Sequential deep learning models for human skeleton-based gait recognition

by

Zhanibek Darimbekov

Submitted to the School of Engineering and Digital Sciences
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

Apr 2022

Author: . . . .
Zhanibek Darimbekov
School of Engineering and Digital Sciences
Apr 29, 2022

Certified by . . . .
Nguyen Anh Tu
Assistant Professor
Thesis Supervisor

Certified by . . . .
Min-Ho Lee
Assistant Professor
Thesis Co-supervisor

Accepted by . . . .
Vassilios D. Tourassis
Dean, School of Engineering and Digital Sciences

Sequential deep learning models for human skeleton-based gait recognition

by

Zhanibek Darimbekov

Submitted to the School of Engineering and Digital Sciences
on Apr 29, 2022, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Gait is the walking posture and dynamics of a person and is considered a unique biometric of a person. Based on this biometric and using various algorithms one can recognize the person's identity with high precision. Although many conventional machine learning and deep learning methods that learn a person's identity based on their gait have been proposed, there still exist some typical limitations. While conventional machine learning methods are error-prone to the background noise, more advanced methods based on CNN models do not well capture the temporal dependencies in terms of inter-frame correlation for gait recognition. This work addresses the limitations of previous works by employing deep sequential models and demonstrating their effectiveness and efficiency in learning gait information using 3D human skeletal data. We propose a person identification model, that uses both the joint coordinates and features derived from them, including joint distance, joint orientation, and joint velocity. Sequential models based on Long Short-Term Memory and Transformer networks capture the spatial correlations of skeletal joints within a single frame and the temporal dependencies and dynamics of the joints throughout a sequence of frames. The effectiveness and efficiency of the proposed methods, the impact of data augmentation methods, a combination of derived gait features were studied and analyzed. The experimental results show that the proposed models achieve high person identification accuracy on the UPCV Gait (98.26%), and KS20 VisLab Multi-View (90.86%) datasets, which are competitive to the previous state-of-the-art methods.

Thesis Supervisor: Nguyen Anh Tu
Title: Assistant Professor

Thesis Supervisor: Min-Ho Lee
Title: Assistant Professor

Contents

Chapter 1

Introduction

1.1 Overview and motivation

Gait recognition refers to a complex of technologies and algorithms whose aim is to identify persons based on their body posture and its dynamics, also known as gait. Recognizing people using gait is more convenient than using other biometrics like face, retina, or fingerprint because can be done from distance without the active collaboration of people and it is difficult to imitate others' gait. One of the applications of gait-based person identification algorithms is video surveillance in public areas for security purposes. Such surveillance systems assist in preempting suspicious events, providing awareness to the security personnel, and re-identifying people after some time in other places.

1.2 Problem statement

Depending on the gait data modality, the latest competitive methods for gait-based person identification can be categorized into two groups, i.e. silhouette-based [4] [9] and skeleton-based [31]. Silhouette-based methods extract silhouette features directly from video sequences with human gait. However, this approach is sensitive to view angle and pose since information on some body parts can be missed when occluded by other body parts. Skeleton-based methods model human skeletal joints from images with the help of 3D depth cameras such as Kinect [15, 23] or human pose estimation algorithms such as OpenPose [3] and AlphaPose [10]. The human body is modelled as a set of coordinates of a human skeletal joints, which can consequently be used as a key feature for person identification. Although, this methods is computationally expensive and require high-quality data, after extracting skeletal data, one can work with a compact and yet efficient data in further processing steps. There are also hybrid methods [2] that integrate both silhouette and skeleton information for gait recognition.

However, the calculation and use of the mean and standard deviation features to capture temporal information and the overall architecture of CNN proposed in [15] may limit the performance of person identification in practice. First, the proposed person identification model requires a preliminary stage of extracting 3D joint coordinates using Kinect cameras. Second, the model size remains large, having 8.6 million parameters, which may lead to overfitting the training data, especially when the dataset is small-size and not challenging i.e. has few subjects and include the limited view variations.

1.3 Aims and objectives

In this work, we propose an alternative gait recognition method that also employs the human skeleton data. In particular, we process the features derived from the skeletal coordinates in a sequential manner benefiting from the architecture of the LSTM and Transformer networks. To the best of our knowledge, this is the first work investigating the use of a Transformer network for gait-based person identification using human skeleton data.

1.4 Key contributions

The key contributions of this work are the following:

- A compact sequential deep learning models for gait recognition using human skeletal data.

- Investigation of identifying ability of the gait features derived from raw joint coordinates.

Chapter 2

Related work

2.1 Conventional methods

The gait features are usually extracted from frame sequences within sample videos of a walking person. The approaches that make use of the raw image data and extract features directly from the image frames are very sensitive to the person's position, camera viewing angle, and scale. One particular example of such approach is Gait Energy Image (GEI) [12], which is a single image template constructed by averaging the binary silhouette over a range of frames. There are many variations of GEI as Frame Difference Energy Image (FDEI) [5], Pose Energy Image (PEI) [27] and Histogram Of Flow Energy Image (HOFEI) [24].

2.2 Deep learning-based methods

2.2.1 Convolutional neural networks

For the gait recognition tasks, the most recent works has applied several classifiers backed with deep learning architectures including several variations of CNNs, that has showed remarkable performance in other computer vision tasks. Majority of these works make use of gait information learned from image sequences in the form of 3D skeletal data with body joint coordinates collected from depth sensors. By manipulating these skeletal data, one can model the human movement by several simple geometric features, such as distance between the joints, joint orientation [15], and joint displacement.

2.2.2 Recurrent neural networks

Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) [13] are among the most popular and effective deep learning methods that learn the internal dependencies within a sequential data in wide range of applications [7,11,21,30]. Many works address the task of gait recognition by analyzing the gait as a sequence of features in a continuous video that captures a human movement. In particular, [28] proposed an LSTM-based human action learning model from skeletal data. First, the model splits the skeletal joints into five groups termed body parts, namely the torso, two hands, and two legs). Then it processes them in a sequential manner and captures the temporal information required for identifying an action. Each body part is processed by a separate LSTM that learns the temporal information required for identifying an action. In other work, [36] developed a convolutional Long Short-Term Memory (Conv-LSTM) for gait recognition based on RGB image sequences of human silhouettes. The method is constructed in such way that the refined silhouette features generated by convolutional layers are transformed into vector sequences as the input of LSTM layers. Although the model proves its effectiveness by beating the previous similar methods, there remains complexities regarding the size of the learning model.

2.2.3 Attention mechanism

The attention mechanism is proposed by [1] as an enhancement of the previous encoder decoder-based neural machine translation networks. Since then, it has been widely used in many other natural language processing [30, 33] and computer vision tasks [34, 37]. Recently, several works [6, 14, 16] have employed the attention mechanism for gait analysis. In particular, [14] proposed a lightweight convolutional neural network architecture for gait recognition using wearable devices that employ the attention mechanism that detects the important channels within the network and simplify the network complexity. In another work, [6] has proposed dual-stream neural network based on Vision Transformer (ViT) [8] to recognize people through radar gaits. Moreover, [16] has proposed convolutional neural network joint attention mechanism (CJAM) which combines a CNN and Transformer [33] networks, where consecutive image frames are encoded via the CNN and the outputs are then passed into the Transformer for gait classification.

In our work, we incorporate the former stages of the gait feature extraction techniques used in Huynh-The et al. with sequential learning models, including LSTM and Transformers, which in turn reduces the model complexity and decreases the training time. In contrast to [16], we feed the skeletal features directly to the Transformer network skipping the stage of encoding through CNN.

Chapter 3

Proposed methodology

3.1 Overview

In this section, we present the overall design of our proposed framework and describe the process of extracting gait features from $3D$ human skeletal joint coordinates and training sequential deep learning models build upon LSTM and Transformer networks for person identification. We extract several gait features from the raw joint coordinates, starting with normalizing and aligning the joint coordinates, followed by data augmentation steps. We then generate joint-based and geometric gait features. The former include the normalized joint coordinates and joint velocity features and the latter include joint distance and joint orientation features. Person identification models based on LSTM and Transformer are then trained on multiple different combinations of the extracted gait features. Finally, the trained models can be used to predict a person identity based on 3D skeletal data. The overall scheme of the proposed gait-based person identification method is shown in Figure 3-2.

3.2 Extracting gait features

Human skeletal joint coordinates can be collected either directly from $3D$ depth cameras or estimated from raw videos by human pose estimation algorithms as OpenPose [3] or AlphaPose [10]. In general, the acquired data are the $2D$ or $3D$ coordinates of a predefined set of skeletal joints, usually 18 to 25 joints, depending on the configurations of the 3D depth cameras and human pose estimation algorithms. Given a full set of skeletal joints $S = \{j_{i=1:n}\}$, where n is the number of body joints, each joint is defined either as a point with coordinates (x, y, z) in the $3D$ space R^3 or as a point with coordinates (x, y) in the $2D$ space R^2. For our experiments, we will use only 3D joint coordinates extracted through Microsoft Kinect depth cameras. $3D$ skeletal data are usually extracted using the Microsoft Kinect depth cameras and are available as sequences of body joint coordinates in $3D$ space. In this work, we use $3D$ skeletal data from UPCV Gait [17–19] and KS20 VisLab Multi-View Kinect Skeleton [23, 25] datasets.

Having extracted raw joint coordinates in a single frame, a person can be theoretically identified based on these gait features. However, the gait information in one frame is not enough to provide accurate identification and is very sensitive to small changes in joint positions. In order to build a more robust and effective identification model that captures temporal information and the gait dynamics, we derive a set of secondary gait features. In particular, we engage the intra-frame geometric features i.e. joint distance and joint orientation, and inter-frame temporal features i.e. joint velocity. Overall, in order to identify persons based on their gait, we set up four types of features, namely, normalized joint coordinates (JC), joint velocity (JV), joint distance (JD), and joint orientation (JO).

3.2.1 Normalized joint coordinates

We apply a preliminary normalization step to enforce the skeletal data to be in a single format and scale. We shift the joints in such a way that the coordinates of the joint for the 'base of the spine', labelled as 12 in Figure 3-1, is placed in the origin. Then, we align the skeleton by rotating the joints so that the vector connecting the joints labelled as 2 and 12 was parallel to Y-axis. Finally, we scale all the coordinates so that the distance between the joints labelled as 1 and 2 was unit size. Normalized joint coordinates (JC) will be used a baseline gait feature in experiments.

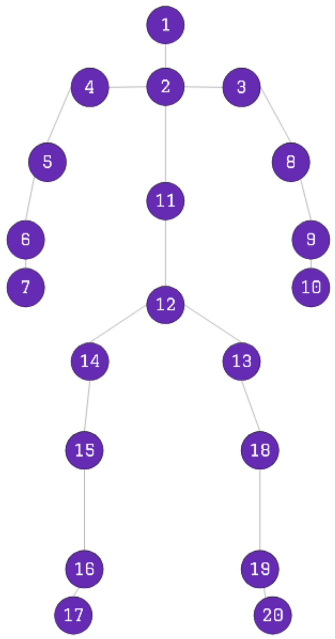


Figure 3-1: Model of a human skeleton represented as a set of 20 body joints estimated by using the Microsoft Kinect v1 depth sensor.

3.2.2 Joint Distance and Joint Orientation

Joint distance (JD) and joint orientation (JO) are the geometric features of gait and are extracted from a single human skeletal model. Both of these features are derived from joint coordinate information following the same methodology introduced in [15]. In particular, we calculate the Euclidean distance between two arbitrary joints in three planes corresponding to $x = 0$, $y = 0$, and $z = 0$ respectively. Specifically, the distance values of two joints i and j are calculated as follows:

$$\begin{aligned}\sigma_x(i,j) &= \sqrt{(y_j - y_i)^2 + (z_j - z_i)^2}, \\ \sigma_y(i,j) &= \sqrt{(x_j - x_i)^2 + (z_j - z_i)^2}, \\ \sigma_z(i,j) &= \sqrt{(y_j - y_i)^2 + (x_j - x_i)^2}.\end{aligned} \tag{1}$$

Hence, the joint distance feature between two arbitrary joints is defined as $s = [\sigma_x \; \sigma_y \; \sigma_z]$.

Furthermore, we calculate the joint orientation as an angle between the joint-

joint vector and three coordinate axes, $\overrightarrow{Ox}$, $\overrightarrow{Oy}$, and $\overrightarrow{Oz}$. Precisely, angles between joint-joint vector $\overrightarrow{ji}$ and three coordinate axes are defined as follows:

$$\begin{aligned} \tau_x(\overrightarrow{ji}, \overrightarrow{Oy}) &= \cos^{-1}(\frac{\overrightarrow{ji} \cdot \overrightarrow{Oy}}{||\overrightarrow{ji})|| \times ||\overrightarrow{Oy})||}), \\ \tau_y(\overrightarrow{ji}, \overrightarrow{Oz}) &= \cos^{-1}(\frac{\overrightarrow{ji} \cdot \overrightarrow{Oz}}{||\overrightarrow{ji})|| \times ||\overrightarrow{Oz})||}), \\ \tau_z(\overrightarrow{ji}, \overrightarrow{Ox}) &= \cos^{-1}(\frac{\overrightarrow{ji} \cdot \overrightarrow{Ox}}{||\overrightarrow{ji})|| \times ||\overrightarrow{Ox})||}). \end{aligned} \tag{2}$$

Hence, the joint orientation feature between two arbitrary joints is defined as $t = [\tau_x \ \tau_y \ \tau_z]$.

3.2.3 Joint Velocity

Joint velocity (JV) measures displacement of joints between consecutive skeleton frames. In contrast to the geometric features, which extract joint information from a single skeleton frame, the joint velocity feature involves multiple skeleton frames, thereby capturing the dynamics of the skeletal joints throughout the continuous frame sequence. Concretely, for a single joint i, the joint velocity at time t is defined as follows:

$$\begin{aligned} \upsilon_x(t) &= |S_x^t(i) - S_x^{t-1}(i)|, \\ \upsilon_y(t) &= |S_y^t(i) - S_y^{t-1}(i)|, \\ \upsilon_z(t) &= |S_z^t(i) - S_z^{t-1}(i)|, \end{aligned} \tag{5}$$

where $S_x^t(i), S_y^t(i), S_z^t(i)$ are the $3D$ coordinates of a joint i of skeleton S at time t. The joint velocity feature for a single joint i at time t (i.e. t^{th} frame of a skeleton sequence) is defined as $v(t) = [\upsilon_x(t) \ \upsilon_y(t) \ \upsilon_z(t)]$.

With the n-joint skeleton S at time t, we retrieve n joint velocity features. Furthermore, we construct $n(n-1)/2$ joint distance features corresponding to every pair of joints in a skeleton. The number of orientation features is also $n(n-1)/2$ due to

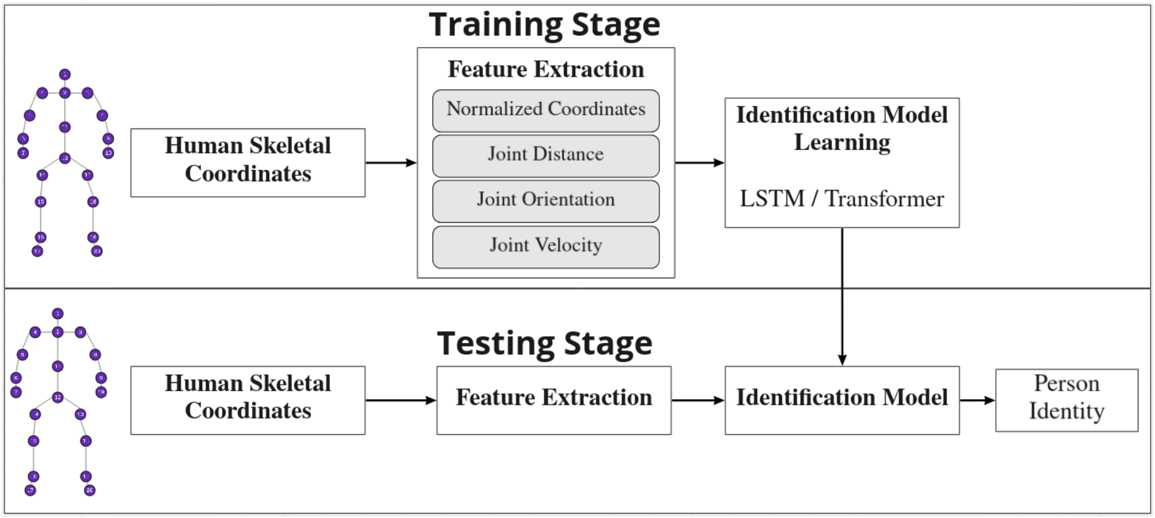


Figure 3-2: Overview of the gait-based person identification method.

the similar definition. All the derived features will be flattened prior to being fed into the identity learning models.

3.3 Person identification model learning

In this section, we present the detailed design of the proposed person identity learning models based on LSTM and Transformer networks, which learns underlying relationship between the extracted gait features and the person identity in a sequential manner.

3.3.1 Long Short Term Memory

Recurrent neural network (RNN) is a type of neural network architecture that is primarily used for modelling arbitrary length sequential data. Given a sequence $X = (x_0, x_1, ..., x_T)$ consisting of T input features, for a timestamp t, the $n-th$ RNN unit produces an output based on both current input feature x_t and state h_{t-1} of the previous unit. Formally, given an input feature x_t and previous RNN unit's state

h_{t-1}, an RNN unit is usually implemented as follows:

$$\begin{aligned} h_t &= g_1(W_{hh}h_{t-1} + W_{hx}x_t + b_h) \\ y_t &= g_2(W_{yh}h_t + b_y) \end{aligned} \tag{6}$$

where W_{hh}, W_{hx}, and W_{yh} are the unit's internal parameters, g_1 and g_2 are activation functions, e.g. tanh or sigmoid, that are used to squash the unit's output values in a range $[0, 1]$, b_h and b_y are the biases, h_t is a state of the unit and y_t is an output of the unit. For some tasks, where a single label is predicted only the output of the last unit is used.

Despite RNNs were used in many natural language processing [29] and computer vision tasks [22], there remain difficulties with learning long-term dependencies within a long sequence of data, mainly because of the vanishing or exploding gradients problem [26]. Long short-term memory (LSTM) [13] is a specific kind of Recurrent Neural Networks (RNNs), which is designed to solve the above-mentioned problems by learning what information should be preserved and what can be forgotten. LSTM achieves this thanks to an extended architecture of a unit, which consists of several steps of computations that update the internal unit's state and output. Formally, given an input feature x_t and previous LSTM unit's state C_{t-1}, and output h_{t-1}, the current unit state and output are calculated according to the following recurrent equation:

$$\begin{aligned} f_t &= \sigma(W_{fh}h_{t-1} + W_{fx}x_t + b_f) \\ i_t &= \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i) \\ C_t^{'} &= \tanh(W_{ch}h_{t-1} + W_{cx}x_t + b_c) \\ C_t &= f_t C_{t-1} + i_t C_t^{'} \\ o_t &= \sigma(W_{oh}h_{t-1} + W_{ox}x_t + b_o) \\ h_t &= o_t \tanh(C_t) \end{aligned} \tag{7}$$

where C_* are the units' state information passed into the next units and h_* are the units' outputs. f_t, i_t, and o_t are the internal processing units, that learn the parameters required for filtering out the input and output information. W_* are the units' internal parameters, σ (sigmoid) and tanh are activation functions that are used to squash the unit's output values in a range $[0, 1]$, and b_* are the biases.

By its nature, gait captures human pose information in space and its motion in time. This information can be represented as a frame sequence where each frame contains spatial information of a human pose. Multiple frames contain the temporal information that captures dynamics of a group of skeletal joints and of a human body in general. Processing the human pose information sequentially, we can effectively learn the spatial correlations within a single frame and temporal relations throughout multiple frames. Based on this intuition, we propose a person identity learning model based on LSTM architecture. We train the LSTM model on a set of features selected from the features described in section 3.2. Specifically, with sequence size S and K-joint 3D skeletal coordinates, the dimensionalities of different gait features are given as below:

- Normalized skeletal joint coordinates, where each of the S input units of the LSTM accepts $K \times 3$ sized vectors.

- Joint distance features, where each of the S input units of the LSTM accepts $0.5 \times K \times (K - 1) \times 3$ sized vectors.

- Joint orientation features, where each of the S input units of the LSTM accepts $0.5 \times K \times (K - 1) \times 3$ sized vectors.

- Joint velocity features, where each of the S input units of the LSTM accepts $K \times 3$ sized vectors.

3.3.2 Transformer

Transformer is a special type of neural network architecture based on attention mechanism [1]. In our implementation of the attention-based person identification model, we follow the original Transformer [33] design with some modifications that will be listed below. The overview of the model for gait sequence classification is displayed in Figure 3-3. Self-attention (SA) is the fundamental operation of the transformer architecture, which is designed to be able to learn the internal information within a given sequence of data. As the network processes each item within the input sequence, self-attention examines items in other positions in the input sequence for information that can help to come up with a better encoding for this item. Self-attention is a weighed sum over all values $\mathbf{v}$ computed for each element in an input sequence $\mathbf{s} \in R^{N \times D}$. The attention weights A_{ij} are based on a compatibility function of the query q_i with the corresponding key k_j. Formally, the self-attention is computed according to the following equations:

$$\begin{aligned} [\mathbf{q}, \mathbf{k}, \mathbf{v}] &= \mathbf{s}U, \\ U &\in R^{D \times 3D_h} \\ A &= Softmax(qk^T/\sqrt{D_h}) \\ SA(\mathbf{s}) &= Av \end{aligned} \tag{8}$$

Multi-head attention is an extension of a self-attention, where multiple self-attention operations a.k.a heads are computed.

$$\begin{aligned} MSA(\mathbf{s}) = [SA_1(\mathbf{s}), SA_2(\mathbf{s}), ...SA_k(\mathbf{s})]V, \\ V \in R^{kD_h \times D} \end{aligned} \tag{9}$$

We adopt the multi-head attention and general encoder design as depicted in Figure 3-3. We use 4 heads of attention inside an encoder and use a single encoder. The output of the encoder is passed to the classifier with a single dense layer and softmax outputting the predicted label.

The gait features are extracted from a video sequence and preserve their sequential nature. The order information, namely the position of the gait feature within an underlying sequence should also be preserved while being processed by Transformer network. Positional Encoder layer placed before the Multi-Head Attention serves for

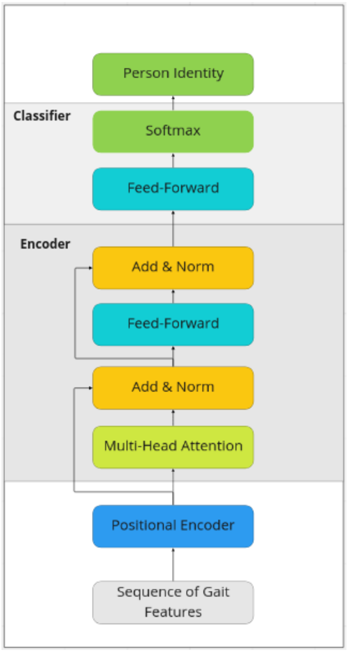


Figure 3-3: Person identity classifier with Transformer Encoder

this purpose, adding the position information to the input gait features.

Furthermore, we will train on several different combinations of the gait features by concatenating them to a single vector. The detailed comparison of the performance for each combination of features is presented in Subsection 4.3.2.

3.4 Data augmentation for person identification

In order to increase the number of gait sequences for a single person, we have employed two types of data augmentation techniques, flipping and rotation. First, knowing that the human body is vertically symmetrical, we can double the size of the gait sequences, by flipping the images with gait frames vertically. Second, having a single skeleton with 3D body joint coordinates, we can generate a set of similar synthetic skeletons by slightly rotating the joints. While rotating the 3D skeleton joints, we have adopted the methodology introduced in [35] and the reader is referred to this work for technical details.

Moreover, in order to increase the number of sequences per person, we follow the windowed sequence extraction method employed in [15], where a skeleton sequence is partitioned into multiple sequences with an overlapping rate of 80%. The reader is referred to [15] for details regarding this process.

Chapter 4

Experimental results and discussion

In this section, we evaluate the proposed 3D gait-based person identification methods on UPCV Gait [17–19], KS20 VisLab Multi-View Kinect Skeleton [23, 25] datasets.

4.1 Experimental Setup

In this section, we conduct the following experiments:

- Compare the rank-1 identification accuracy of the proposed methods with several baselines and previous approaches on two datasets.
- Analyze performance of the proposed method under different combinations of gait features.
- Evaluate the proposed networks' complexity in competition with several modern existing models in the field of gait-based person identification.

4.2 Datasets

UPCV Gait: The dataset is a benchmark dataset for pose based gender and identity recognition. It consists of human pose sequences for 30 persons walking in a direct path with a normal speed. The sequences were extracted through Microsoft Kinect v1 from a side view and consists of 55 to 120 frames depending on the walking speed.

There are 5 sequences for each person, totalling 150 sequences. Three sequences of each person are randomly selected for training while the remaining sequences left for testing, following the same evaluation protocol used in [15].

KS20 VisLab Multi-View Kinect Skeleton: The dataset is a collection of pose sequences of 20 persons captured using Microsoft Kinect v2 from five viewpoints, including left lateral at 0°, left diagonal at 30°, frontal at 90°, right diagonal at 130°, and right lateral at 180°. In total, it contains 300 pose sequences, with 3 sessions per person at particular viewpoint. Two sequences per person in a particular viewpoint are randomly selected for training while the remaining sequences left for testing, following the same evaluation protocol used in [15].

4.3 Results and Discussions

In this chapter, the numerical results are presented along with the discussion of advantages and drawbacks of the proposed methods.

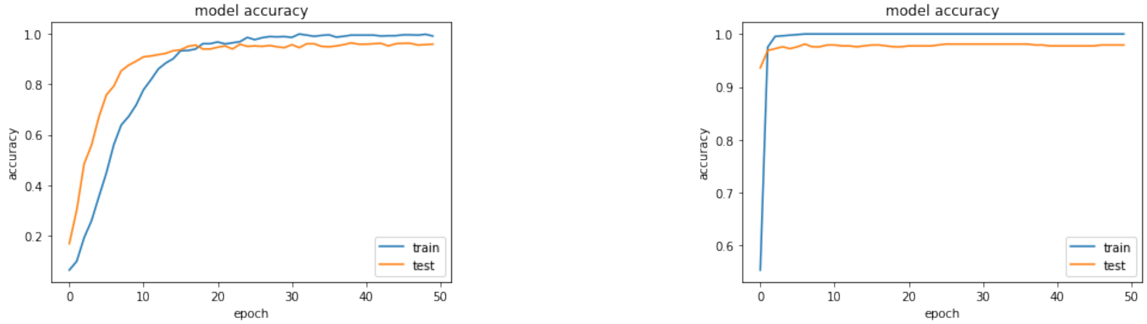
4.3.1 Method comparison

We evaluate the performance of the proposed person identification methods using a rank-1 accuracy rate and in terms of processing time and model complexity. We compare the proposed methods against the baseline and previous methods for 3D skeletal gait recognition. In performance comparison, we include a list of baseline methods like k-Nearest Neighbors (kNN), Support Vector Machines (SVM), and Random Forest (RF), which are also trained and evaluated on the gait features extracted from the datasets mentioned above. We also compare with several deep learning-based methods such as Covariance Dissimilarity [20], SRC in Dissimilarity Space [32], Euclidean-Riemannian Fusing [18], Context-Unaware Score-level Fusion [23], and Context-Aware Score-level Fusion [23]. We also included the ST-CNN model [15], which is trained on descriptive statistics of joint distance and joint orientation features. The detailed quantitative results of the experiments are demonstrated in Table 4.1.

We report the results of the proposed LSTM and Transformer models with several different combinations of the gait features i.e. joint coordinates (JC), joint distance (JD), joint orientation (JO), and joint velocity (JV). The methods were trained on the augmented dataset, where the windowing technique is used as described in Section 3.4. We have additionally applied the rotation and flipping techniques to the training set for the methods employing the combination of gait features, including JC, JV, JD, and JO, labeled **(rot. & flip.)** in Table 4.1.

Baseline models, including kNN, SVM, and RF were trained and evaluated to test the validity and effectiveness of the gait features. While kNN achieves an accuracy rate of above 60% for the UPCV1 Gait dataset, SVM and RF methods show superior performance at 94.24% and 94.56% respectively. SVM is an effective baseline model because of its ability to generalize well under a limited number of training samples and high dimensions. RF is composed of multiple random decision trees that help to efficiently train a generalizable classification model. However, the baseline models fail to capture temporal dependencies and dynamics of sequential data simply by their design. They will not preserve their performance for higher-dimensional features and are not robust for intra-class variances introduced by multiple viewpoints as in KS20.

To address these issues, we have designed and modelled the LSTM and Transformer methods, the effectiveness of which is proved to be superior to conventional machine learning methods. Specifically, the LSTM method evaluated with different feature combinations on the UPCV Gait dataset has achieved an accuracy rate of over 80%, except for the joint velocity (JV) feature, showing its highest rate with the joint orientation feature at 92.56%. Similar metrics for the KS20 VisLab dataset did not extend beyond 85% due to the fact that the dataset contains gait sequences from multiple views, thereby introducing intra-class variations. Combination of three best performing features, JC, JD, JO, further improves the accuracy of LSTM up to 95.50% and 72.17% on UPCV Gait and KS20 VisLab respectively. However, the performance of the LSTM remains inferior to the previous methods, achieving around 4% less accuracy rate on UPCV Gait and KS20 datasets. Only the join distance feature shows a high accuracy rate which is close to the state-of-the-art performance by Context-Aware Score-level Fusion [23] (88.67%). Joint coordinates, joint velocity,



(a) Learning curve of Transformer with normalized joint coordinates from UPCV Gait dataset i.e JC + Transformer

(b) Learning curve of Transformer with the combination of gait features from UPCV Gait dataset i.e JC, JD, and JO + Transformer

Figure 4-1: Learning curves of the Transformer model with different features from UPCV Gait dataset

and joint orientation must be thoroughly evaluated together with the LSTM model in order to determine what stages of training LSTM pose limitations in learning discriminative characteristics of the human gait. In particular, one can evaluate the model with gait sequences with more frames, in order to capture dynamics at a longer time range.

The Transformer achieves much higher accuracy rates compared to the LSTM, showing up to 98.26% for UPCV Gait and 90.86% for KS20 VisLab datasets. This supports our hypothesis that the self-attention layers help better capture the temporal dependencies within a sequence of gait features. Moreover, Transformer method in combination with JC, JD, and JO features is superior to other previous methods evaluated in KS20 VisLab dataset, outperforming previous methods by around 3.0%. However, on UPCV Gait dataset, Transformer's highest accuracy rate, 98.26%, is around 0.6% less than that of ST-CNN.

4.3.2 Performance sensitivity

We also have investigated the contribution of every feature type separately with the Transformer model. Specifically, we have trained and tested the model separately with normalized joint coordinates (JC), joint distance (JD), joint orientation (JV),

Table 4.1: Method and Feature Comparison on UPCV Gait and KS20 VisLab Multi-View Kinect Skeleton Datasets.

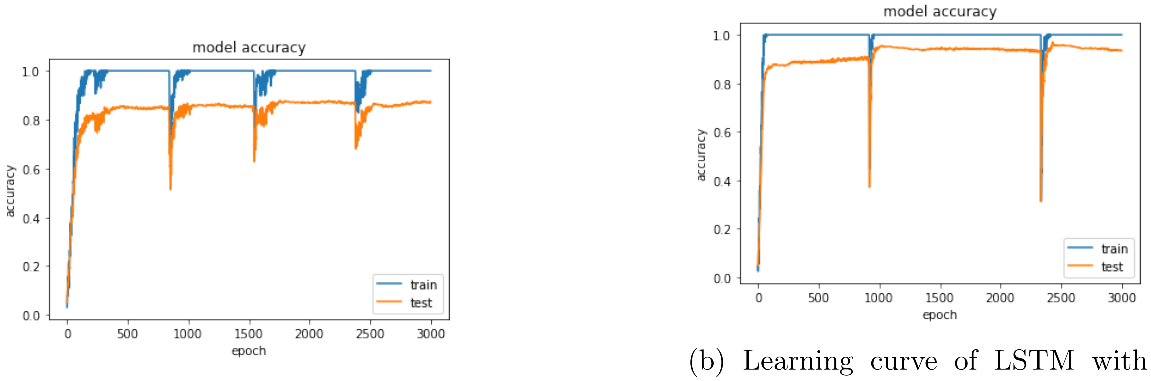
Method	Accuracy (%) UPCV Gait	Accuracy (%) KS20 VisLab
JC, JV, JD, and JO + kNN	61.14%	38.56%
JC, JV, JD, and JO + SVM	94.24%	88.37%
JC, JV, JD, and JO + RF	94.56%	88.21%
Covariance Dissimilarity [20]	89.60%	NA
SRC in Dissimilarity Space [32]	94.50%	NA
Euclidean-Riemannian Fusing [18]	95.67%	NA
Context-Unaware Score-level Fusion [23]	NA	79.33%
Context-Aware Score-level Fusion [23]	NA	88.67%
ST-CNN [15]	98.86%	87.63%
JC + LSTM	84.42%	67.60%
JV + LSTM	67.99%	34.34%
JD + LSTM	89.79%	84.34%
JO + LSTM	92.56%	69.13%
JC, JD, JO + PCA(512) + LSTM	93.24%	69.13%
JC, JV, JD, JO + PCA(512) + LSTM	92.56%	70.34%
(rot. & flip.) JC, JD, JO + PCA(512) + LSTM	95.50%	71.95%
(rot. & flip.) JC, JV, JD, JO + PCA(512) + LSTM	94.46%	72.17%
JC + Transformer	96.71%	86.73%
JV + Transformer	75.45%	25.86%
JD + Transformer	97.40%	90.43%
JO + Transformer	97.23%	88.26%
JC, JD, JO + Transformer	97.40%	87.17%
JC, JV, JD, JO + Transformer	96.42%	89.52%
JC, JD, JO + PCA(512) + Transformer	97.40%	90.65%
JC, JV, JD, JO + PCA(512) + Transformer	96.23%	89.52%
(rot. & flip.) JC, JD, JO + PCA(512) + Transformer	**98.26%**	**90.86%**
(rot. & flip.) JC, JV, JD, JO + PCA(512) + Transformer	97.23%	90.65%

and joint velocity (JV) features. The models trained on the joint distance and orientation features show the highest predictive accuracy on both of the datasets compared to joint coordinates and joint velocity features. The reason for this might be that the joint distance and orientation features better captures the mutual dynamics of multiple pairs of joints throughout the gait sequence. The detailed results of the experiments are shown in Table 4.1. While the Transformer model with joint coordinates requires around 50 epochs to achieve its highest accuracy, the same method with the the combination of the gait features requires around 8 epochs to achieve the same performance. This is mainly because joint distance and joint orientation serve as a well-constructed and intuitive features of human gait. The learning curves are displayed in Figure 4-1.

After augmenting the training data by flipping the joint coordinates horizontally and applying slight rotation in 3D as described in Section 3.4, we have increased the training set size by 56 times. Then we have trained and evaluated the LSTM and Transformer methods on the combinations of gait features (JC, JV, JD, JO). The accuracy of the methods has increased by around 1%, as shown in the rows labelled (rot. & flip.) in Table 4.1. The improved methods are still inferior to ST-SNN [15] on UPCV Gait dataset, Transformer showing 98.26% accuracy against 98.86% of ST-CNN. However, on KS20 VisLab Transformer outperforms the previous methods, advancing up to 90.86% against 88.67% of Context-Aware Score-level Fusion [23].

4.3.3 Complexity analysis

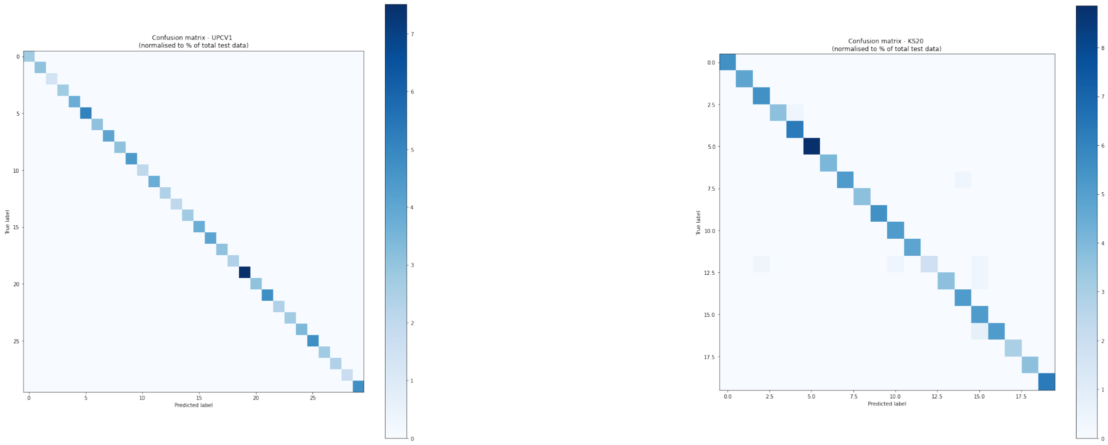
In the last experiment, we benchmark the network complexity by measuring the training time based on a system equipped with NVIDIA GeForce GTX 1650. The training time per epoch on UPCV Gait dataset of the Transformer model with different gait features is plotted on Fig 4-4. All the methods were trained under the same configuration for 50 epochs. Although, some features take more time to train because of their complexity (number of dimensions, etc), all the features except JC show the accuracy rate higher than 96% as shown in 4.1. For instance, by concatenating the joint distance, joint orientation, and joint coordinates features, we trained the Trans-



(a) Learning curve of LSTM with normalized joint coordinates from UPCV Gait dataset i.e. JC + LSTM

(b) Learning curve of LSTM with the combination of gait features from UPCV Gait dataset i.e JC, JD, and JO + LSTM

Figure 4-2: Learning curves of the LSTM method with different feature combinations from UPCV Gait dataset



(a) UPCV Gait

(b) KS20 VisLab Multi-View

Figure 4-3: Confusion matrices of person identification results by the Transformer models on UPCV Gait and KS20 VisLab Multi-View Kinect Skeleton datasets

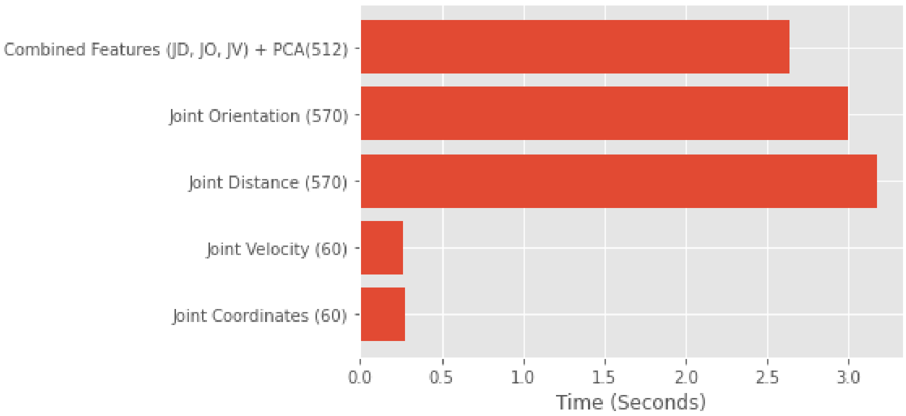


Figure 4-4: Training time per epoch. Transformer with different features from UPCV Gait dataset

former model for 50 epochs and obtained above 97% accuracy, so did we by using only the joint orientation feature in almost three times less time. When multiple features are combined by concatenating, the size of a single feature vector increases appropriately. For instance, for UPCV Gait dataset, the size of the feature vector will be 1260 after concatenating JC, JV, JO, JD, whose sizes are 60, 60, 570, and 570 respectively. In order to keep the complexity of the model small we have applied Principal Component Analysis (PCA) to reduce the dimensionality of the combined feature vectors down to 512, to be comparable to the dimensionality of JO and JD features. More importantly, this trick had no negative effect on the model's performance and preserved the accuracy rates almost identical to the ones achieved with original dimensions.

In contrast to Transformer model, LSTM model requires much more time to train and converge to show a stable performance. For instance, as depicted in Figure 4-1a and Figure 4-2a, LSTM reached the accuracy rate over 95% after 300 epochs, while Transformer achieved the same performance after 50 epochs.

Using a combination of joint coordinates, distance and orientation features with the Transformer model we achieved a leading performance in a few minutes of training. However, the UPCV Gait and KS20 Multi-View datasets contain very few number of subjects (20 and 30), and in order to investigate further the complexity and capabilities of the LSTM and Transformer model, we should evaluate the methods on more challenging datasets with skeletal joints. For instances. those with more subjects and those collected under unconstrained environments.

4.3.4 Model deployment

The proposed models have several limitations. First, the human skeleton-based gait recognition models require the availability of algorithms that would provide the joint coordinates data. Second, the current versions of the models were trained on the datasets that were collected in a constrained environment, where the human gait sequences were recorded from a single or a limited number views in a relatively short distances. In real-world systems, depending on the distance and the viewpoint of the cameras, the scale and quality of the collected data will vary hugely, which may reduce the models' effectiveness unless the model was trained on similar gait datasets. Moreover, it should be noted that in an unconstrained environment, cameras may capture several persons in a single session, so preliminary steps for partitioning different skeletal data may also be required. We plan to evaluate the proposed sequential models on more challenging datasets. For instance, skeletal gait datasets with a large population, recorded from multiple views and collected under unconstrained environments. In absence of the 3D depth cameras, we can collect skeletal data using real-time pose estimation algorithms as Open Pose or Alpha Pose, which could be more affordable and equally effective. Moreover, we can place several RGB cameras to capture subjects from multiple viewpoints, which in turn helps to improve the detection and further recognition accuracy.

Chapter 5

Conclusion

In this paper, we have investigated several gait features derived from 3D body joint coordinates. Using these gait features, we then trained a compact deep sequential models based on LSTM and Transformer networks. The proposed models fully gain the spatial correlation of body joints within a single frame and the temporal dynamics of joints from frame to frame. We have also introduced data augmentation steps, which enriched the dataset when working with a limited number of gait sequences for an individual. According to the experimental results, the proposed models achieved high person identification results on the UPCV Gait (98.26%) and KS20 VisLab Multi-View (90.86%) datasets. Our further work will focus on testing our models on more challenging datasets and involve more advanced gait features while keeping the model compact enough to stay efficient.

Bibliography

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Ning Cai, Shiling Feng, Qing Gui, Lei Zhao, Huadong Pan, Jun Yin, and Bin Lin. Hybrid silhouette-skeleton body representation for gait recognition. In *2021 13th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pages 216–220. IEEE, 2021.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[4] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8126–8133, 2019.

[5] Changhong Chen, Jimin Liang, Heng Zhao, Haihong Hu, and Jie Tian. Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, 30(11):977–984, 2009.

[6] Shiliang Chen, Wentao He, Jianfeng Ren, and Xudong Jiang. Attention-based dual-stream vision transformer for radar gait recognition. *arXiv preprint arXiv:2111.12290*, 2021.

[7] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based

model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020.

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[11] Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018.

[12] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2005.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Haohua Huang, Pan Zhou, Ye Li, and Fangmin Sun. A lightweight attention-based cnn model for efficient gait recognition with wearable imu sensors. *Sensors*, 21(8):2866, 2021.

[15] Thien Huynh-The, Cam-Hao Hua, Nguyen Anh Tu, and Dong-Seong Kim. Learning 3d spatiotemporal gait feature by convolutional network for person identification. *Neurocomputing*, 397:192–202, 2020.

[16] Pengtao Jia, Qi Zhao, Boze Li, and Jing Zhang. Cjam: Convolutional neural network joint attention mechanism in gait recognition. *IEICE TRANSACTIONS on Information and Systems*, 104(8):1239–1249, 2021.

[17] Dimitrios Kastaniotis, Ilias Theodorakopoulos, George Economou, and Spiros Fotopoulos. Gait-based gender recognition using pose information for real time applications. In *2013 18th International Conference on Digital Signal Processing (DSP)*, pages 1–6. IEEE, 2013.

[18] Dimitris Kastaniotis, Ilias Theodorakopoulos, George Economou, and Spiros Fotopoulos. Gait based recognition via fusing information from euclidean and riemannian manifolds. *Pattern Recognition Letters*, 84:245–251, 2016.

[19] Dimitris Kastaniotis, Ilias Theodorakopoulos, Christos Theoharatos, George Economou, and Spiros Fotopoulos. A framework for gait-based recognition using kinect. *Pattern Recognition Letters*, 68:327–335, 2015.

[20] MS Naresh Kumar and R Venkatesh Babu. Human gait recognition using depth camera: a covariance based approach. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–6, 2012.

[21] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, Puneet Agarwal, et al. Long short term memory networks for anomaly detection in time series. In *Proceedings*, volume 89, pages 89–94, 2015.

[22] Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu. Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3639–3655, 2017.

[23] Athira Nambiar, Alexandre Bernardino, Jacinto C Nascimento, and Ana Fred. Context-aware person re-identification in the wild via fusion of gait and anthropometric features. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 973–980. IEEE, 2017.

[24] Athira Nambiar, Jacinto C Nascimento, Alexandre Bernardino, and José Santos-Victor. Person re-identification in frontal gait sequences via histogram of optic flow energy image. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 250–262. Springer, 2016.

[25] Athira M Nambiar, Alexandre Bernardino, Jacinto C Nascimento, and Ana LN Fred. Towards view-point invariant person re-identification via fusion of anthropometric and gait features from kinect measurements. In *VISIGRAPP (5: VISAPP)*, pages 108–119, 2017.

[26] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[27] Aditi Roy, Shamik Sural, and Jayanta Mukherjee. Gait recognition using pose kinematics and pose energy image. *Signal Processing*, 92(3):780–792, 2012.

[28] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[29] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *ICML*, 2011.

[30] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[31] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: graph convolutional network for skeleton-based gait recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2314–2318. IEEE, 2021.

[32] Ilias Theodorakopoulos, Dimitris Kastaniotis, George Economou, and Spiros Fotopoulos. Pose-based human action recognition via sparse representation in dissimilarity space. *Journal of Visual Communication and Image Representation*, 25(1):12–23, 2014.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[35] Pichao Wang, Wanqing Li, Chuankun Li, and Yonghong Hou. Action recognition based on joint trajectory maps with convolutional neural networks. *Knowledge-Based Systems*, 158:43–53, 2018.

[36] Xiuhui Wang and Wei Qi Yan. Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. *International journal of neural systems*, 30(01):1950027, 2020.

[37] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.