, as its input, segmented gait samples, the so-called gait cycles. The biometrics system implemented the issue of identification, and it is used to predict one of 100 labels specifying the identification number of the experiment participant.

A CNN neural network with an attention mechanism was used as a classifier [6]. This type of model has achieved very good results in previous work [7,8]. Segmented measurement readings of a three-axis accelerometer and a three-axis gyroscope (6 channels in total) interpolated to a fixed length of 128 were used as input data. Typical cross-entropy was used as the cost function, and the ADAM algorithm was used to

optimize the network weights. A constant learning rate of 0.01 and a number of 200 learning epochs were used. The details of the network architecture are presented graphically in Figure 3.

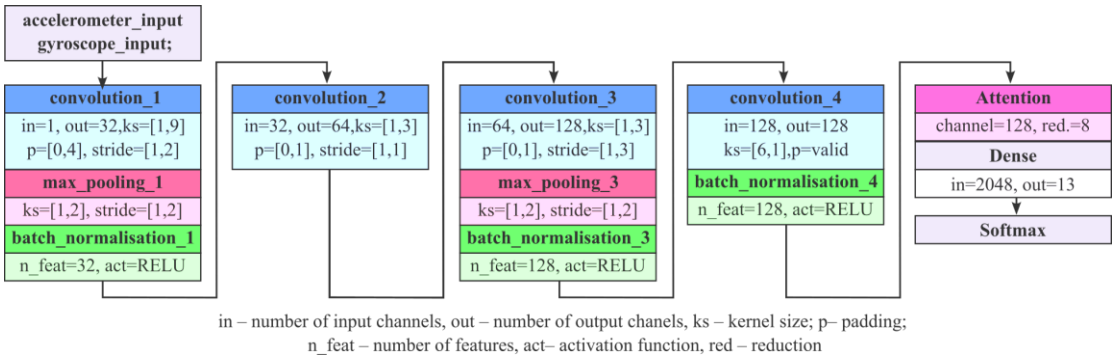

Fig. 3. CNN with attention mechanism network architecture.

4. Results

The results of the developed biometrics systems were validated using 20-time repeated simple validation. Each experiment was reproduced many times to minimize the influence of the randomness of the initial network weights. The research was carried out in two basic options - without normalisation and with normalisation of the input data.

The study was conducted in the baseline case (without synthetic samples) and with data generated by the LSTM-MDN as well as timeVAE and RHVAE models. For the first model, the number of modelled distributions is M ={1, 2, 3, 4}. As part of the conducted research, the effect of the amount of data generated on the effectiveness of the decision model was examined in the cases of {15, 30, 60, 120, 240} synthetic samples per participant. In addition, for each of the LSTM-MDN` models, the possibility of increasing the variance in the range {1, 4, 8, 16, 32} was verified.

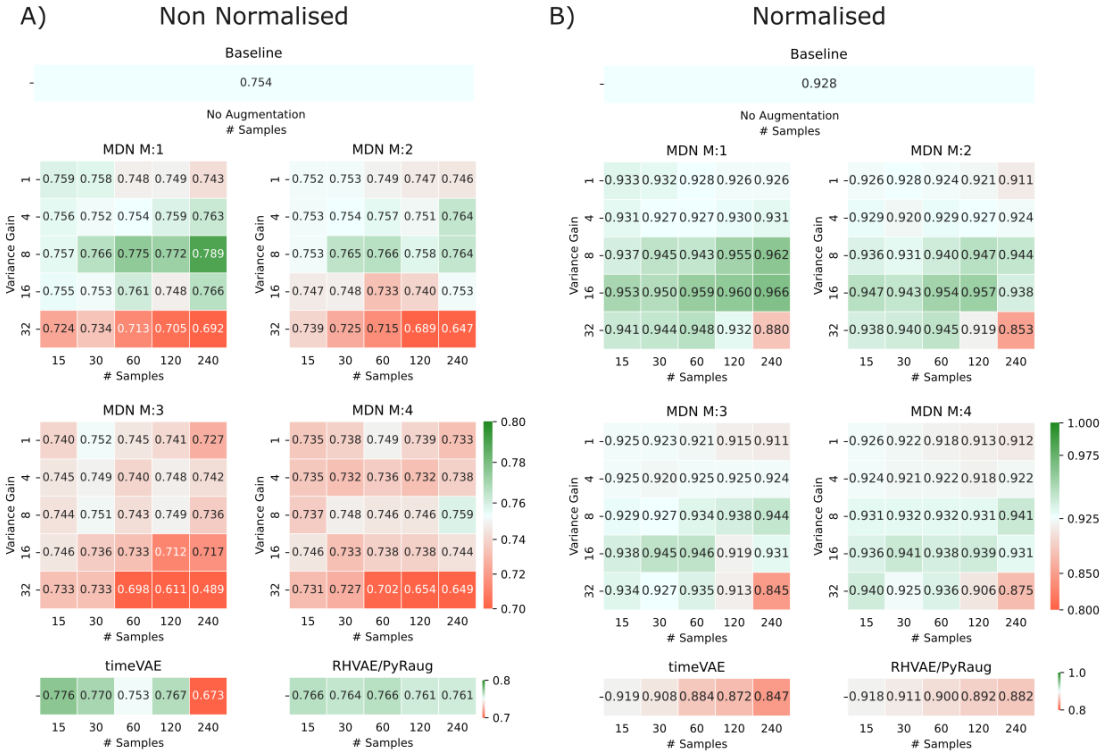

Fig. 4. F1-score measure of biometrics systems for: A) non normalised and B) normalized input data

Figure 4 presents the results of the F1-score measure in the form of a heatplot visualization, where the numerical value corresponds to the median result. The graph has two subgraphs for non-normalised A) and normalised data B). In each of them there is a total number of 7 subplots. The first row contains a visualization of the baseline case

without synthetic samples. The next two rows contains four graphs that show the results for data generation using MDN models differing in the number of modelled normal distributions. MDN models have the ability to gain variance after training, so these charts have an additional five rows. It should be noted that the variance amplification takes place before frequency filtering, which means that amplification, e.g. 32, will not actually generate 32 greater variance. Last row presents the results of using the timeVAE and RHVAE models.

Several relationships can be observed in Figure 4 A): In the baseline case (without synthetic samples), a metric of 0.754 F1-score was achieved; typically, the best results are observed with the number of 60 and 120 generated samples; the timeVAE model enabled to increase the F1-score to 0.776 and the RH-VAE model to 0.766; MDN models allow achieving good results mainly when a single normal distribution is modeled (maximum value 0.789), or in $M = 2$ case (maximum value 0.766); the greatest observed increase in metrics is visible for $M = 1$, when the variance gain is equal to 8;the use of MDN models in $M = \{3, 4\}$ normal distributions cases only worsens the results of the biometrics system.

Figure 4 B), shows that for the normalised data, the F1 score for the baseline is significantly higher (0.928) compared to the non-normalised data (0.754).The illustration provides evidence for the following conclusions: Synthetic sample generation using timeVAE and RHVAE models only causes degradation of biometrics system metrics; In the $M = 1$ normal distributions case, the maximum change in effectiveness is observed maximum F1-score is equal to 0.966, whereas in $M = 2$ normal distributions cases is lower - 0.957. The maximum increase to the level of 0.966 is observed for $M = 1$ normal distributions, with a sixteen gain in variance.

4. Conclusions and Future Works

As part of the work, a biometric gait system was built and validated using an original data corpus of 100 subjects. Measurement data from the accelerometer and gyroscope of a sensor located in the area of the right thigh were used as the source of input data. The evaluation of the developed solution was carried out in a cross-over configuration in which training and test samples were collected on different days. The main focus of the work concerned the use of synthetic samples generated by timeVAE, RHVAE, and LSTM-MDN models to enrich the training set and consequently improve the generalization properties of the decision module.

For non-normalized data, the use of models from the autoencoder group allowed to achieve very good results: 0.776 F1-score for timeVAE and 0.766 for RHVAE, respectively. While in the case of normalized data, these models allowed to generate samples that only reduced the metrics of the biometrics system (0.919 F1-score for timeVAE and 0.918 for RHVAE, respectively).In the case of LSTM-MDN models, the system efficiency increased from 0.754 to 0.789 F1-score for the case of unnormalized data (Figure 4 A) and from 0.928 to 0.966 F1-score for normalized data (Figure 4 B). Models of this type can be considered the most profitable.

The carried out experiments showed that the use of synthetic samples can have a positive impact on gait biometrics system metrics. At the same time, it was shown that in the case of unnormalized data, an increase in effectiveness was observed from 0.764 to 0.789 of the F1-score measure, with the base effectiveness of normalized data being 0.928. The experiments performed (Figure 4) indicate that synthetic data can improve the effectiveness of models, but it cannot replace pre-processing elements. The addition of synthetic samples should be the last element of work in the process of developing a biometrics system.

Despite the fact that synthetic samples based on LSTM-MDN models have achieved promising results in this study, some shortcomings of the developed approach can also be noticed. First of all, in the case of modeling more than one normal distribution, in each step t there is a random selection of the current distribution. In the next step, a sample is generated using Gaussian sampling. For 128 gait samples, this would force a maximum

of 127 possible 'jumps' between the modelled distributions. This approach results in a wide variety of data being generated, which in the case of behavioral biometrics is not necessarily an advantage. Our further ideas in this area will be to develop an aggregation approach with a limited number of draws of distributions. The last element that would be worth extending is the comparison of effectiveness in the case of using samples produced by the generative model and those obtained by perturbing existing samples, i.e. the issue of data augmentation.

Acknowledgment

This work was supported by grant 2021/41/N/ST6/02505 from Białystok University of Technology and funded with resources for research by National Science Centre, Poland. For the purpose of Open Access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

References

1. Wang, J., Chen, Y., Gu, Y., Xiao, Y., Pan, H.: SensoryGANs: An Effective Generative Adversarial Framework for Sensor-based Human Activity RecognitionInt. Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 1-8 (2018)
2. Li, X., Luo, J., Younes, R.: ActivityGAN: generative adversarial networks for data augmentation in sensor-based human activity recognition. In Adjunct Proc. of the ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing and Proc. of the ACM Int. Sym. on Wearable Computers Association for Computing Machinery, 249–254 (2020)
3. Graves, A.: Generating Sequences With Recurrent Neural Networks. arXiv:1308.0850 (2013)
4. Alzantot, M., Chakraborty, S., Srivastava, M.: SenseGen: A deep training architecture for synthetic sensor data generation, 2017 IEEE Int. Conf. on Pervasive Computing and Communications Workshops (2017)
5. Desai, A., Freeman, C., Wang, Z., et al.: Timevae: A variational auto-encoder for multivariate time series generation. arXiv:2111.08095 (2021)
6. Chadebec, C., Thibeau-Sutre, E., Burgos, N., et al.: Data Augmentation in High Dimensional Low Sample Size Setting Using a Geometry-Based Variational Autoencoder. arXiv:2105.00026 (2021)
7. Sawicki, A., Saeed, K.: Application of Generative Models to Augment IMU Signals in Gait Biometrics. Dependable Computer Systems and Networks. DepCoS-RELCOMEX 2023. Lecture Notes in Networks and Systems, vol. 737 (2023)
8. Sawicki, A., Saeed, K.: Gait-Based Biometrics System, Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track. ECML PKDD 2023. Lecture Notes in Computer Science, vol. 14175 (2023)
9. Sawicki, A., Grabowski, D.: Application of Mixture Density Network for sample generation in behavioral biometrics, applied to 23rd International Conference on Computer Information Systems and Industrial Management Applications (2024)
10. Perceptrion Neuron 32 official website, https://neuronmocap.com/products/perception_neuron/ (acc. on 16.01.2019)
11. Kinect for Windows official website. Available online: https://www.microsoft.com/en-us/kinectforwindows /(accessed on 28.05.2015)
12. Sawicki, A., Saeed, K.: Application of LSTM Networks for Human Gait-Based Identification. Theory and Engineering of Dependable Computer Systems and Networks. DepCoS-RELCOMEX. Advances in Intelligent Systems and Computing,1389 (2021)
13. Subramanian, R., Sarkar, S.: Evaluation of Algorithms for Orientation Invariant Inertial Gait Matching, in IEEE Transactions on Information Forensics and Security, 14 (2), 304-318 (2018)
14. Luo, Y., Coppola, S.M., Dixon, P.C.et al.: A database of human gait performance on irregular and uneven surfaces collected by wearable sensors. Scientific Data, 7 (2020)
15. Huang H., Zhou P., Li Y., et al.: A Lightweight Attention-Based CNN Model for Efficient Gait Recognition with Wearable IMU Sensors, Sensors 21, 2866 (2021)