View-Independent 3D Gait Recognition Using Sequence-Based Siamese Networks

Mikolaj Klimek* and Bogdan Kwolek

AGH University of Krakow, 30 Mickiewicza, 30-059 Kraków, Poland
{miklimek,bkw}@agh.edu.pl

Abstract. Gait recognition is a rapidly evolving area of computer vision research. In recent years, emphasis has been placed on the approaches based on multi-camera datasets and the three-dimensional data derived from them. However, a notable research gap persists between gait recognition results obtained from precise motion capture data and those achieved using marker-less approaches. The synchronized GPJATK dataset used in this study enables this issue to be addressed by validating methods that rely on approximated joint positions obtained via linear triangulation against ground-truth motion capture data. We propose a sequence-based Siamese framework for view-independent 3D gait recognition. Using low-dimensional representations derived from similarity learning, the proposed approach achieves a Rank-1 person identification of 91.79% on triangulated data and 98.74% when evaluated using ground-truth motion capture data.

Keywords: Computer vision · Gait recognition · Siamese neural networks · Motion capture · Triangulation.

1 Introduction

Walking is a fundamental human activity that is performed by nearly all individuals as a primary mode of locomotion. It is defined as a method of locomotion that involves the use of the two legs, alternately, to provide both support and propulsion. As the walking person moves forward, one limb act as a source of support while the other limb moves itself to a new support site. Then the limbs reverse their roles. The literature identifies numerous variables that can influence gait characteristics, including illnesses, injuries and fatigue, social and cultural context, type of footwear, training and muscle development, and the terrain on which the person is moving. This manner or style of walking is defined as gait whereas a single sequence of events repeated by each limb multiple times is called a gait cycle [4–6].

Due to its unique characteristics, gait analysis might be useful in distinguishing one person from another [12]. Unlike other popular biometric methods, such as facial and fingerprint recognition, vision-based gait recognition can be performed from a greater distance. This approach can be employed, for instance, to

* Corresponding author: miklimek@agh.edu.pl

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

identify criminals responsible for offenses captured on surveillance cameras, while individual elements of gait analysis may prove useful in medicine, physiotherapy, sports, gaming and human-robot interactions [15].

The existing literature categorizes gait recognition methods into two main groups: model-based (skeleton-based) and appearance-based (model-free) approaches [13–15]. The main difference lies in the methodological approach – the first method utilizes extracted key joints of a skeletal model, whereas the second relies on the full body shape or extracted silhouettes. Gait recognition methods based on skeletal representations have a potential to be more robust to variations in viewpoint and clothing than those relying on silhouette-based approaches. However, accurately estimating the positions of the joints of a moving person remains a challenging task.

In the field of gait-based person identification, popular methods include Convolutional Neural Networks [23, 29], Graph Convolutional Networks [26], Recurrent Neural Networks [28], Long Short-Term Memory networks [25], and architectures based on Transformers [27, 29]. State-of-the-art approaches are capable of achieving Rank-1 accuracy significantly exceeding 90% on large-scale datasets. For a long time, gait recognition relied primarily on 2D representations of gait; however, recently there has been a shift towards methods using 3D data, which are more robust to covariate factors present in real-life scenarios, such as different clothing, camera angles, or more complex environments [21]. However, creating a solution that is completely independent of such external factors remains a challenge [15].

While 3D gait datasets exist – collected using depth sensors such as Kinect [19] or motion capture systems [20–22] – research in this field typically uses larger, more widely recognized multi-camera datasets [16–18]. These datasets allow for the determination of 3D human poses through various techniques, such as triangulation and machine learning models. A research gap still exists between studies based on datasets containing precise 3D ground-truth data and those employing 3D data reconstructed from images using alternative approaches.

The GPJATK [1] synchronized dataset used in this study provides a unique setting by incorporating multi-view video data from four distinct perspectives along with synchronized motion capture data. This specific structure enables the validation of video-derived joint positions against ground-truth motion capture data for the same subjects. The GPJATK dataset makes it also possible to evaluate the accuracy of a model pretrained on motion capture ground-truth data against 3D data obtained through triangulation or modern machine learning techniques. This dataset is relatively small, as it comprises only 32 participants, with 4 to 10 recorded sequences per individual. Despite being challenging, training models on such limited datasets might be valuable, given the difficulty of collecting large-scale datasets with accurately annotated 3D ground-truth joint positions. To date, no effective neural network-based methods have been developed for view-independent 3D gait recognition based on comparable scale datasets. Previous work including the GPJATK dataset has focused mainly on minimizing joint position estimation errors and performing identification using

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

popular machine learning algorithms, often incorporating PCA-based components [1, 2].

2 Background and Relevant Work

In this research, the GPJATK dataset [1] was used. It offers a unique integration of two data sources, containing both synchronized video data from four perspectives for each sequence and corresponding three-dimensional data acquired through professional motion capture techniques. It comprises 166 gait sequences recorded from 32 individuals. Each sequence includes four RGB video streams (in AVI format) and motion capture data acquired using markers attached to the subjects. The video streams were recorded at a frame rate of 25 frames per second with a resolution of 960×540 pixels, while the motion capture data was collected at a rate of 100 frames per second. In addition to the recordings, the dataset also includes intrinsic and extrinsic camera parameters, allowing the triangulation to be performed.

For six of the 32 participants, clothing changes were introduced during the data collection process. In these cases, the gait sequences recorded with the first clothing set are designated as sequences 1-4, whereas the sequences recorded with the second clothing set are designated as sequences 5-8. For each clothing set, four walking sequences were recorded: straight and diagonal walks, each from right to left and from left to right. In addition, for seven participants, additional sequences were recorded in which individuals wore backpacks. Motion capture data are not available for these particular sequences, so these sequences were not used in this investigation.

The GPJATK dataset has been thoroughly examined with respect to its applicability for 2D gait analysis in research [3]. The authors demonstrated that pose estimation using OpenPose can provide estimates of human gait parameters with sufficient accuracy to detect changes in gait patterns. The reported results indicate a temporal difference of 0.02 s for gait cycle parameters and 0.049 m for step length. Moreover, the mean error in sagittal plane hip and knee angles between motion capture data and OpenPose estimates was 4.0° and 5.6°, respectively. Other widely adopted pose estimation algorithms, such as YOLO, have not been assessed in this context.

Using the GPJATK dataset, previous work [1], which employed Annealed Particle Swarm Optimization (APSO), reports mean 3D pose estimation errors of 36.8 mm for the head, 37.7 mm for the torso, 43.4 mm for the tibiae, and approximately 28.5 mm for the forearms, relative to ground-truth motion capture data. However, studies evaluating the accuracy of gait feature extraction have not been conducted for 3D data obtained via triangulation or other available methods. No applications of this dataset for neural network-based person recognition are known – previous research has made use of Principal Component Analysis [1], Swarm intelligence [2], and other classical techniques.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

3 Methods

3.1 Contribution

The primary contributions of this research involve the development of an effective neural network-based method for view-independent 3D gait recognition, which utilizes low-dimensional representations of gait cycles obtained through similarity learning. It was evaluated using the publicly available GPJATK dataset. Furthermore, the study provides a comparative analysis between results obtained from motion capture-based data and those derived from triangulated 3D data, focusing on strategies to minimize the performance gap between these two distinct approaches. The proposed view-independent 3D gait recognition pipeline follows a structured five-step process summarized in Fig. 1:

- **Pose detection** – utilizing YOLO26 to detect joint positions within video sequences.
- **3D pose reconstruction** – applying linear triangulation based on camera parameters to obtain 3D joint coordinates.
- **Feature extraction** – calculation of gait features from the acquired 3D positions
- **Similarity learning** – training a Siamese neural network (with 1D-CNN or LSTM backbone) to quantify similarity between gait patterns.
- **Identification** – performing rank person identification based on the learned Siamese embeddings.

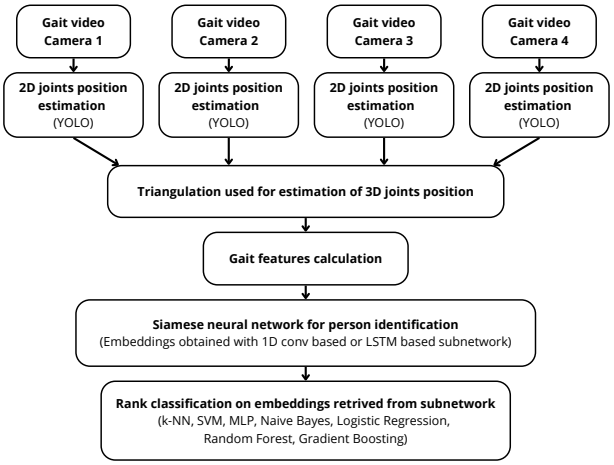


Fig. 1: Computational pipeline for gait analysis based on synchronized views from four cameras.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

3.2 Pose Detection and Triangulation

For each video in every gait sequence, the positions of key joints were estimated using the largest available YOLO26 [11] model for pose estimation. The human pose model utilized in motion capture (MoCap) differs slightly from the one used in YOLO26 pose estimation. Skeletons for both data sources are shown in Fig. 2a. The main differences involve the spine and toe markers, which are present in the MoCap datasets but are not included in YOLO skeletal model. In contrast, YOLO provides more detailed head markers, featuring distinct points for the nose, eyes, and ears. Despite these differences, both models share a core set of markers – shoulders, hips, knees, ankles, elbows, and wrists – which form the basis for feature extraction in this study, with their positions approximately coinciding across models (e.g., the MoCap foot markers are placed slightly lower than the YOLO ankle markers, but this should not have a significant impact on the calculated parameters).

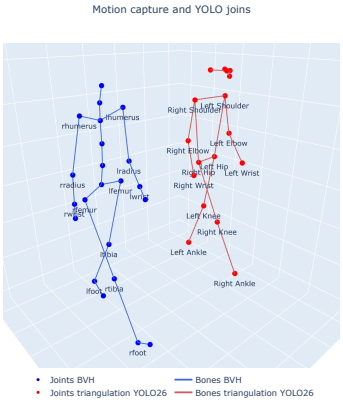
Taking advantage of the fact that each gait sequence was recorded using four synchronized cameras and that the corresponding camera parameters are available, linear triangulation [8] was applied to estimate the 3D coordinates of each joint. The dataset provides the parameters required to construct a Tsai camera model [7], including geometry parameters defining sensor dimensions, intrinsic parameters that specify internal optical properties (focal length, radial distortion, principal point, and skew), and extrinsic parameters that describe the camera's translation and rotation. The intrinsic parameters form the camera calibration matrix, while the extrinsic parameters define the rotation and translation matrices that compose the extrinsic matrix. The camera projection matrix combines both intrinsic and extrinsic parameters and can be used to transform a 3D point in world coordinates directly into 2D homogeneous image coordinates. The mean triangulation errors for each joint, compared to it's position in motion capture data are presented in Table 1.

Triangulation based on YOLO detections most accurately predicts the positions of the ankles, elbows, and knees, with a mean error below 70 mm compared to the motion capture reference data. In contrast, wrist positions are estimated with the lowest accuracy, with a mean error of nearly 100 mm. The average error in the position of the right and left joints is the biggest for the elbow and is approximately 6 mm.

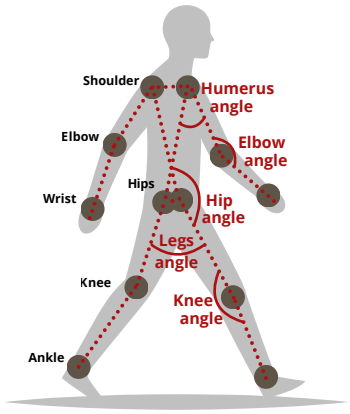
Table 1: Comparison of mean triangulation errors for selected joints (in millimeters).

Joint	Left	Right
Shoulder	76.06	77.73
Hip	82.77	80.61
Knee	68.66	68.19
Ankle	57.15	56.89
Elbow	60.92	54.88
Wrist	99.79	95.75

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35



Comparison of motion capture and YOLO26.



Gait angles and joints position schema.

Fig. 3: Gait features and skeletal comparison.

3.3 Gait Features Extraction

Based on previous studies [1, 2] and after confirming that a single step takes approximately 29 frames on average, each gait sequence was segmented – using foot height – into subsequences representing an approximate full stride (a complete gait cycle defined as the interval between two successive initial contacts of the same foot), with a fixed length of 32 frames. Additional experiments were conducted using a Butterworth filter applied to each dimension of the triangulated joint positions. The best results were shown to be obtained using a relatively aggressive filter with a cutoff frequency of 0.625 and a second-order filter. Less aggressive filtering produced results closer to those obtained without filtering, while stronger filtering reduced the visibility of temporal variations in gait features.

In gait recognition, a person-independent descriptor is a feature representation that captures a person's walking characteristics while remaining robust to external covariate factors such as different clothing, carrying conditions, or viewing angles. To represent gait characteristics using this descriptor, the following features were extracted:

- **Change in center of gravity height** – change in the distance between the center of gravity (defined as the center of the pelvis) and the ground.
- **Lateral pelvic tilt** – angle in the frontal (coronal) plane between a global horizontal axis and the transverse line connecting the two hip joints.
- **Pelvis rotation** – angle in the transverse (horizontal) plane between the participant trajectory and the line connecting the hip joints.
- **Distances between pairs of joints** – ankles, knees, elbows, and wrists.
- **Joint angles** – left and right knee, hip, humerus, and elbow angles.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

– **Leg angle** – defined as the angle between the extensions of the femur lines projected onto the sagittal plane.

Figure 2b illustrates the key joint positions along with their angular features described above on the sagittal plane. These parameters were calculated for each of the 32 frames within a stride, resulting in a feature representation of size 32×16 for each gait sequence. In addition, a reduced feature subset was considered that excluded upper limb joints. In this case, the dataset did not include distances between the wrist and elbow pairs, nor the humerus and elbow angle features, resulting in respectively smaller gait representation.

Given that the motion capture data was recorded at 100 Hz and the corresponding videos at 25 Hz, a comparable reference dataset was obtained by resampling the motion capture data. Every fourth frame was selected, with an appropriate offset, based on the metadata provided with the dataset.

To assess the impact of three-dimensional information on gait recognition, similar features were also extracted from 2D joints position in sagittal plane. These were computed using YOLO detections from a single camera view in which the participant walked parallel to the camera. However, in this analysis, the pelvis rotation and pelvic tilt features had to be omitted.

3.4 Representation of Person Similarity Using Siamese Neural Networks

To address the challenge of gait sequence similarity, Siamese Neural Networks [9] (SNN) were used. The SNN architecture processes two inputs, each represented as a matrix with the first dimension corresponding to the sequence length (32) and the second to the number of features, through two identical backbone subnetworks with shared weights. These subnetworks map the inputs into 32-dimensional embedding vectors, which are then compared using the Euclidean distance metric. The standard size of embedding equal to 32 has been used in this research - larger embeddings might slightly improve the quality of person identification, but may result in lower person similarity performance. If the distance between two embeddings falls below a predefined threshold (set to 0.5 during all experiments both in training and evaluation phase), the sequences are classified as belonging to the same participant. A lower threshold used for evaluation improves model precision but reduces recall, whereas a higher threshold increases recall at the expense of precision, so it could be adjusted depending on the needs. In this research, two distinct architectures were evaluated as the Siamese backbone:

– **1D Convolutional (1D-CNN) based neural network** model employs a series of three 1D convolutional layers, followed by adaptive average pooling and a final dense layer to produce the embedding.
– **Long Short-Term Memory (LSTM) based network** [24, 25] utilizes a two-layer LSTM with dropout to extract temporal features from the sequential data. The final hidden state is then mapped to the embedding space via a single dense layer.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

Both models were trained using a contrastive loss function [10], which is defined as $L = (1 - y) \cdot D^2 + y \cdot \max(0, m - D)^2$, where:

- $y \in \{0, 1\}$ is the label (0 = similar, 1 = dissimilar),
- D is the distance between items in pair (Euclidean distance in this case),
- m is a margin value that defines the minimum distance required between dissimilar pairs.

This approach encourages the network to minimize the distance between embeddings of similar pairs, while ensuring that dissimilar pairs are separated by at least a specified margin.

During the training of Siamese neural networks, the Adam optimizer was used with batches consisting of 32 pairs. Experimental results indicate that embeddings more suitable for gait recognition can be obtained even with relatively low learning rate values. For 3D data (both ground-truth and triangulated), the Conv1D-based Siamese neural network was trained with a learning rate of 1e-6 for 100 epochs, while the LSTM-based Siamese neural network used a learning rate of 1e-5 and was trained for 150 epochs. For 2D data, the best performance was achieved with a learning rate of 1e-5 for the Conv1D backbone and 1e-4 for the LSTM backbone. In both 2D cases, the training process was completed within 20 epochs.

4 Experimental Results

4.1 Evaluation Protocol

For evaluation, we used the dataset comprising all sequences for which motion capture data were collected (152 sequences, split into individual strides as described earlier). The Siamese Neural Network models were evaluated using 4-fold cross-validation. Given that the dataset consists of sequences from 32 participants, each fold was partitioned so the training set contained 24 participants, while the remaining eight were reserved for test dataset.

To ensure a robust learning process, the following pairing strategy was implemented – for each participant positive pairs were generated by creating all possible pairs with samples presenting this participant. To obtain a balanced dataset, the same number of negative pairs as positive pairs were created by randomly selecting a sample representing this participant and a sample representing another participant from the same set. This procedure was applied to each participant in the selected dataset. To increase the versatility of the model and prevent overfitting, the negative pairs in the training set were regenerated at the beginning of each training epoch. Since the problem at this stage reduces to binary classification on a balanced dataset, Accuracy, Precision, Recall, and the Area Under the ROC Curve (AUROC) were calculated for each classifier in each fold using a test set with randomly generated negative pairs.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

Following the training of the Siamese Neural Network (SNN) for person similarity prediction, the trained backbone can be utilized to generate low-dimensional vector embeddings representing the gait sequence. Using these embeddings, individual identification can be performed using standard machine learning models, including K-Nearest Neighbors (k-NN), Support Vector Classifier (SVM), Multi-Layer Perceptron (MLP), Gaussian Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GradBoost). The Correct Classification Rate (CCR) was calculated for each classifier. CCR is defined as the ratio of correctly classified cases to the total number of test set samples. In addition to the standard Rank-1 measure, Rank-2 and Rank-3 rates were determined, indicating the percentage of correct identifications appearing within the top two and top three predictions of classifier, respectively. This evaluation procedure was applied to each test set during the SNN training process, employing 4-fold cross-validation.

4.2 Person Similarity Performance

Table 2 presents the performance results obtained during training using the full feature set, including those derived from upper limb marker positions. Baseline training with motion capture (MoCap) data demonstrates that presented approach accurately determines whether two sequences represent the same individual with an accuracy of 99.38% using a 1D convolution based backbone and 97.99% with an LSTM subnetwork.

Table 2: Person similarity performance using all features calculated from ground-truth, triangulated 3D and raw 2D data.

	1D Conv				LSTM			
Data type	**Accuracy**	**Precision**	**Recall**	**AUROC**	**Accuracy**	**Precision**	**Recall**	**AUROC**
MoCap data (Ground truth)	99.38	99.96	98.80	100.0	97.99	96.14	100.0	99.94
Triangulated data	95.72	99.88	91.54	99.55	95.12	92.02	98.80	99.22
Triangulated data, Butterworth	84.29	99.44	68.96	99.17	95.79	92.23	100.0	99.60
2D data	90.61	91.65	89.35	97.92	72.76	68.41	84.55	85.28
2D data, Butterworth	81.93	83.78	79.18	92.04	70.39	66.71	81.43	83.31

When using the triangulated dataset, the accuracy of similarity predictions decreases by 3.66 percentage points for the Siamese neural network with a 1D-CNN backbone and by 2.87 percentage points for the LSTM based version. Further data smoothing via a Butterworth filter results in a decline in prediction accuracy by more than 10 percentage points for 1D-CNN and an increase of a few tenths of a percentage point for the LSTM backbone. For 1D-CNN backbone and Butterworth-smoothed data, a large number of false negatives is observed,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

resulting in a low recall of 68.96%. All models evaluated on 3D data achieved an AUROC exceeding 99%.

For 2D data, the 1D convolution based backbone shows a significant performance drop – accuracy is 5.11 percentage points lower on raw data compared to 3D results, with an additional decrease following Butterworth filtering. With the LSTM backbone, the accuracy gap between triangulated 3D data and 2D data is even larger and exceeds 20 percentage points, both with and without the Butterworth filter.

Table 3: Person similarity performance using only features not involving hands calculated from ground-truth, triangulated 3D and raw 2D data.

	1D Conv				LSTM			
Data type	**Accuracy**	**Precision**	**Recall**	**AUROC**	**Accuracy**	**Precision**	**Recall**	**AUROC**
MoCap data (Ground truth)	97.92	98.15	97.68	99.87	97.02	94.38	100.0	99.81
Triangulated data	90.29	98.68	81.67	98.23	92.01	86.98	98.80	98.57
Triangulated data, Butterworth	85.64	93.22	76.86	96.85	94.31	89.79	100.0	99.02
2D data	84.97	83.30	87.47	95.31	82.36	77.48	91.23	89.72
2D data, Butterworth	79.96	79.58	80.59	90.58	75.95	70.34	89.73	86.98

Table 3 summarizes the results obtained using the same methodology, but with upper limb joint features excluded from the datasets. For the SNN with a 1D convolution backbone, accuracy was 1.46 percentage points lower using MoCap data, decreased by 5.43 for raw triangulated data, and increased by 1.35 percentage points after Butterworth filtered was applied.

In the case of the LSTM backbone, performance across all data types was slightly worse than with all features, with the largest difference occurring for raw triangulated data at more than 2 percentage points. Excluding upper limb related features led to improved performance on 2D data for the LSTM backbone, while degrading the performance of the 1D-CNN backbone.

Figure 5 compares the t-SNE visualizations of test set embeddings across one cross-validation iteration, highlighting three dataset types (motion capture, raw triangulated data, and triangulated data with a Butterworth filter) and two subnetwork architectures (1D convolution and LSTM based).

4.3 Persons identification based on Siamese neural network embeddings

Table 4 summarizes the person identification results obtained using the motion capture dataset. When utilizing the full feature set and the 1D convolution based backbone, the highest Rank-1 scores were achieved by the Logistic Regression classifier (98.74%) and the MLP (98.50%). Most classifiers, except k-NN, Naive

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

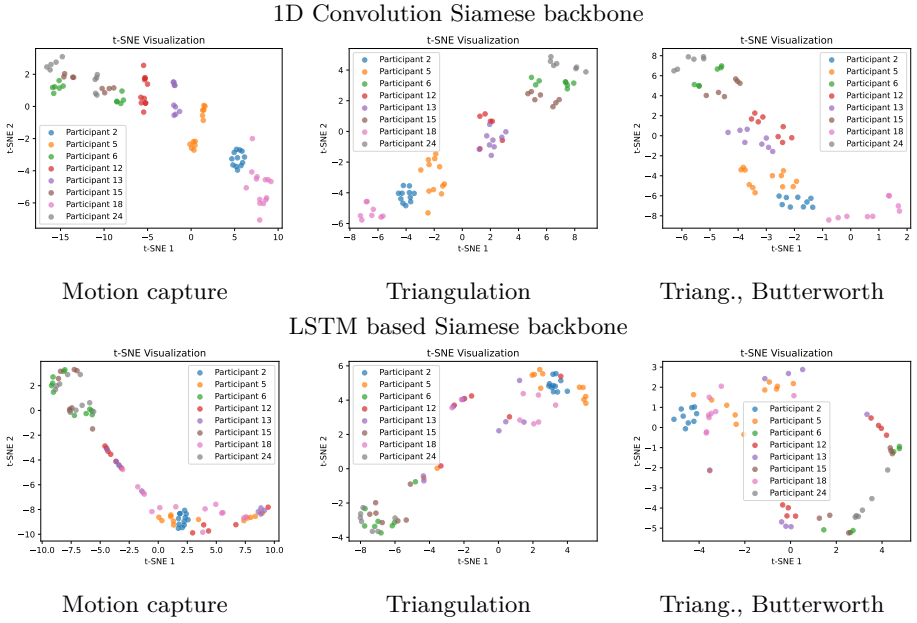


Fig. 5: t-SNE visualization of Siamese embeddings (x/y: 2D t-SNE coordinates); better-performing models show clearer participant separation.

Bayes, and Gradient Boosting, achieved at least 99% Rank-2, while only Gradient Boosting remained below 99.5% at Rank-3. In contrast, when the LSTM based backbone was employed on the same dataset, the maximum CCR reached 73.66% with the MLP. This classifier also proved to be the most effective for Rank-2 and Rank-3, achieving 89.31% and 95.79%, respectively. Other classifiers performed significantly worse. When training without wrist and elbow joints related features, the Rank-1 score for the 1D-CNN embeddings dropped to 97.27%, while the LSTM backbone dropped to 71.11%.

The left side of Table 5 presents the training results for the triangulated data (both raw and Butterworth-filtered). Using embeddings from the 1D-CNN subnetwork, the raw data achieve a CCR of 89.90% with Logistic Regression, Rank-2 scores reaching almost 98% with the same classifier, and Rank-3 scores exceeding 99% for SVM, LR and RF. After applying the Butterworth filter, the Random Forrest classifier achieved a CCR of 91.79%. LR and MLP achieved more than 98% in Rank-2, while both methods, together with RF, exceeded 99% in Rank-3. The highest Rank-1 score obtained with triangulated data is 6.95 percentage points lower than the best result achieved using the motion capture dataset.

The presented method outperforms the techniques described in [1] and [2], where the highest Rank-1 values reached 80.13% and 87.05%, respectively, using datasets without participants after changing clothes. As mentioned previously, the dataset used in this research includes all individuals from the GPJATK

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

Table 4: Top-K person identification comparison on embeddings obtained from motion capture data.

	All features					
	1D Conv			LSTM		
Model	**Rank-1**	**Rank-2**	**Rank-3**	**Rank-1**	**Rank-2**	**Rank-3**
k-NN	90.78	98.50	99.75	60.90	82.90	91.81
SVM	98.00	99.51	99.76	54.75	77.37	92.08
MLP	98.50	**100.0**	**100.0**	**73.66**	**89.31**	**95.79**
NB	92.53	95.77	99.51	52.99	75.89	91.08
LR	**98.74**	**100.0**	**100.0**	69.64	84.85	95.06
RF	97.50	99.75	**100.0**	70.64	88.82	95.03
GradBoost	87.80	92.55	95.03	64.16	84.32	90.81
	Features not involving hands					
	1D Conv			LSTM		
Model	**Rank-1**	**Rank-2**	**Rank-3**	**Rank-1**	**Rank-2**	**Rank-3**
k-NN	86.61	98.75	99.50	52.94	78.87	86.82
SVM	93.53	98.76	**99.51**	51.48	74.89	85.83
MLP	95.52	98.99	99.50	**71.11**	**84.06**	90.79
NB	90.34	95.78	98.75	45.00	69.94	79.60
LR	**97.27**	**99.00**	99.50	66.85	82.81	**91.04**
RF	93.30	97.27	98.75	66.40	82.09	88.54
GradBoost	80.36	89.52	92.25	60.41	76.36	85.33

dataset for whom motion capture ground-truth validation was possible, including several individuals recorded in two different clothing sets.

Consistent with the motion capture results, the person identification performance on triangulated data using the LSTM subnetwork was significantly lower than that of the 1D-CNN, peaking at 52.91% CCR for raw data and 65.22% for filtered data. The performance gap between datasets with and without upper-limb features is greater in triangulated data than in motion capture data. For triangulated dataset with reduced features set, the highest achieved Rank-1 was 74.02% (using raw data and the 1D convolution based backbone).

The right side of Table 5 presents the CCR results obtained using two-dimensional data. In this configuration, performance metrics are markedly lower, with no classifier reaching a 50% Rank-1 score. The peak performance of 47.66% was achieved on raw data using the 1D-CNN backbone. Interestingly, for 2D data, the performance difference between the full feature set and the set excluding upper limb features is significantly less visible than in the 3D data scenarios.

5 Summary and Conclusions

In this work, a model-based, view-independent 3D gait recognition approach based on Siamese neural networks with two types of backbone architecture is proposed and evaluated with triangulated and ground-truth motion capture data. The experimental results yield the following key conclusions:

– The proposed Siamese based approach demonstrates high effectiveness in distinguishing between motion sequences belonging to different individuals.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

Table 5: Top-K person identification comparison on embeddings obtained from YOLO-based data. The first column reports results for 3D triangulated data, while the second column shows results for raw 2D data.

Triangulated data

All features

	1D Conv			LSTM		
Model	**Rank-1**	**Rank-2**	**Rank-3**	**Rank-1**	**Rank-2**	**Rank-3**
k-NN	76.12	92.95	98.46	45.94	67.87	80.14
SVM	83.47	96.06	99.09	49.27	71.32	84.13
MLP	86.14	96.92	98.77	51.06	72.22	86.17
NB	81.24	92.96	97.23	44.41	68.20	85.04
LR	**89.90**	**97.87**	99.39	48.92	73.38	**87.77**
RF	84.01	96.30	**99.68**	**52.91**	**73.68**	87.10
GradBoost	70.23	81.52	87.73	48.62	66.64	76.49

All features, Butterworth filter applied

	1D Conv			LSTM		
Model	**Rank-1**	**Rank-2**	**Rank-3**	**Rank-1**	**Rank-2**	**Rank-3**
k-NN	79.54	94.66	98.37	49.41	76.18	87.05
SVM	85.78	94.74	97.71	41.46	72.53	90.34
MLP	90.45	98.03	99.01	**65.22**	**80.84**	90.43
NB	88.85	95.77	97.73	51.04	71.23	88.16
LR	91.41	**98.37**	**99.04**	57.64	78.55	**91.72**
RF	**91.79**	97.39	99.01	63.22	80.14	90.45
GradBoost	73.27	85.77	90.42	56.75	73.62	80.60

Features not involving hands

	1D Conv			LSTM		
Model	**Rank-1**	**Rank-2**	**Rank-3**	**Rank-1**	**Rank-2**	**Rank-3**
k-NN	54.47	77.70	87.18	28.07	45.83	61.03
SVM	62.02	83.18	93.83	26.37	41.95	55.70
MLP	71.25	**85.61**	90.54	32.49	52.29	65.14
NB	60.46	77.65	87.39	21.38	36.70	54.09
LR	**74.02**	85.32	**93.25**	**33.49**	**55.39**	**67.31**
RF	62.31	79.78	89.31	30.89	50.14	61.98
GradBoost	55.37	69.46	75.85	25.72	44.31	58.99

Features not involving hands, Butterworth

	1D Conv			LSTM		
Model	**Rank-1**	**Rank-2**	**Rank-3**	**Rank-1**	**Rank-2**	**Rank-3**
k-NN	57.11	78.58	89.89	41.94	64.96	76.80
SVM	58.06	81.89	90.82	38.56	58.02	72.00
MLP	59.12	82.51	91.41	46.45	**69.25**	78.81
NB	57.48	83.27	91.15	36.69	61.30	77.21
LR	**63.30**	**86.18**	**94.40**	47.53	69.20	**83.06**
RF	62.11	84.58	93.77	**49.57**	70.86	79.37
GradBoost	56.15	72.35	78.32	41.50	60.72	74.53

YOLO 2D data

All features

	1D Conv			LSTM		
Model	**Rank-1**	**Rank-2**	**Rank-3**	**Rank-1**	**Rank-2**	**Rank-3**
k-NN	38.13	58.41	70.61	28.34	44.87	55.92
SVM	29.35	49.82	60.21	18.27	31.66	45.01
MLP	41.50	64.04	76.79	37.68	48.91	58.86
NB	30.35	52.89	65.16	24.88	47.73	59.45
LR	**47.66**	**67.04**	**83.50**	**38.85**	**51.01**	**62.22**
RF	40.31	60.84	72.93	28.98	47.71	58.21
GradBoost	35.36	52.00	64.74	26.00	39.83	57.46

All features, Butterworth filter applied

	1D Conv			LSTM		
Model	**Rank-1**	**Rank-2**	**Rank-3**	**Rank-1**	**Rank-2**	**Rank-3**
k-NN	34.83	52.87	62.40	34.81	48.19	63.44
SVM	31.39	46.26	58.97	24.30	39.53	55.23
MLP	44.83	60.99	**72.91**	35.86	58.62	71.49
NB	38.99	54.64	65.67	35.07	51.77	64.28
LR	**46.03**	**61.32**	72.40	35.16	**58.98**	**72.83**
RF	41.46	57.73	71.56	**41.39**	57.08	66.61
GradBoost	41.44	55.23	63.39	31.93	46.12	62.27

Features not involving hands

	1D Conv			LSTM		
Model	**Rank-1**	**Rank-2**	**Rank-3**	**Rank-1**	**Rank-2**	**Rank-3**
k-NN	33.51	50.69	68.55	35.47	57.57	71.24
SVM	27.04	44.65	56.82	29.25	48.30	61.45
MLP	40.98	63.52	76.82	**45.44**	**66.92**	**77.38**
NB	35.17	53.83	67.80	30.64	55.80	70.32
LR	**43.64**	**63.73**	**77.52**	41.17	64.76	74.61
RF	35.76	51.22	69.66	44.82	64.67	77.30
GradBoost	31.28	49.62	59.10	39.81	54.95	67.52

Features not involving hands, Butterworth

	1D Conv			LSTM		
Model	**Rank-1**	**Rank-2**	**Rank-3**	**Rank-1**	**Rank-2**	**Rank-3**
k-NN	32.32	45.84	56.72	27.99	43.68	64.69
SVM	24.83	43.41	57.66	25.65	41.35	55.72
MLP	31.01	48.04	61.05	32.40	**59.99**	**74.33**
NB	29.45	48.03	60.10	28.11	48.18	60.58
LR	**37.75**	**54.33**	**68.14**	**36.79**	55.20	70.13
RF	32.87	48.07	55.80	30.95	53.89	69.18
GradBoost	28.46	48.69	62.59	28.65	44.75	59.51

– Segmenting gait sequences into individual steps and training a Siamese neural network with dynamically generated negative pairs enables effective learning even when using a relatively small dataset.
– The transition from 2D sagittal plane data to 3D features (via triangulation or motion capture) results in a substantial improvement in recognition performance – differences of up to several tens of percentage points are observed for the proposed datasets.
– Low-dimension embeddings generated by the Siamese neural network backbones are highly effective for person identification, and the proposed solution remains robust to covariate factors, including different camera positions and participant clothing. The 1D convolution based backbone proves to be a significantly more robust feature extractor than the LSTM based alternative,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

achieving higher performance across almost all scenarios. Applying Butterworth filter to triangulated data also has a positive impact on the obtained results and allows for the minimization of the discrepancy between the results obtained from triangulated 3D data and those derived from motion capture ground-truth data.
– Among the machine learning models evaluated for person identification based on Siamese low-dimensional embeddings, the Logistic Regression frequently delivers the most accurate results. Nevertheless, Random Forest and Multi-Layer Perceptron also emerged as competitive and viable options for this task. Good performance is difficult to obtain without incorporating features related to the upper limb.

Acknowledgements

This work was supported by Polish National Science Center (NCN) under research grant 2024/55/B/ST6/01580.

References

1. Kwolek, B., Michalczuk, A., Krzeszowski, T. et al. (2019). "Calibrated and synchronized multi-view video and motion capture dataset for evaluation of gait recognition." Multimed. Tools Appl., 78, 32437–32465.
2. Krzeszowski, T., Wiktorowicz, K. (2020). "Combined Regularized Discriminant Analysis and Swarm Intelligence Techniques for Gait Recognition." Sensors 2020, 20, 6794.
3. Stenum J, Rossi C, Roemmich RT. (2021) "Two-dimensional video-based analysis of human gait using pose estimation. PLoS Comput Biol."
4. Whittle, M. W. (2014). "Gait analysis: an introduction." Butterworth-Heinemann.
5. Dicharry, Jay. (2010). "Kinematics and Kinetics of Gait: From Lab to Clinic." Clinics in sports medicine. 29. 347-64.
6. Kharb, Ashutosh & Saini, Vipin & Jain, Y & Dhiman, Surender & Tech, M & Scholar,. (2011). "A review of gait cycle and its parameters." IJCEM Int. J. Comput Eng Manag. 13.
7. Tsai, R. Y. (1985). "An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision." EEE Int. Conf. on Comp. Vis. and Patt. Rec.
8. Hartley R, Zisserman A. (2004) Multiple View Geometry in Computer Vision. 2nd ed. Cambridge University Press.
9. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I.M., LeCun, Y., Moore, C., Säckinger, E., & Shah, R. (1993). "Signature Verification Using A "Siamese" Time Delay Neural Network." Int. J. Pattern Recognit. Artif. Intell., 7, 669-688.
10. Hadsell, Raia & Chopra, Sumit & Lecun, Yann. (2006). "Dimensionality Reduction by Learning an Invariant Mapping." 1735 - 1742. CVPR
11. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, (2016) "You Only Look Once: Unified, Real-Time Object Detection," in 2016 EEE Int. Conf. on Comp. Vis. and Patt. Rec. (CVPR), Las Vegas, NV, USA, pp. 779-788
12. Boyd, Jeffrey & Little, J.J. (2003). "Biometric Gait Recognition." LNCS, Springer. 3161. 19-42.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35

13. Kusakunniran, W. (2020). "Review of gait recognition approaches and their challenges on view changes." IET Biometrics, 9(6), 238-250.
14. C. Shen, S. Yu, J. Wang, G. Q. Huang and L. Wang. (2025)."A Comprehensive Survey on Deep Gait Recognition: Algorithms, Datasets, and Challenges," in IEEE Trans. on Biom., Beh., and Id. Sc., vol. 7, no. 2, pp. 270-292.
15. Sethi, D., Bharti, S., & Prakash, C. (2022). "A comprehensive survey on gait analysis: History, parameters, approaches, pose estimation, and future work." Artificial Intelligence in Medicine, 129, 102314.
16. Takemura, N., Makihara, Y., Muramatsu, D. et al. (2018). "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition." IPSJ T. Comput. Vis. Appl. 10, 4
17. C. Ionescu, D. Papava, V. Olaru and C. Sminchisescu, (2014). "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," in IEEE Ttrans. on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1325-1339.
18. Shiqi Yu, Daoliang Tan and Tieniu Tan, (2006). "A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition," 18th Int. Conf. on Pattern Recognition (ICPR'06), pp. 441-444
19. Andersson, V., & Araujo, R. (2015). "Person identification using anthropometric and gait data from kinect sensor." AAAI Conf. on AI, Vol. 29, No. 1.
20. Balazia, M., & Sojka, P. (2018). "Gait recognition from motion capture data." ACM Ttrans. on Multimedia Computing, Communications, and Applications (TOMM), 14(1s), 1-18.
21. J. Zheng et al. (2022). "Gait Recognition in the Wild with Dense 3D Representations and A Benchmark," in 2022 IEEE/CVPR, pp. 20196-20205
22. Vielemeyer, J., Tronicke, L., Schreff, L. et al. (2026). "A Full-Body Motion Capture Gait Dataset of Healthy Young Adults Walking Ramps Up and Down." Sc. Data 13, 18.
23. Alotaibi, M., & Mahmood, A. (2017). "Improved gait recognition based on specialized deep convolutional neural network." Computer Vision and Image Understanding, 164, 103-110.
24. Yucer, S., & Akgul, Y. S. (2018). "3D human action recognition with siamese-LSTM based deep metric learning."
25. Kwon, J.; Lee, Y.; Lee, J. (2021). "Comparative Study of Markerless Vision-Based Gait Analyses for Person Re-Identification." Sensors 2021, 21, 8208.
26. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., & Rigoll, G. (2021). "Gaitgraph: Graph convolutional network for skeleton-based gait recognition." In 2021 IEEE (ICIP), pp. 2314-2318.
27. H. -M. Hsu, Y. Wang, C. -Y. Yang, J. -N. Hwang, H. L. U. Thuc and K. -J. Kim, (2022). "GAITTAKE: Gait Recognition by Temporal Attention and Keypoint-Guided Embedding," IEEE (ICIP), Bordeaux, France, 2022, pp. 2546-2550
28. Ghosh, R. (2022). "A Faster R-CNN and recurrent neural network based approach of gait recognition with and without carried objects. Expert Systems with Applications", 205, 117730.
29. Mogan, J. N., Lee, C. P., Lim, K. M., Ali, M., & Alqahtani, A. (2023). Gait-CNN-ViT: Multi-model gait recognition with convolutional neural networks and vision transformer. Sensors, 23(8), 3809.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_35