2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Marker-free Human Gait Analysis using a Smart Edge Sensor System

Eva Katharina Bauer[1], Simon Bultmann[2], and Sven Behnke[2]

***Abstract*— The human gait is a complex interplay between the neuronal and the muscular systems, reflecting an individual's neurological and physiological condition. This makes gait analysis a valuable tool for biomechanics and medical experts. Traditional observational gait analysis is cost-effective but lacks reliability and accuracy, while instrumented gait analysis, particularly using marker-based optical systems, provides accurate data but is expensive and time-consuming.**
In this paper, we introduce a novel markerless approach for gait analysis using a multi-camera setup with smart edge sensors to estimate 3D body poses without fiducial markers. We propose a Siamese embedding network with triplet loss calculation to identify individuals by their gait pattern. This network effectively maps gait sequences to an embedding space that enables clustering sequences from the same individual or activity closely together while separating those of different ones. Our results demonstrate the potential of the proposed system for efficient automated gait analysis in diverse real-world environments, facilitating a wide range of applications.

I. Introduction

Locomotion, particularly walking, is an essential ability for humans. It is learned at an early age and is usually a subconscious process. Moving on two legs gives us independence and makes physical activities like running possible.

Because of its complexity and uniqueness, gait analysis attracts significant interest from both experts in biomechanics and medical professionals. Observing and analyzing changes in gait can not only provide insight into neurological and physiological conditions, but can also aid in the development and evaluation of individualized treatments [1], [2].

The clinical use of gait analysis mainly focuses on observational gait analysis performed with human eye and brain. This method is simple and cost-efficient but suffers from low validity, reliability, and responsiveness. Instrumented gait analysis on the other hand promises accurate and reliable gait data for medical use [2]. Marker-based optical motion capture systems are considered the gold standard in instrumented gait analysis. They provide a high level of precision, but are expensive and time-consuming to use [3].

In this work, we instead employ a markerless approach to capture human movement, developed in previous work [4]–[6]. A multi-camera system is used to estimate the 3D poses of multiple persons in real time, as illustrated in Fig. 1 (a, b). We refer to the camera nodes as *smart edge sensors*, as they contain an integrated inference accelerator for local, on-board, semantic image interpretation. Unlike marker-based systems that require applying markers to the participant's body (e.g. wearing a marker suit), the smart edge sensors detect human joints by inferring heatmaps of human body keypoints from the camera images. These keypoint detections are streamed to a central backend, where the pose estimates of each camera are fused into a 3D skeleton per observed person. The fusion process incorporates priors on typical bone lengths to enhance the accuracy of the pose estimation [4], [5]. The proposed system enables the capture of human motion sequences and gait measurements in real time, making gait analysis accessible for wider use in diverse, real-world environments.

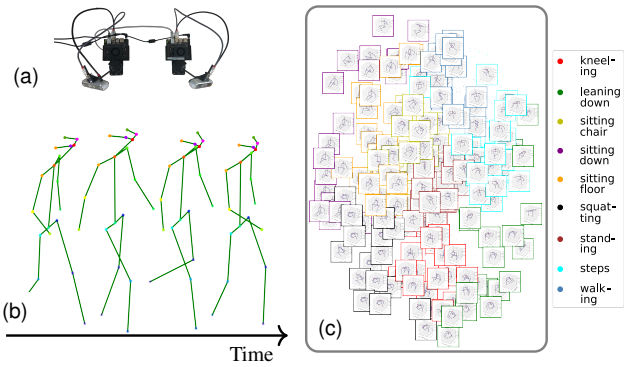


Fig. 1. Sample movement sequence (b) captured with smart edge sensors (a) consisting of a Nvidia Jetson Orin compute board and an Intel RealSense RGB-D camera. (c) t-SNE visualization of activities from Human 3.6M dataset based on the learned gait embedding.

Another significant challenge in clinical gait analysis is the wide range of parameters affecting human gait. Efficiently capturing gait patterns and accurately recognizing individuals based on them could simplify the differentiation between patient groups. Identifying similar gait characteristics across different patients may indicate a shared underlying condition, aiding in the selection of appropriate treatments. Integrating machine learning methods into this analysis facilitates recognizing individual gait patterns and clustering similar traits or activities across different individuals [7]–[10].

In this work, we design a Siamese embedding network based on TriNet [11] for the identification of individuals by their walking patterns and for the differentiation of activities. The network takes gait sequences as input and maps them into an embedding space, using a ResNet 18 backbone [12]. We train the network using the Triplet Loss [13] on L_2-distances in the embedding space to ensure that motion sequences from the same person or activity are positioned close together, while sequences from different individuals or actions are pushed apart (cf. Fig. 1 (c)).

This research has been funded by the Federal Ministry of Education and Research of Germany under grant no. 01IS22094A WEST-AI.

[1]Hochschule Koblenz, RheinAhrCampus, Remagen, Germany; {ebauer1}@hs-koblenz.de,
[2]Autonomous Intelligent Systems group, University of Bonn, Germany; {bultmann,behnke}@cs.uni-bonn.de

In summary, our main contributions in this paper are:

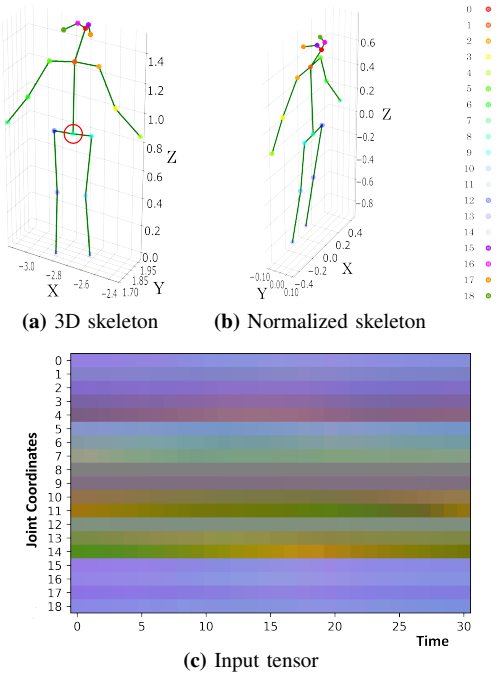
- We propose a novel deep learning-based framework to cluster human walking patterns in an embedding space to identify similar gait patterns or activities.
- We collect accurate gait data from multiple subjects in a real-world environment using a smart edge sensor network to capture human movement without the need to apply markers or sensors to the body.
- We quantitatively and qualitatively evaluate the proposed integrated system for human motion capture and gait analysis using the collected real-world data and the Human 3.6M database [14].

II. Related Work

The development of approaches for human gait pattern analysis with artificial neural networks started in 1993 with Holzreiter et al. [15] proposing a three-layer neural network with connected units to distinguish between healthy and pathological gait patterns using data captured by force measurement platforms. In 2002, Schöllhorn et al. [16] proposed to determine individual movement characteristics using self-organizing maps and data from force platforms. They compared the performance using time-continuous and time-discrete data concluding that continuous data leads to more accurate and stable results. In 2006 Han et al. [17] introduced the *gait energy image* (GEI) to fuse human motion in a single image using component and discriminant analysis to learn gait features.

Supervised machine learning approaches were introduced in 2007 by Lu et al. [18]. Their approaches are based on Independent Component Analysis (ICA) and Principal Component Analysis (PCA). As input, images of moving persons are reduced to binary silhouettes by background subtraction and then characterized by mathematical methods. Although their recognition accuracy was promising, the method cannot be reliably used in real-world environments. Many recent deep-learning frameworks for in-the-wild multi-view multi-person 3D pose estimation exist, e.g. using cross-view epipolar constraints [19], volumetric representations [20], [21], or direct regression [22], most of them centralized approaches that incur heavy computational load. We instead use a distributed, real-time capable approach [4], where images are processed locally on each smart edge sensor.

Alotaibi et al. [23] proposed a specialized deep Convolutional Neural Network (CNN) architecture for gait recognition. They used the CASIA-B dataset [24] and an eight-layer CNN to recognize individuals by their gait. In our work, we employ the widely used and efficient ResNet [12] backbone instead. In 2023, Taha et al. [25] introduced an auto-encoder to recognize biological and physiological characteristics of individuals, like gender, age, and weight. They used Inertial Measurement Units (IMUs) in both outsoles of the shoes and a marker-based motion capture system to collect gait data of people walking on a treadmill. After the learning process, they clustered the embeddings using K-Means, achieving a high identification accuracy. In contrast, we employ a marker-free optical recording setup, making the capture process significantly more time-efficient and accessible.



(a) 3D skeleton **(b)** Normalized skeleton

(c) Input tensor

Fig. 2. Example of the 3D skeleton model with $J = 19$ joints (a), coded by color. The red circle marks the root joint used for normalization (b); (c) RGB-representation of one $(J, T, 3)$ input tensor with $T = 30$ frames.

Schroff et al. [26] also attempt to recognize and cluster people by specific traits. They recognize the faces of people by reducing the input data to an embedding vector and clustering embeddings of pictures of the same person. They propose the semi-hard mining process for the triplet loss calculation and achieve accurate clustering, benefiting from their large dataset. We take up the semi-hard mining strategy in our training process but process the more privacy-preserving gait sequence data, as all image processing happens locally on the smart edge sensors in our system.

Fink et al. [27] attempt to distinguish between people with rare bone diseases and healthy controls by analysing their gait. In contrast to our approach, they use smartphone sensors to test a simpler but less cost-effective way for gait analysis. They employ a two-sided Mann-Whitney-U test to compare individual gaits, demonstrating the potential to identify pathological gait patterns, albeit with high intergroup variability in the measured parameters.

III. Method

A. Capture Space Setup

To capture gait data from participants, 25 smart edge sensors with RGB-D cameras are deployed throughout a $\sim$240 m^2 lab space, mounted at $\sim$2.5 m height [4], [5]. Each smart edge sensor estimates the 2D poses of detected persons from the RGB image stream by inferring a heatmap to locate $J = 19$ keypoints defined at the major joints of the human skeleton including nose, ears, and eyes. The CNN inference for the person detection and pose estimation runs locally on each sensor board using lightweight model architectures

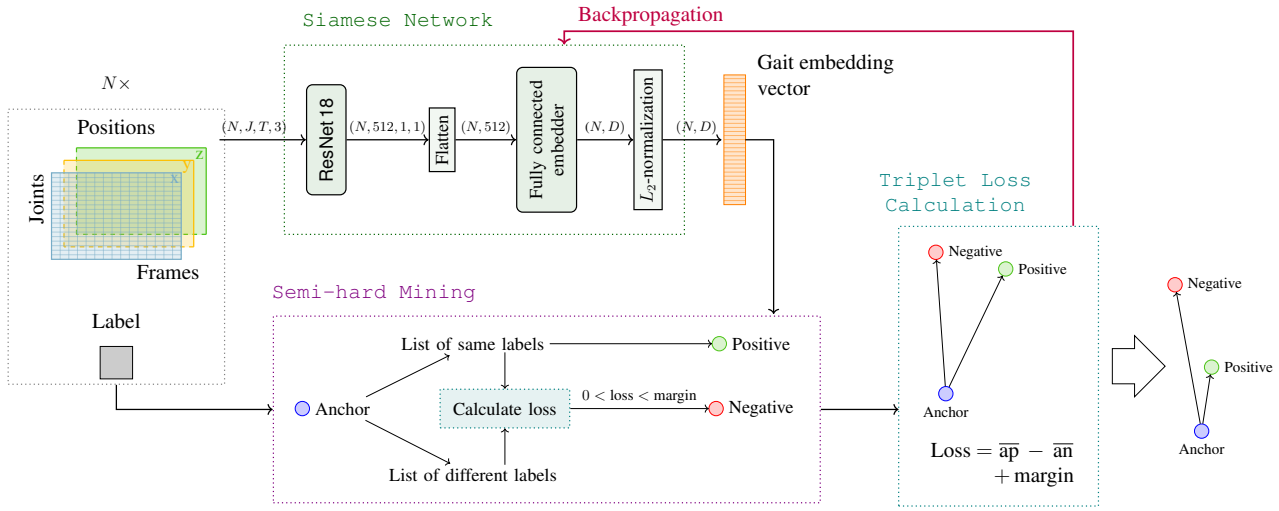


Fig. 3. Gait embedding network with *semi-hard* mining used for the training process. The input is a data batch of N tensors that hold the 3D position of J joints over sequences of T frames. The input is processed by the Siamese network which computes D-dimensional embedding vectors. Every embedding vector is matched with positive samples of the same person and negative samples of a different one to form triplets. The negative sample is selected by searching for embeddings close to the anchor vector to make the learning process more efficient. With the selected triplets, the loss is calculated and backpropagated through the network. After training, the network is able to map the same person close by and different ones apart.

TABLE I
PARTICIPANTS IN THE DATA COLLECTION.

Age	Height [cm]	Gender	
30 ± 10	176 ± 8	19 Male	3 Female

and embedded inference accelerators [4], [5]. Subsequently, only the estimated 2D pose coordinates are transmitted to a backend system, where multiple camera views are fused to estimate a 3D skeleton model for each observed person, as illustrated in Fig. 2 (a). Leveraging prior knowledge of human skeletal dimensions (e.g. bone lengths), the estimated pose coordinates are refined and then re-projected back to the sensor boards in a semantic feedback loop to further improve the accuracy of the pose estimation [4], [5].

B. Data Collection

For data collection, 22 persons without pathological or neuronal gait disorders volunteered to participate in the experiment. They were instructed to walk naturally, as they would in their everyday environment, for a duration of one minute. During this time, their joint positions were tracked and recorded by the above-described smart edge sensor system. Further details on the participants are given in Tab. I.

The resulting gait data from 20 participants was used to train a neural network for gait sequence embeddings. The data of each subject was divided into a larger training set (90%) and a smaller validation set (10%). The gait data from the remaining two participants was completely excluded from the training and validation process and was used exclusively for testing the performance of the final model.

C. Data Preprocessing

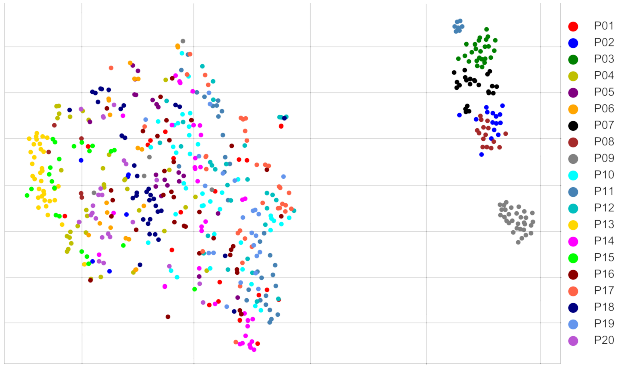
After data collection, the resulting gait tracks were filtered to exclude parts where the camera system failed to detect all participant's joints, as noisy, incomplete observations could lead to unrealistically long or distorted limb representations.

The coordinates of the remaining gait tracks were normalized to be independent of the subject's height, absolute position in the capture space, and orientation. For this, the height of each person is scaled to 1 and the root coordinate, located at the pelvis, is subtracted from each skeleton. Next, to normalize the orientation, the joint coordinates are rotated to align the x-axis with the connection of the left and right hip and the z-axis with the direction of the neck and pelvis (cf. Fig. 2 (b)). This ensures that all participants are constantly oriented as if walking straight forward, and the computed embeddings are independent of the absolute position and walking direction.

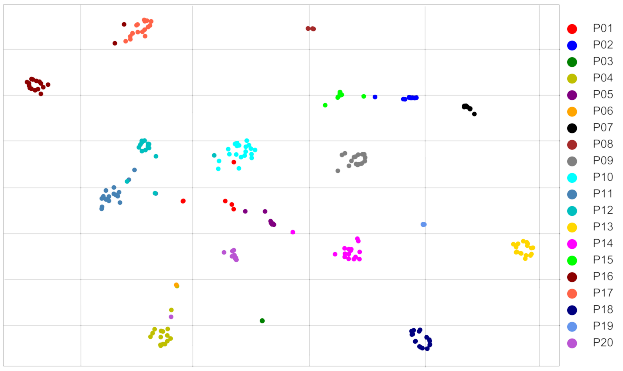
The resulting 3D joint coordinates for each frame are then represented as a matrix $X \in \mathbb{R}^{J\times 3}$, with $J = 19$ the number of tracked keypoints. For a gait sequence, the joint coordinates of a subject are then stacked into tensors with the shape of $(J, T, 3)$, illustrated in Fig. 2 (c). Each tensor includes the gait coordinates of a one-second walking sequence, resulting in a sequence length of $T = 30$ at a 30 Hz capture frame rate. Finally, the tensors are matched with a label indicating the participant's identification number.

D. Gait Embedding Network

The network architecture used to identify individuals by their walking pattern is based on a TriNet Siamese network [11] using a lightweight ResNet 18 [12] backbone for efficient computation. We further employ a semi-hard mining strategy for efficient triplet selection during training. The network architecture and training process are illustrated in Fig. 3. As input, the network receives data batches of the preprocessed and normalized gait sequences, each consisting of N tensors with dimensions $(J, T, 3)$. The ResNet 18 backbone maps each input tensor to a $(512, 1, 1)$ feature vector. This feature vector is flattened and projected to a D-dimensional embedding space using a fully connected linear layer. Finally, the embedding vector is L_2-normalized.



(a) raw input data



(b) final model

Fig. 4. t-SNE projections of the validation data. Each dot represents a validation sequence with Participant ID coded by color. (a) raw input data, (b) embeddings computed with final model after $10,000$ training epochs.

The network outputs a data batch of N D-dimensional gait sequence embedding vectors. As detailed in Sec. IV, we choose a batch size of $N = 64$ and an embedding dimension of $D = 32$ in our experiments. To enhance the network's performance, we initialized the ResNet 18 model with ImageNet-pretrained weights.

E. Triplet Loss with Semi-hard Mining

To enhance the speed and efficiency of the training process, we employ semi-hard negative mining for online triplet selection [26]. In this process, first, a batch of input gait sequences is sampled from the training data and processed with the network to compute the embedding vectors. Each output embedding then serves as a possible anchor to form triplets. For this, the anchor is associated with a positive example, which is an embedding vector from the same person as the anchor, and a negative example. The negative match is selected by calculating the loss values of all possible negative pairs, as shown in Eq. (1):

$$\text{Loss} = \overline{\text{ap}} \; - \; \overline{\text{an}} \; + \; \text{margin}. \tag{1}$$

The Euclidean distance between the anchor and a negative example ($\overline{\text{an}}$) is subtracted from the distance between the anchor and the positive example ($\overline{\text{ap}}$). Finally, a pre-selected margin $\delta = 0.2$, is added to this calculation. After calculating the losses of every possible negative match in the data batch given a pair of anchor and positive example, triplets with a loss greater than zero, but less than the margin are selected.

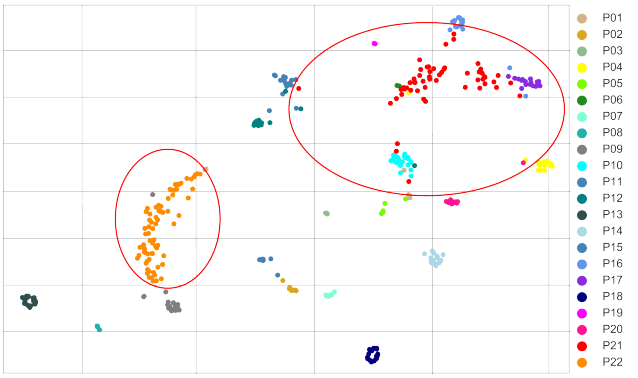


Fig. 5. t-SNE plot for validation of the model with gait data of two unknown participants 21 (red) and 22 (orange), highlighted with red circles.

The negative sampling is repeated for all $\overline{\text{ap}}$ pairs. This *semi-hard* mining strategy [26] selects negative samples that are further away from the anchor than the positive sample but still hard, as the embedding distance is close to the $\overline{\text{ap}}$ distance. In contrast, selecting only the hardest negative can lead to model collapses, as confirmed in Tab. II.

The model's loss is computed by averaging the losses of all found triplets, calculated as in Eq. (1). This loss is then back-propagated through the network to adjust and improve its weights. The network's ultimate goal is to produce embedding vectors that have small distances when representing the same person by learning to differentiate their gait patterns.

IV. Evaluation

To evaluate the clustering ability of the proposed network, we apply the K-means algorithm to the embedding vectors computed from the validation data. We calculate the Adjusted Rand Index (ARI) score [28] based on the resulting clusters for quantitative evaluation of the clustering accuracy. Further, we conduct ablation studies on the design choices of our network. To illustrate the network's clustering ability, we employ the t-SNE projection [29] to display the similarities of the embedding vectors in a 2D scatter plot. After completing the training process, the final network was additionally tested using the gait data of two previously unseen participants to assess the generalization capabilities of our system.

To evaluate our approach on activity clustering when dealing with larger data variety, we trained and tested the network using the popular Human 3.6M dataset [14].

A. Clustering Performance

The t-SNE plot in Fig. 4 compares the direct clustering of the raw input data to the network's clustering performance. During training, the network improves its ability to recognize different gait patterns, learning to separate the participants into distinct clusters. In the raw data, Fig. 4 (a), two clusters of several participants are visible, but a clear separation of individual subjects is not possible. Using the gait embedding vectors of the trained network, Fig. 4 (b), nearly every participant has a separate cluster.

To evaluate the generalization performance of our network, we test it with gait data of two unknown subjects, completely

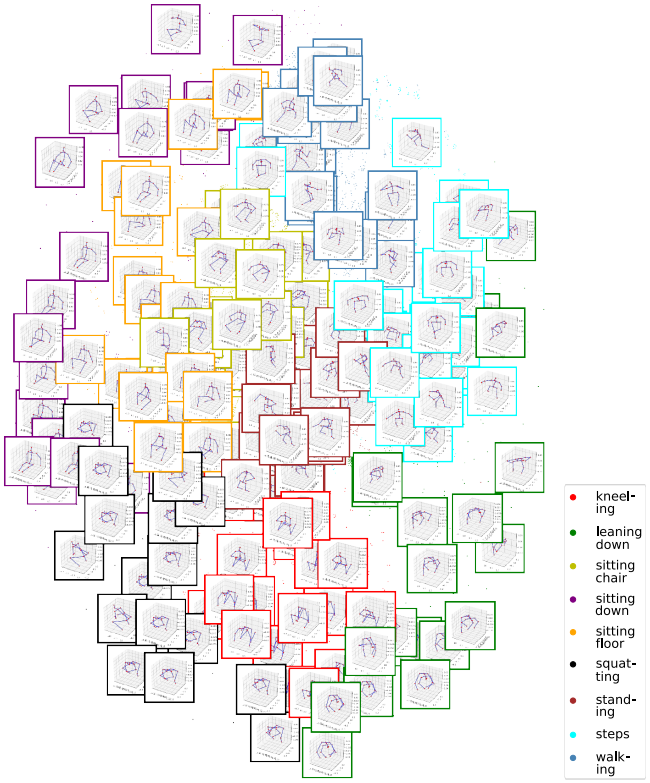


Fig. 6. t-SNE plot of activities from Human 3.6M (zoom into Fig. 1 (c)).

unseen during the training process, as shown in Fig. 5. The model successfully maps the unknown gait sequences close to each other, although they are slightly more spread out compared to the validation data of known participants. Notably, Participant 21 has several gait sequences incorrectly assigned to clusters of nearby participants.

Further experiments are performed with the public Human 3.6M database [14], where sequences are clustered according to the activity labels of Tanke et al. [30]. The clustering is visualized in Fig. 6, showing the first frames of the validation sequences. Different actions are nicely separated and similar activities lie closely together in the embedding space, i.e. sitting to the top-left, walking or standing to the top-right, and kneeling or crouching to the bottom. The evolution of the ARI score during training is shown in Fig. 7, with a final score of 73.5% after 750 epochs.

B. Ablation Studies

The proposed neural network was initially trained with several different parameters to find the most efficient choices. The tested parameters include the triplet selection method, the size of the embedding dimension, the size of the data loader batches, and the number of frames in each movement sequence. Parameters like the learning rate, $\text{lr} = 10^{-4}$, and the margin, $\delta = 0.2$, were selected as recommended in [26]. Tables II–IV list the resulting ARI scores.

Table II shows the model's clustering accuracy for different triplet selection procedures. *Semi-hard* mining achieves the highest accuracy, with 80.3% after 1,000 training epochs, while hard mining results in a model collapse with only 3.4%

TABLE II

CLUSTERING PERFORMANCE FOR TRIPLET SELECTION METHODS.

Method	random	semi-hard mining	hard mining	raw data
ARI score	77.3%	**80.3%**	3.4%	17.3%

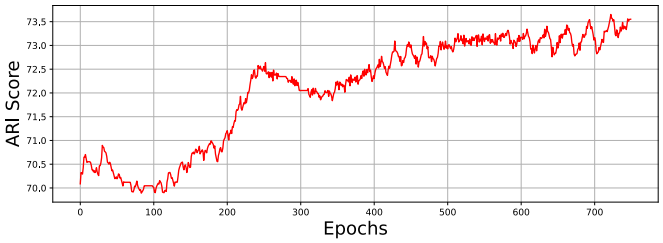


Fig. 7. ARI score evolution during training for H3.6M activity clustering.

accuracy. Using a random selection procedure, the accuracy is lower than with semi-hard mining, at 77.3%, still reaching a reasonable level of performance, significantly better than clustering the raw data directly (17.3%).

Comparing the different embedding dimensions D in Table III, the resulting ARI scores are fairly similar. With $D = 32$, the model achieves an accuracy of 80.3%, 0.3% higher than with $D = 64$. Increasing the embedding dimension further ($D = 128$) results in a decrease in ARI score to 77.6%. Additionally, training the model with larger embedding dimensions takes longer and requires more computational resources. Modifying the batch sizes shows little impact on model performance, as apparent in Table III. The best result of 80.3% is achieved with a batch size of 64. A batch size of 32 leads to an accuracy of 79.1%, while a batch size of 128 results in a lower accuracy of 77.2%.

Comparing the ARI scores after 1,000 and 10,000 epochs in Table IV reveals an improvement of approx. 10% for sequence lengths T of 30 and 45 frames, and 6.3% for $T = 15$. The highest accuracy, 87.8%, was achieved when the model completed 10,000 training epochs with a sequence length of 30 frames. It is also noticeable that for a training time of 1,000 epochs, the accuracy decreases as the sequence length increases. However, when training for 10,000 epochs, the ARI score improves when the sequence length is doubled to $T = 30$ frames but decreases when the length is increased again. Notably, all the scores after 10,000 epochs are higher than those after 1,000. It seems plausible that the model requires more training time to recognize patterns when processing longer input sequences that give more context, leading to increased accuracy. Visualizing the participants during these sequences suggests that 15 frames roughly correspond to one step, depending on the participant's walking speed. On the other hand, training the model with a sequence length of $T = 45$ results in decreased accuracy. This may be due to the reduction in the total number of input tensors as the sequence length increases while the available data is limited. To determine whether a higher sequence length could further improve the model, additional gait measurements over longer periods could be beneficial.

For the final model, the input data loader uses a batch size of $N = 64$, where each batch consists of gait sequences

TABLE III

CLUSTERING PERF. FOR EMBEDDING DIMENSIONS AND BATCH SIZES.

Emb. dim.	32	64	128	32		
Batch size	64			32	64	128
ARI score	**80.3%**	80.0%	77.6%	79.1%	**80.3%**	77.2%

TABLE IV

CL. PERF. FOR INPUT SEQUENCE LENGTHS AND TRAINING ITERATIONS.

Seq. length	15	30	45
ARI score (1,000 epochs)	**80.3%**	78.9%	74.1%
ARI score (10,000 epochs)	86.9%	**87.8%**	84.6%

of length $T = 30$ frames. The network outputs a $D = 32$-dimensional embedding vector. During the training process, triplets were selected using the *semi-hard* mining method, and the model was trained for 10,000 epochs. The average inference time for a batch of gait sequences was measured at 8.14 ms on an NVIDIA RTX 2080 Ti GPU, where our model uses 500 MB of GPU memory.

V. CONCLUSIONS

In this work, we proposed a novel approach for human gait clustering using a TriNet-based Siamese network to distinguish between differnet motion patterns. Gait data was captured with a smart edge sensor network providing accurate marker-less motion capture. This significantly speeds up gait analysis and improves accessibility to digital gait data compared to traditional marker-based systems, offering an innovative solution to gait analysis in sports and medicine.

While the clustering performance on raw gait data is poor, our model achieved an ARI score of 87.8% with gait embedding vectors computed from 30-frame sequences, effectively differentiating all participants. Testing on unknown participants showed promising generalization, though with slightly less tight clustering than known subjects. We further demonstrated the network's application in activity clustering using the Human3.6M dataset [14].

Improving the dataset by including longer sequences and more participants could enhance performance, potentially requiring additional training iterations. A smaller, less cluttered measurement space could reduce joint displacement errors and unusable sequences. Including participants with diverse gait characteristics, such as elderly or injured, could further test and enhance the system's robustness and applicability.

REFERENCES

[1] M. W. Whittle, "Clinical gait analysis: A review," *Human Movement Science*, vol. 15, no. 3, pp. 369–387, 1996.

[2] A. A. Hulleck, D. Menoth Mohan, N. Abdallah, M. El Rich, and K. Khalaf, "Present and future of gait assessment in clinical practice: Towards the application of novel trends and technologies," *Frontiers in Medical Technology*, vol. 4, p. 901331, 2022.

[3] M. Moro, G. Marchesi, F. Hesse, F. Odone, and M. Casadio, "Markerless vs. marker-based gait analysis: A proof of concept study," *Sensors*, vol. 22, no. 5, p. 2011, 2022.

[4] S. Bultmann and S. Behnke, "Real-time multi-view 3D human pose estimation using semantic feedback to smart edge sensors," in *Robotics: Science and Systems (RSS)*, 2021.

[5] ——, "3D semantic scene perception using distributed smart edge sensors," in *Int. Conf. on Intelligent Autonomous Systems (IAS)*, 2022.

[6] B. Pätzold, S. Bultmann, and S. Behnke, "Online marker-free extrinsic camera calibration using person keypoint detections," in *44th DAGM German Conf. on Pattern Recognition (GCPR)*, 2022, pp. 300–316.

[7] J. Slemenšek, I. Fister, J. Geršak, B. Bratina, V. M. van Midden, Z. Pirtošek, and R. Šafarič, "Human gait activity recognition machine learning methods," *Sensors*, vol. 23, no. 2, p. 745, 2023.

[8] P. Khera and N. Kumar, "Role of machine learning in gait analysis: a review," *Journal of Medical Engineering & Technology*, vol. 44, no. 8, pp. 441–467, 2020.

[9] J. Hummel, M. Schwenk, D. Seebacher, P. Barzyk, J. Liepert, and M. Stein, "Clustering approaches for gait analysis within neurological disorders: a narrative review," *Digital Biomarkers*, vol. 8, no. 1, pp. 93–101, 2024.

[10] F. Horst, S. Lapuschkin, W. Samek, K.-R. Müller, and W. I. Schöllhorn, "Explaining the unique nature of individual gait patterns with deep learning," *Scientific Reports*, vol. 9, no. 1, p. 2391, 2019.

[11] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *preprint arXiv:1703.07737*, 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[13] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research (JMLR)*, vol. 10, no. 2, 2009.

[14] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2014.

[15] S. H. Holzreiter and M. E. Köhle, "Assessment of gait patterns using neural networks," *Journal of Biomechanics*, vol. 26, no. 6, pp. 645–651, 1993.

[16] W. Schöllhorn, B. Nigg, D. Stefanyshyn, and W. Liu, "Identification of individual walking patterns using time discrete and time continuous data sets," *Gait & Posture*, vol. 15, no. 2, pp. 180–186, 2002.

[17] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2005.

[18] J. Lu and E. Zhang, "Gait recognition for human identification based on ica and fuzzy svm through multiple views fusion," *Pattern Recognition Letters*, vol. 28, no. 16, pp. 2401–2411, 2007.

[19] K. Bartol, D. Bojanić, and T. Petković, "Generalizable human pose triangulation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 11 018–11 027.

[20] H. Tu, C. Wang, and W. Zeng, "VoxelPose: Towards multi-camera 3D human pose estimation in wild environment," in *Europ. Conf. on Computer Vision (ECCV)*, 2020, pp. 197–212.

[21] H. Ye, W. Zhu, C. Wang, R. Wu, and Y. Wang, "Faster VoxelPose: Real-time 3D human pose estimation by orthographic projection," in *Europ. Conf. on Computer Vision (ECCV)*, 2022, pp. 142–159.

[22] T. Wang, J. Zhang, Y. Cai, S. Yan, and J. Feng, "Direct multi-view multi-person 3D pose estimation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 13 153–13 164.

[23] M. Alotaibi and A. Mahmood, "Improved gait recognition based on specialized deep convolutional neural network," *Computer Vision and Image Understanding*, vol. 164, pp. 103–110, 2017.

[24] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *18th International Conference on Pattern Recognition (ICPR)*, vol. 4, 2006, pp. 441–444.

[25] K. Taha, P. D. Yoo, Y. Al-Hammadi, S. Muhaidat, and C. Y. Yeun, "Learning a deep-feature clustering model for gait-based individual identification," *Computers & Security*, vol. 136, p. 103559, 2024.

[26] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[27] S. Fink, M. Suppanz, J. Oberzaucher, M. A. Castro, O. Fernandes, and I. Alves, "Gait characterization in rare bone diseases in a real-world environment-a comparative controlled study," *Gait & Posture*, 2024.

[28] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.

[29] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[30] J. Tanke, C. Zaveri, and J. Gall, "Intention-based long-term human motion anticipation," in *International Conference on 3D Vision (3DV)*, 2021, pp. 596–605.