View Invariant Gait Recognition For Person Re-Identification in a Multi Surveillance Camera Environment

By

Md Mahedi Hasan

1014312019

MASTER OF SCIENCE

IN INFORMATION AND COMMUNICATIN TECHNOLOGY

Institute of Information and Communication Technology

Bangladesh University of Engineering and Technology

Dhaka, Bangladesh

January 2019

CANDIDATE'S DECLARATION

This is to certify that the work presented in this thesis, titled, "View Invariant Gait Recognition For Person Re-Identification in a Multi Surveillance Camera Environment", is the outcome of the investigation and research carried out by us under the supervision of Name of the SupervisorN-2.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Md Mahedi Hasan
1014312019

CERTIFICATION

This thesis titled, **"View Invariant Gait Recognition For Person Re-Identification in a Multi Surveillance Camera Environment"**, submitted by Md Mahedi Hasan, Roll No.: 1014312019, Session: 2019, has been accepted as satisfactory in partial fulfillment of the requirement for the degree of MASTER OF SCIENCE in Information and Communication Technology in January 2019.

BOARD OF EXAMINERS

Signature

Name of the Supervisor chairman
Designation
Address

Signature

N-2 Member
D-2
A-2

ACKNOWLEDGEMENT

First and foremost, I express my deepest gratitude to **Almighty Allah** for bestowing His blessings on me and enabled me to successfully accomplish this work.

I would like to express my sincere appreciation and gratitude to my supervisor **Dr. Hossen Asiful Mustafa**, who is a great mentor leading me to the research field of computer vision and deep learning. His cooperation and guidance has been a constant source of encouragement in shaping this work throughout my master thesis period. Without his great help and invaluable suggestions I would hardly finish this thesis.

I will always appreciate my classmates and brothers **Md Abdul Aowal, Abu Noman**, and **Imran Khan** for their firm-backing and support that has greatly helped me in accomplishing the research work. I am also highly grateful for the support of my roommate cum brother **Abdullah Al Mahmud**, who always tolerated my frustration. We studied together and shared a lot of discussion which were very helpful for this research.

Finally, I want to dedicate the essence of my purest respect to my parents and to my colleagues for providing me with support and continuous inspiration throughout my years of study and through the process of writing this thesis. This accomplishment would not have been possible without them. Thank you.

Dhaka
January 2019

Md Mahedi Hasan

Contents

List of Figures

List of Tables

List of Algorithms

ABSTRACT

Recognizing individual people from gait is still a challenging problem in computer vision research due to the presence of various covariate factors like varying view angle, change in clothing, walking speed, and load carriage, etc. Most of the earlier works were based on human silhouettes which have proven to be efficient in recognition but are not invariant to change in illumination and clothing. In this research, to address this problem, we present a simple yet effective approach for robust gait recognition using a recurrent neural network (RNN). Our RNN network with GRU architecture is very powerful in capturing the temporal dynamics of the human body pose sequence and perform recognition. We also design a low-dimensional gait feature descriptor based on the 2D coordinates of human pose information which is proven to be not only invariant to various covariate factors but also effective in representing the dynamics of various gait pattern. The experimental results on challenging CASIA A and CASIA B gait datasets demonstrate that the proposed method has achieved state-of-the-art performance on both single-view and cross-view gait recognition which prove the effectiveness of our method.

Chapter 1

Introduction

Biometrics refers to automatic identification or authentication of people by analyzing their physiological and behavioral characteristics. Physiological biometrics is related to the shape of body parts such as face, fingerprints, shape of the hand, iris, retina, etc., which are not subject to change due to aging. It is now used as the most stable means for authenticating and identifying people in a reliable way. However, for efficient and accurate authentication, these traits require cooperation from the subject along with a comprehensive controlled environmental setup. Hence, these traits are not useful in surveillance systems. Behavioral biometrics such as signatures, gestures, gait, and voice, etc., is related to a persons behavior. But, these traits are more prone to change depending on factors such as aging, injuries, or even mood.

Gait recognition is a behavioral biometric modality that identifies a person based on walking posture of that person. In contrast to other biometrics such as face and fingerprint, it is a non-invasive technique for identifying an individual which is hard to copy. A unique advantage of gait as a biometric is that it offers recognition at a distance and at low-resolution images; consequently, gait biometric signature is now considered as the only likely identification technology suitable for access control, covert video surveillance, criminal investigation, and forensic analysis and the method is not vulnerable to spoofing attacks and signature forgery.

Due to the advantages of gait recognition, the past two decades have witnessed significant improvements of its algorithms. However, unfortunately, there still exist many challenges that need to be addressed for robust gait recognition. It has been observed that the performance of gait recognition is highly affected by different intraclass variations in people's appearance such as clothing and carrying variation, and in environment such as variations in illumination, walking surface, and view angle, etc. These factors can drastically reduce the performance of gait recognition. Traditional body appearance-based methods in gait recognition are sensitive to these covariate factors since the extraction of human silhouettes is affected by by the changes in lighting. Moreover, when the shape of the human body and appearance change substantially, the performance of appearance-based methods severely degrades. Therefore, these methods are not completely robust toward these covariate change.

In contrast, model-based gait recognition exploits features based on the shape of human body parts and the dynamics of the motion of each of these parts. The salient advantage of the model-based approach is that, as opposed to silhouette-based approaches, it can efficiently handle many covariate changes such as view angle, body appearance and shape, so, these methods show robustness toward these variations. But, due to heavy computational cost required for modeling human body accurately, these methods were not as popular as appearance-based ones.

Modern deep learning-based algorithms have recently gained increasing popularity while achieving outstanding performance in many computer vision tasks like person re-identification, pose estimation, and action recognition, etc. Furthermore, advancement on human body pose estimation can significantly assist in accurately modeling different body parts required for model-based gait recognition. On the other hand, recurrent neural networks have also achieved a promising performance in many sequence labeling tasks. The reason behind their effectiveness for sequence-based tasks lies in their ability to capture long-range dependencies in a temporal context from sequence. RNNs have been successfully employed to achieve state-of-the-art results in many vision-based tasks like human emotion detection and action recognition.

In this work, we propose a model-based gait recognition method where we consider human 2D pose data as our effective gait features. As body pose is proven not to depend on people body appearance and shape, and is invariant to change of clothing and carrying conditions. Additionally, as gait can be considered a time series of walking postures, body pose information has a powerful capacity to capture the temporal pattern of gait. Therefore, the proposed method will be less affected by the variation of covariate factors. It is also worth mentioning that, in this work we didn't use 3D pose data as our gait feature: firstly, computing 3D poses is computationally expensive, and secondly, most of the 3D pose estimation algorithms from 2D RGB images often require multiple views, and hence multiple cameras, rendering the technique unsuitable for surveillance. Again, recovering 3D pose from a single RGB images is an ill-posed problem and often causes large pose estimation errors.

Compared to other gait covariate, view is the most important factor severely affecting gait recognition performance. To handle view variation efficiently, gait algorithms have generally been studied under three experimental setups: single-view, multi-view, and cross-view setup. In single-view gait recognition, both probe and gallery gaits are kept within same view angle, where in cross-view gait recognition, the probe and gallery gaits are kept in different views; and in multi-view gait recognition, multiple views of gallery gaits are combined to recognize a probe gait under a specific view.

Thus, the key to our proposed method is to develop a pose-based deep recurrent neural network for robust gait recognition by modeling the temporal dynamics associated with human gait. Most of the descriptors proposed in the literature for gait recognition often lead to a high dimensional feature space, which can be computationally expensive to map. In this work, we designed a lower dimensional spatio-temporal feature descriptor from 2D pose estimation for improved performance at a reduced computational cost. Our gait descriptor is a concatenation of four different kinds of features which are robust to view variation. We demonstrate the effectiveness of our proposed method through extensive experiments on two public benchmark datasets: the CASIA A and CASIA B gait dataset [6]. Our method achieved state-of-the-art performance on these two challenging gait datasets in both single-view and cross-view recognition, providing better results as compared to other methods proposed in the literature.

For multi-view gait recognition, we also propose a two-stage network in which we first determine the walking direction, i.e. the viewpoint angle of the camera using a 3D convolutional network and later identify the subject using proposed RNN based temporal network trained on that particular angle. Our proposed two-staged network is far simpler and efficient in terms of time and space while outperforming present state-of-the-art networks on multi-view gait recognition.

The main contributions of our work are summarized as follows:

- We propose a novel RNN network with GRU architecture and devise several strategies to effectively train the network for robust gait recognition.
- The proposed pose-based RNN network achieves the best results on two challenging benchmark datasets CASIA A and CASIA B by outperforming other prevailing methods in single-view gait recognition at a significant margin.
- We consider 2D coordinates of body pose to design a novel gait feature descriptor which is invariant to covariate factors and achieved comparable performance to the methods which require to calculate gait energy image (GEI) or expensive 3D poses for gait descriptors.

1.1 What is Gait Recognition

1.2 Importance of Gait Recognition

1.3 Challenges Gait Recognition

Chapter 2

Literature Review

Over the last two decades, several methods have been studied to develop a robust gait recognition system [7]. However, robust recognition is still challenging due to the presence of large intraclass variations in a person's gait which substantially changed the performance. In this section, we briefly discuss the literature of the two categories of existing gait recognition techniques: appearance-based and model-based methods. Next, we review some of the recent deep learning-based gait recognition approaches which are closely related to our work.

2.1 Appearance-based Methods

Most of the previous work following this approach [8–10] used human silhouette masks as the main source of information and extracted features that show how these mask change. The most popular gait representation employed in such work is gait energy image (GEI) [8], a binary mask computed through aligning and averaging the silhouettes over the complete gait cycle. Though there are many other alternatives for GEI, e.g., gait entropy image (GEnI) [9], and gait flow image (GFI) [10], due to its in-sensitiveness of incidental silhouettes error, it has been considered as the most stable gait features. It can achieve good performance under controlled and cooperative environments, but does not show robustness when the view angle and clothing condition change.

In order to reduce drastic change of the shape of GEI, Huang *et al.* [11] fused two new gait representation: shifted energy image and the gait structural profile to increase the robustness to some classes of structural variations. But, the performance of this method is not good enough due to the loss of temporal information while calculating GEI. In [12], GaitSet has been proposed where a gait is regarded as a set consisting of independent frames rather than a template or sequence. Though it handled cross-view conditions very well, it is not good enough in handling cross-carrying and cross-clothing conditions.

2.2 Model-based Methods

Model-based methods [13–16] are based on the extraction and modeling of the human body structure as well as the local movement pattern of these parts. Therefore, this approach is often built with a structural and a motion model to capture both static as well as dynamic information of gait. For example, In [13], Yam *et al.* developed an automated model-based approach to recognize people using walking as well as running gait by analyzing the leg motion. They used the Biomechanics of human locomotion and coupled oscillators and employed a bilateral symmetric and an analytical model to successfully extract the leg motion. Ariyanto *et al.* [14] employed a structural model including articulated cylinders for fitting the 3D volumetric subject data at each joint to model the lower legs. In [15], authors presented a model-based approach where they captured the discriminatory features of gait by analyzing the leg and arm movements. For recognition, they used K-nearest neighbor classifier and Fourier components of the joint angle.

So, Model-based approaches are generally invariant to various intraclass variations like clothing, carrying and view angle variations, etc. However, the main drawback of this approach is the extraction process of body parameters like height, knee, and torso which is computationally expensive and highly dependent on the quality of the video.

2.3 Deep Learning for Gait Recognition

Due to its powerful feature learning abilities, convolutional neural networks (CNNs) have achieved great success in object recognition task in recent years. Several CNN-based gait recognition methods [17–22] have been proposed which can automatically learn robust gait features from the given training samples. Additionally, using CNNs, we now can execute feature extraction and perform recognition within a single framework using train samples. Wu *et al.* [17] performed cross-view gait recognition by developing three convolutional layer network using the subject's GEI as input. Shiraga *et al.* [18] designed a eight-layered CNN network, GEINet, which consist of two sequential triplets of convolution, pooling, normalization layers, and two fully connected layers for large-scale gait recognition on OU-ISIR database.

In [19], Wolf *et al.* used 3D convolutions for multi-view gait recognition by capturing spatio-temporal features from raw images and optical flow information. A Siamese neural network-based gait recognition system has been developed in [20] where GEI was feed as input. In [21], Yu *et al.* used generative adversarial nets to design a feature extractor in order to learn the invariant features. In [22], they further improved the GAN-based method by adopting a multi-loss strategy to optimize the network to increase the inter-class distance and to reduce the intraclass distance at the same time.

2.4 Pose-based Gait Recognition

In recent years, there has been a huge interest in the study of deep learning-based approaches for the task of real-time pose estimation from image and video. The task of pose estimation mainly involves localizing the keypoints of human figure to estimate the locations of different body parts. It can broadly be classified into two categories: single-person and multi-person pose estimation. To recognize multi-person pose, Cao *et al.* [1] developed a deep CNN-based regression method to estimate the association between anatomical parts in the image. Their bottom-up method achieved state-of-the-art performance on multiple benchmark datasets. In this work, we employed their pretrained model to get an accurate 2D pose estimation on our experimental dataset.

With the advent of the pose-estimation algorithms in computer vision, the recognition of human gait based on pose information has received much more attention [16, 23, 24] due to its effective representation of gait features and robustness toward covariate condition variations. Feng *et al.* [16] used the human body joint heatmap to describe each frame. They fed the joint heatmap of consecutive frames to long short term memory (LSTM). Their gait features are the hidden activation values of the last timestep. In [23], Liao et al. constructed a temporal-spatial network (PTSN) to extract the spatial-temporal features of gait from 2D human pose information. Authors in [24], employed 3D pose estimation in their PoseGait network to extract the spatial-temporal gait features and achieved better performance compared with 2D pose estimation.

Again, some of the most successful approaches for human action recognition employ RNNs [25, 26] to effectively model the temporal sequences of human skeleton data. Song *et al.* [25] proposed an end-to-end spatial and temporal attention model with LSTM for human action recognition from skeleton data. In [26], Du *et al.* proposed an end-to-end hierarchical RNN network for skeleton-based action recognition. They divided the human skeleton into five different parts and then separately feed them into five sub-networks.

Our approach to gait recognition is similar to these approaches. In this study, we have proposed a simple RNN architecture that effectively models the discriminative gait features in a temporal domain.

Chapter 3

Methodology

In this section, we are going to discuss the proposed framework and its main components in detail. The proposed method is efficient and computationally inexpensive compared to other methods proposed in literature.

3.1 Overview

The workflow of our proposed network is illustrated in Figure 3.1 Many strategies have been taken to design and train the network for robust recognition. Firstly, we designed a novel spatio-temporal gait feature descriptors based on the 2D human poses estimated from raw video frames using an improved OpenPose [1] algorithm. Thereafter, the gait descriptors were fed into a 2-layer BiGRU network which models the descriptors to recognize the subject ID.

3.2 Forming Spatio-Temporal Features

3.2.1 2D Body Joints Feature

As every joint of the human body does not have a significant role in gait pattern, they cannot improve gait recognition accuracy. Some joints perform even worse. So, among the 25 body joints estimated from OpenPose algorithm we searched out those joints which have a rich and discriminative gait representation capacity. Cunado *et al.* [27] used the human leg-based model as they found that change of human leg contains the most important features for gait recognition. In our study, we found that knee along with the joints located in feet show more robustness than any other body joints because they do not change while people

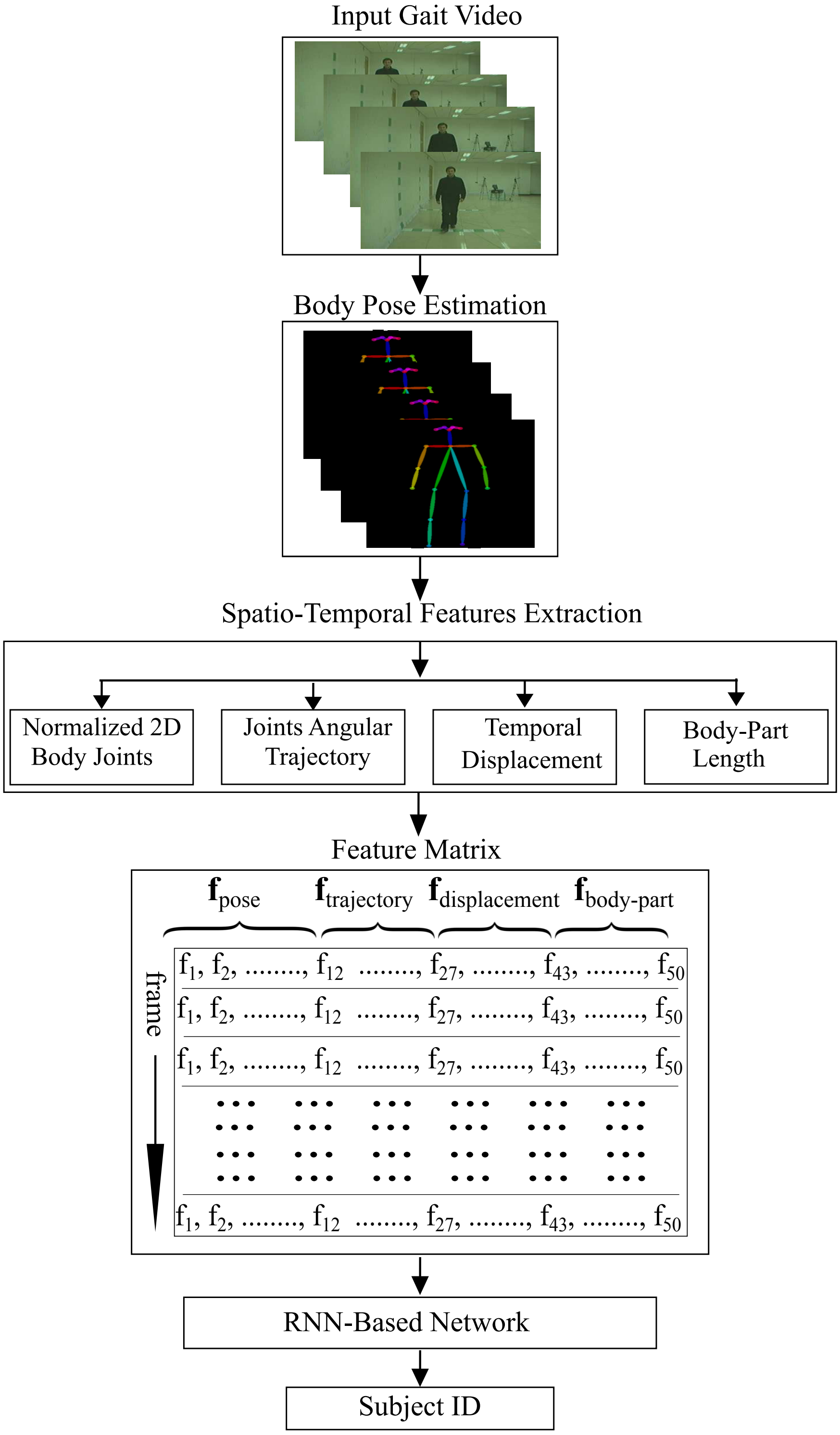


Figure 3.1: The overview of the proposed framework for gait recognition. 2D human poses were first extracted from raw video frames using improved OpenPose [1] algorithm. Effective body joints were then selected from each subject pose estimation and a timestep of 28 frame pose sequence was formed to feed into a temporal network. The temporal network identified the subject by modeling the gait features.

are walking in cloths or carrying bags. For example, hip joints get wider in coat than normal condition. Again, in some gait videos, some subjects put their hands into their coat pocket, which they cannot do in normal walking. This situation significantly changes the joint coordinates. Hence, we do not consider hip or any other body joints above it.

Consequently, in our work, as shown in Figure 3.2(a) we selected 6 body joints (RKnee, Rankle, RBigToe, LKnee, LAnkle, LBigToe) to form our effective pose features. Thus, we have 12-dimensional pose feature vector, $\mathbf{f}_{pose}$, for a single frame.

$$\mathbf{f}_{pose} = [x_1, y_1, x_2, y_2, \ldots\ldots, x_6, y_6]^T \tag{3.1}$$

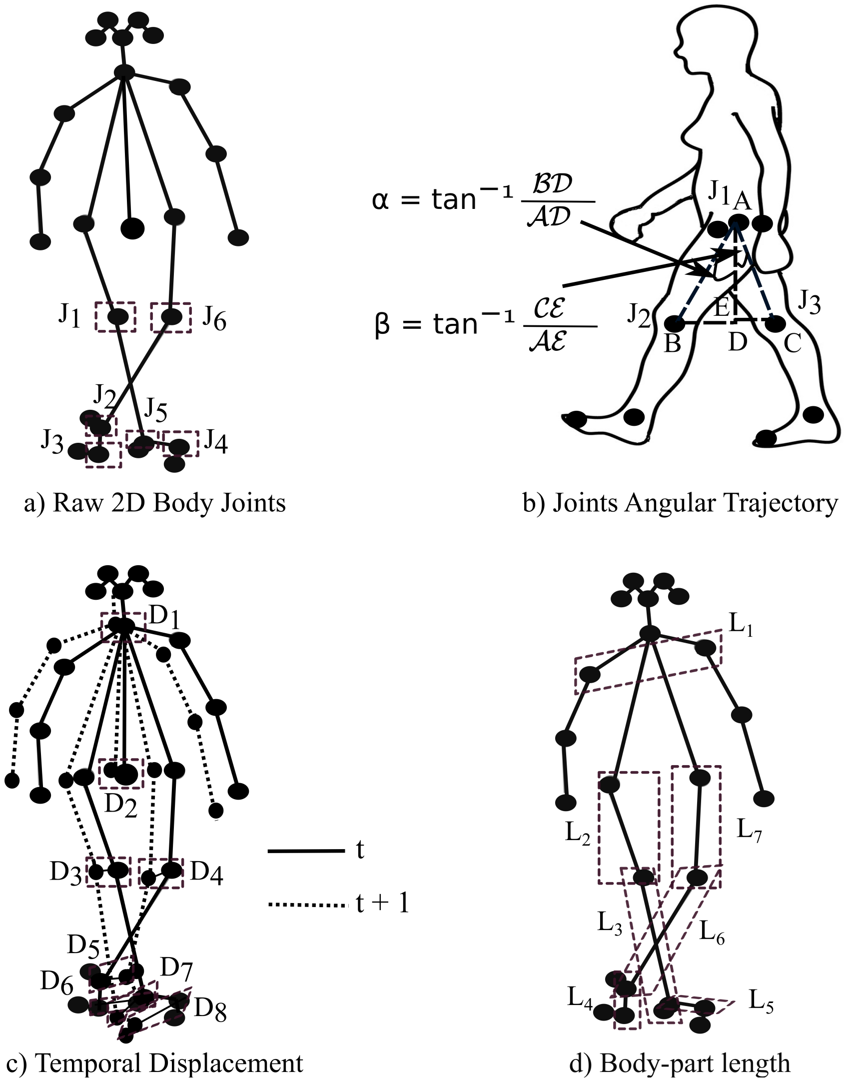
It is necessary to normalize the pose sequence data with regard to the subject position in frame, size, or speed of walking to get improved performance. In different gait datasets, as people walk through the fixed camera, the size of the subjects body changes due to change in the distance between the subject and the camera changes. In our study, to find the origin of the coordinate system (J_c) for each subject, we took right, left, and middle of the hip joints and calculated the average of them. Again, to normalize the bodies of different subjects to a fixed size, we took h, the euclidean distance from hip to neck joint, as unit length.

$$\begin{aligned} Jc &= (J_{LHip} + J_{RHip} + J_{MHip})/3 \\ h &= \parallel J_c J_{neck} \parallel_2 \\ J_i^N &= (J_i J_c)/h \end{aligned} \tag{3.2}$$

Here, ${J_i}^N$ be the new coordinate of J_i.

3.2.2 Joints Angular Trajectory

The dynamics of human gait motion can be expressed by the temporal information of joint angles. Hence, discriminative gait features can be found by considering the change in joint-angle trajectory of the lower limbs [28]. Therefore, in this study, we formulated a 15-dimensional joint angle features vector $\boldsymbol{f}_{trajectory}$ by considering five lower limb joint-angle trajectories using the following equations.



a) Raw 2D Body Joints b) Joints Angular Trajectory

c) Temporal Displacement d) Body-part length

Figure 3.2: Different features extraction process of the proposed method. a) 6 effective joints were selected out of 25 body joints as estimated from pose estimation algorithm [1] to form a 12-dimensional pose vector. b) 5 angular trajectories from lower limbs were considered to form a joint-angle feature vector. c) A total of 8 body joints were selected to get temporal displacement features. d) 7 body parts were taken to form a limb length feature vector.

Table 3.1: List of effective joint-angle trajectories with corresponding body joints set for gait angular feature vector

Angular Trajectory	**Body Joints Set**
Hip trajectory	10, 8, 13
Right knee trajectory	11, 10, 9
Left knee trajectory	14, 13, 12
Right ankle trajectory	22, 11, 10
Left ankle trajectory	19, 14, 13

$$\alpha = \begin{cases} \tan^{-1} \frac{|J_{2,x} - J_{1,x}|}{|J_{2,y} - J_{1,y}|} & J_{2y} \neq J_{1,y} \\ \pi/2 & J_{2,y} = J_{1,y} \end{cases}$$

$$\beta = \begin{cases} \tan^{-1} \frac{|J_{3,x} - J_{1,x}|}{|J_{3,y} - J_{1,y}|} & J_{3,y} \neq J_{1,y} \\ \pi/2 & J_{3,y} = J_{1,y} \end{cases} \tag{3.3}$$

$$\theta = \alpha + \beta$$

As shown in Figure 3.2 (b), J_1, J_2, J_3 are the joints which form a set of angular trajectory. In this work, we considered five sets of angular trajectories from the lower limb of human body. Table 3.1 demonstrated the selected angular trajectories with their corresponding body joints. For each trajectory, we took (θ, α, β)s gait features.

$$\mathbf{f}_{trajectory} = [\theta_1, \alpha_1, \beta_1, \theta_2, \alpha_2, \beta_2, \ldots\ldots, \theta_5, \alpha_5, \beta_5]^T \tag{3.4}$$

3.2.3 Temporal Displacement

Our third extracted features were a simple descriptor that preserves temporal information. It stores the local motion features of gait by keeping the displacement information between the two adjacent frames of the pose sequence. Each joint displacement of a frame was then normalized by the total length of all joints. Let, t and $(t + 1)$ are two adjacent frames. The displacement information of the coordinates of any body joint of frame t would be the normalized difference between the corresponding coordinates.

$$\begin{aligned} \triangle x_1^t &= \frac{x_1^{t+1} - x_1^t}{\sum_{i=1}^{8} \parallel J_i^{t+1} - J_i^t \parallel_2} \\ \triangle y_1^t &= \frac{y_1^{t+1} - y_1^t}{\sum_{i=1}^{8} \parallel J_i^{t+1} - J_i^t \parallel_2} \\ \mathbf{f}_{displacement} &= [\triangle x_1, \triangle y_1, \triangle x_2, \triangle y_2 \ldots\ldots, \triangle x_8, \triangle y_8]^T \end{aligned} \tag{3.5}$$

Here, J it is the 2D coordinates of the i_{th} body joint at t_{th} frame in the video and $(\triangle x_1^t, \triangle y_1^t)$ is the displacement of the coordinates of first joint at t th frame of the video. As shown in Figure 3.2 (c), we selected 8 joints (Neck, MHip, RKnee, Rankle, RBigToe, LKnee, LAnkle, LBigToe) to get a 16-dimensional $\mathbf{f}_{displacement}$ vector.

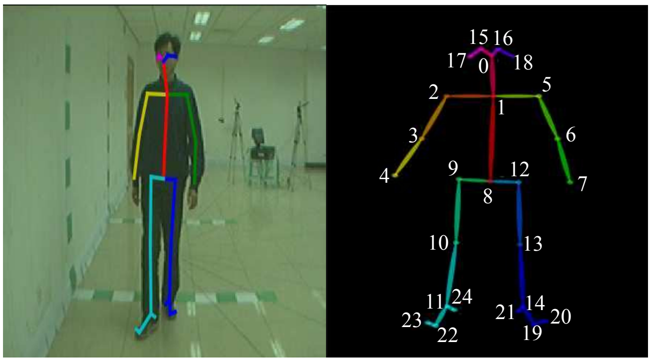


Figure 3.3: Examples of 2D human pose estimation by [1] from RGB images of CASIA dataset (left ones). Detected 25 human body joints with description are shown. (right ones)

3.2.4 Body Part Length Features

These static gait parameters, e.g., the length of the body parts calculated from joints position are also important for gait recognition [28, 29]. They form spatial gait features which make them robust against covariate such as carrying and clothing variation. In this work, we took seven body parts (Figure 3.2 (d)) namely length of the two leg, two feet, two thigh and width of the shoulder which formed a 8-dimensional spatial feature vector $\mathbf{f}_{body-part}$.

3.2.5 Fusion of Features

Many research works have applied techniques to fuse multiple features to get improved performance [24, 28]. Different types of fusion methods are proposed in literature such as feature level fusion, representation level fusion, and score level fusion. In feature level fusion, multiple features of the same frame are concatenated before feeding into a final network and in representation level, fusion each feature vector is firstly fed into a network and the resulting global representations are then concatenated to train a final classifier. For score level fusion, each feature vector is separately fed into the final network which predicts a classification score. Then, the scores from multiple classifiers are fused using an arithmetic mean.

In this study, we found that features level fusion has produced better recognition results in contrast to other fusion techniques or individual feature sets.

3.3 Feature Preprocessing

From 2D pose estimation algorithm [1], we got 25 body joints from each frame (Figure **??**). We took several preprocessing steps to address the problem of missing data due to occlusions. The main strategies are:

- If the origin of the coordinate system can't be calculated due to missing hip joints, the frame should be rejected.
- If more than 1 body joint is missing in between knee and ankle joints of both leg, the frame should be rejected due to having little information.
- In other cases, individual joints were not located in the frame and a position of $[0.0, 0.0]$ was given to that joint.

The above strategies are simpler which do not require any computation and proven to be effective in addressing the missing data problem. Thus, we designed a 50-dimensional spatial-temporal gait feature vector **p** from the raw 2D pose estimation of each frame.

We split a gait video into 28 frames segments. Each 28 frame- segment formed a timestep which can be described by equations (6). Here, **p** is 50-dimension pose vector for each frame; ***T*** is the feature matrix for each timestep; N is the total number of timestep sequence, and ***V*** is the sequence of features for a gait video.

$$\begin{aligned} \boldsymbol{p} &= [f_1, f_2, f_3, \ldots\ldots, f_{50}]^T \\ \boldsymbol{T} &= [\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \ldots\ldots, \boldsymbol{p}_{28}]^T \\ \boldsymbol{V} &= [\boldsymbol{T}_1, \boldsymbol{T}_2, \boldsymbol{T}_3, \ldots\ldots \boldsymbol{T}_N]^T \epsilon \quad \mathbb{R}^{28\times 50} \end{aligned} \tag{3.6}$$

In CASIA dataset, gait videos of different subject have varying timesteps. The number of timesteps in each gait video depends on the total number of frames where a person is detected. Due to the position of the camera, some angles $(0°, 18°, 36°)$ have more person detected frame than other angles $(72°, 90°, 108°)$. Therefore, the total number of timesteps in a gait video is different for different subjects and viewing angles. This eventually makes our train dataset unbalanced. Again, in CASIA B dataset, not all subjects have all gait videos; there are some missing gait videos. To solve the problem, we develop our own balance training set by making each subject pose sequence to have a fixed number timesteps. We first found the subject which had maximum timesteps for a particular gait angle and then augmented other subject's timesteps with that specific length by overlapping their sequences.

3.3.1 Data Augmentation

The performance of deep neural networks is strongly correlated with the amount of available training data. Although CASIA is the largest gait dataset, the standard experimental setup of this dataset (see Table IV) allows us to train only the four normal walking sequence of each subject. Therefore, we need to augment our train data to obtain a stable model. One way to increase the amount of training data is to overlap video clips Therefore, we split the input video into an overlapping sequence of video clips. For every 28 image clip, we overlapped 24 images of the previous clip at almost 85.7% overlapping rate. For example, a particular gait video of 100 frames would be split into the clips $(1-28)$, $(5-32)$, $(9-36)$, ... up to frames $(73, 100)$.

In addition to above technique, we further augment our training data by adding another gait sequence (i.e., 25% increment) by implementing Gaussian noise to a given normal walking sequence.

$$N(j_i) = (x + \tilde{x}, \quad y + \tilde{y}) \tag{3.7}$$

Here, $\tilde{x}$ and $\tilde{y}$ are two random real numbers generated by a normal distribution with zero mean and unit standard deviation. We apply noising (N) into the raw joints position of a training pose data.

3.3.2 Network Architecture

In this research, we experimented with different RNN architectures such as Gated Recurrent Units (GRUs), Long Short Term Memory Units (LSTMs), and Bidirectional Gated Recurrent Units (BiGRU) [2]. Firstly, we designed the proposed network employing all these architectures with one recurrent layer and then, searched for optimum recurrent unit size between 50 to 150. Thereafter, we increased the capacity of the network by adding the second and third layers of hidden units. Finally, we found that, among different RNN-based architectures, 2-layer BiGRU with 80 hidden units performs best.

After input and the second recurrent layer, we placed a batch normalization (BN) [3] layer. At last, a fully connected layer with softmax activation was used to output the subject classes. Figure 3.4 illustrates the architecture of the proposed network.

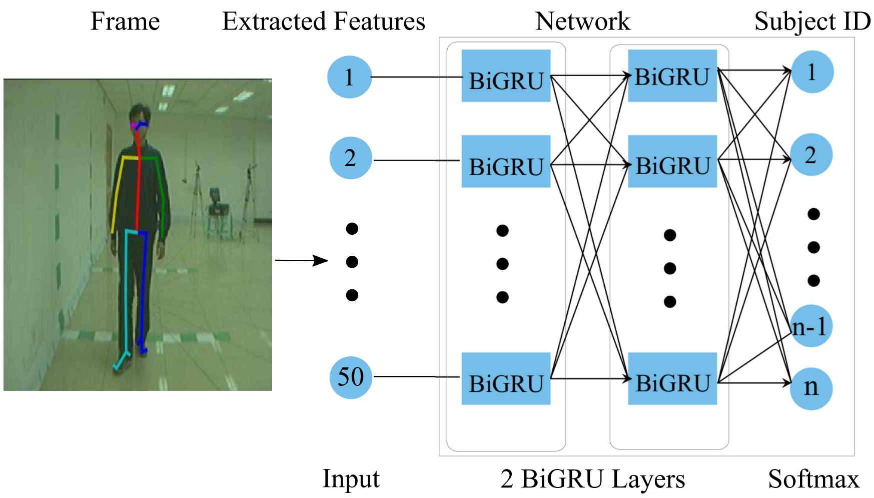


Figure 3.4: Proposed RNN architecture for robust gait recognition. It consists of two Bi-GRU [2] layers each of which consists of 80 GRU cells with one batch normalization and one output softmax layer. The network was fed with 50 dimensional input from 2D pose estimation. Input layer was followed by a batch normalization layer [3]. The output of recurrent layers was also batch normalized to standardize the activations and finally fed into an output softmax layer. For the output layer, the number of the output neuron equals to the number of subjects.

Table 3.2: Training summary of our proposed temporal network.

Hyperparameter	**Value**
Optimizer	Adam [30]
Objective function	Fusion of softmax and center loss
Epochs	450
Initial learning rate	5×10^{-3}
Mini-batch size	256

3.3.3 Training

The training of RNNs allows us to learn the parameters from the sequence. We have employed Adam [30] optimization algorithm with $\beta_1 = 0.9, \beta_2 = 0.999$, which is known to work very well for training recurrent neural networks. We tried several learning rates in our experiment and found out that the best initial learning rate is $(1\text{x}10^{-3})$. We also reduced the learning rate by a factor when it hit a plateau. Reducing the learning rate will allow the optimizer to get rid of the plateaus in loss surface. Table 3.2 summarizes all the hyperparameters settings of our network.

Our network was trained with a 256 batch size and 28 image clips with one timestep. Our network showed some overfitting mostly due to the high learning capacity of the network over data. This overfitting problem has been addressed by adding a batch normalization layer.

We also tried to add dropout layer during training, but that did not help to reduce the overfitting problem. Moreover, it degraded gait recognition performance. Hence, we skip it.

3.3.4 Loss functions

In this work, we found that due to the influence of various covariate factors, intraclass distance related to one subject is sometimes more significant than interclass distance. Now, if we only use the *cross-entropy loss* as our objective function, the resulting learned features may contain large intraclass variations. Therefore, to effectively reduce the intraclass distance, we used *center loss*, introduced by Wen *et al.* [31] for face recognition task. As the training progresses, the center loss learns a center for features of each class and the distances between the features and their corresponding class centers are minimized simultaneously. However, using only center loss may lead the learned features and centers close to zeros due to the very small value of the center loss. Hence, with the fusion of softmax loss (L_s) and center loss (L_c), we can achieve discriminative feature learning by increasing interclass dispersion and compacting intraclass distance as much as possible.

$$
\begin{aligned}
L_s &= -\sum_{i=1}^{m} log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \\
L_c &= \frac{1}{2} \sum_{i=1}^{m} \parallel \boldsymbol{x}_i - \boldsymbol{c}_{y_i} \parallel_2^2 \\
L &= L_s + \lambda L_c + \lambda_\theta \parallel \theta \parallel_2
\end{aligned}
\tag{3.8}
$$

Equations (3.8) describes total loss (L) calculation of our network. where $\boldsymbol{x}_i \epsilon \mathbb{R}^d$ denotes the ith pose sequence which belongs to the y_ith class and $\boldsymbol{c}_{y_i} \epsilon \mathbb{R}^d$ denotes to the y_ith class center of the learned pose features. $W \epsilon \mathbb{R}^{d \times n}$ is the feature dimension of the last fully connected layer and $b \epsilon \mathbb{R}$ is the bias term of the network. The batch size and the class number are m and n respectively. λ, a scalar variable, is set to value 0.01 to balance the two loss functions. $\parallel \theta \parallel_2$ refers to the kernel regularizer for all the parameters of the network with a weight decay coefficient (λ_θ) set to 0.0005 for the experiment.

3.3.5 Post-processing

While training, our proposed temporal network considers each of these video clips as a separate video (see Fig. 3.5). For a given video, the prediction of our model is a sequence of class probabilities for each of the timestep, i.e. 28 frame clips.

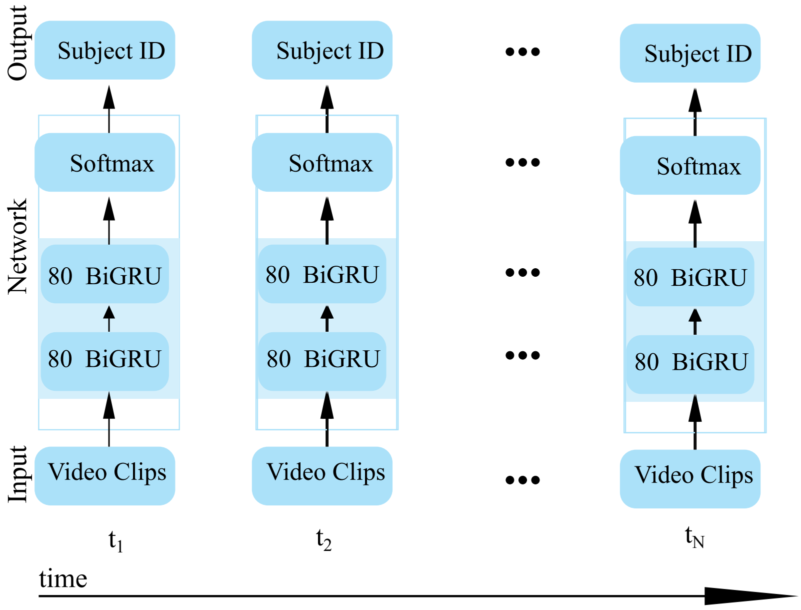


Figure 3.5: Output prediction scheme of our proposed temporal network. Each input clip was considered as a separate video and a sequence of class probabilities was predicted at output. Majority voting scheme was used to process the output to predict the subject ID.

But, while testing, we actually need the subject ID of the complete gait video. Therefore, we used *majority voting scheme* to process this output to predict the subject ID. In this scheme, the subject that receives the highest number of votes over all timesteps of a particular gait video is referred as the predicted class.

Let's consider, $\boldsymbol{s}$ is a vector of n number of subjects. For a particular timestep t of a gait video, input pose sequence vector $\boldsymbol{X}^t \epsilon \mathbb{R}^{28\times 50}$ has an n-dimensional output vector $\boldsymbol{o}^t$.

$$\begin{aligned} \boldsymbol{s}^t &= [s_1, s_2, s_3, ..., s_n]^T \\ \boldsymbol{o}^t &= [o_1, o_2, o_3, ..., o_n]^T \end{aligned} \tag{3.9}$$

Here, $o_i^t = P(s_i|X^t)$ refers the probability of input pose vector $\boldsymbol{X}^t$ belongs to class s_i. Now, we assign the output class s^t to the subject class s_i which have maximum probabilities for the timestep t. As each of our gait videos is divided into a series of timestep sequence (see equation **??**), using majority voting scheme we can have the subject ID. Following equations described the voting scheme.

$$\begin{aligned} s_t &= \arg\max_{s_i} \{o_i^t | 1 \leq i \leq n\} \\ s &= \arg\max\nolimits_{i\in(1,2,...,n)} \sum_{t=1}^{N} s_i^N \end{aligned} \tag{3.10}$$

Here, N is the total number of timesteps in which a gait is split and s is the final predicted

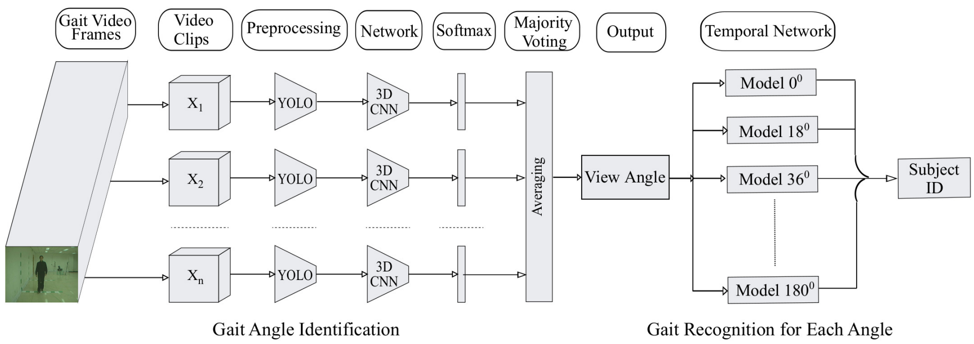


Figure 3.6: Overview of our proposed multi-view gait recognition network scheme. YOLOv3 [4] was used to detect and locate the walking people in video frames. The input of the network was a clip of 16 consecutive frame which was preprocessed and resized to 112×112 to feed into a 3D convolutional network based on C3D [5]. The network used 3D kernels to exploit spatio-temporal dynamics for viewing angle identification. Thereafter, a temporal network, trained on each viewing angle, was performed subject identification by modeling temporal dynamics from input 2D pose sequence

class.

3.4 Multi-View gait recognition

The workflow of our proposed two-stage multi-view gait recognition network is illustrated in Fig. 3.6. In first stage, we trained a 3D convolutional network to estimate the walking direction of the subject by extracting spatio-temporal features from gait video. Thereafter, we performed subject recognition using proposed temporal network which has been trained for that particular angle.

3.4.1 Preprocessing

Firstly, to localize human walking in gait videos, we used YOLOv3, a state-of-the-art real-time object detection algorithm, proposed by Redmon *et al.* [4]. We then cropped each of the person detected frame using the bounding box coordinates found from YOLOv3 algorithm and resized them to 112×112 for our network input. Thereafter, we splitted each gait video into overlapping sequences of 16 consecutive frames within training or test set. There is an overlap of 8 frames (50%) indicating that the samples were gathered using a 16 frame sliding window with a 50% stride.

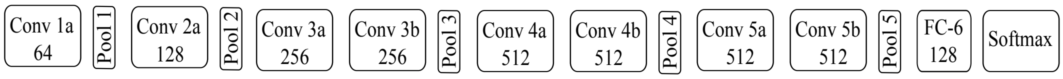


Figure 3.7: Proposed 3D-CNN for video angle identification. Last 3 layers of a pretrained C3D [5] network has been replaced by a fully connected layer of 128 neurons followed a final softmax layer of 11 neurons to classify 11 different walking direction in CASIA-B dataset.

Table 3.3: Training summary of our proposed 3D-CNN network.

Hyperparameter	**Value**
Optimizer	Stochastic gradient descent (SGD)
Objective function	Mean squared error (MSE)
Epochs	70
Initial learning rate	1×10^{-3}
Mini-batch size	12
Momentum	0.92

3.4.2 Network architecture

Identifying walking direction from gait video is somewhat similar to action recognition problem in computer vision. Recently, in action recognition, researcher have started to exploit 3D features in video using 3D-CNN model which extracts features from both spatial and temporal dimensions by performing 3D convolutions. Tran et.al. [5] proposed a 3D convolutional neural network, also known as C3D, which has been widely used for applications like video classification, action recognition, etc. This network has been trained on one of the largest video classification benchmark datasets Sports-1M [32]. The dataset contains 1.1 million sports videos, where each video belongs to one of the 487 sports categories.

The proposed method for our gait angle identification is illustrated in Fig. 3.6. The input of the network was a clip of 16 consecutive frame which was preprocessed and resized to 112×112 to feed into a a 3D-CNN network. We used *majority voting scheme* to process the output to predict the view angle similar to section 3.3.5, i.e. the angle that receives the highest number of votes over all clips are referred as predicted angle of the video.

Successful transfer learning within or across different domain of interest leads to significant improvement in performance due to the amount of jointly learning representations in a shared feature space. In our work, we used a pretrained C3D model and fine-tuned it for our 3D Convolutional network to determine the viewpoint angle from gait videos. Fig. 3.7 shows our proposed 3D convolutional network.

C3D network is composed of 8 convolutional layers, 5 pooling layers, 2 fully-connected layers, followed by a softmax layer at the end. All the 3D convolution kernels are $3 \times 3 \times 3$

with stride 1 in both spatial and temporal dimensions. We removed the last 3 layer from the model and then added a fully connected layer of 128 neurons and a dropout layer of 0.5 to avoid overfitting. Finally, a softmax layer of 11 neuron has been added to classify any given videos into 11 different viewing angles.

3.4.3 Training

We used CASIA-B gait dataset [6] to train our model. We trained the network using 4 normal walking sequences of 100 subject in gallery set of CASIA-B as described in Table 4.2. Our network was trained with a 12 batch size with an initial learning rate 10^{-3} for 70 epochs. Table 3.3 summarizes all hyper-parameters setting of our proposed network.

Chapter 4

Results & Discussion

This section briefly discusses the datasets we used to train and evaluate our model, and the results our proposed algorithm achieved at different experimental setup. As to estimate pose, RGB video frames are required. So, we couldn't evaluate our method to those dataset which only consists of silhouette sequences.

4.1 Dataset

The success of deep learning-based methods greatly depends on the vast amount of labeled train data. Unfortunately, few existing gait databases contain a large number of subjects as well as a variety of covariate factors. Some of the publicly available gait databases are CASIA gait dataset [6], TUM GAID dataset [**?**], OU-ISIR multi-view large population dataset (OU-MVLP) [**?**] and USF HumanID dataset [**?**].

In USF HumanID gait dataset, there are 122 subjects walking outside on two different surfaces of an elliptical path under two different time, viewpoint, clothing, shoes, and carrying conditions. However, not all subjects were filmed under all conditions. TUM GAID dataset is another large dataset for gait recognition which consists of 305 subjects where each subject has 10 videos. But this dataset is not suitable for our multi-view gait recognition as all the videos are recorded from side view angle. The largest dataset available for gait recognition is OU-ISIR multi-view large population dataset (OU-MVLP). It contains 10,307 subjects from 14 viewing angles ranging from $0° - 90°$, $180° - 270°$. Only two sequences are provided, one for the gallery and the other for the probe. But, this dataset is formatted only as a set of silhouette sequence making it different from our approach.

In this study, we used CASIA (both CASIA A and CASIA B) dataset which is one of the largest multi-view gait databases. CASIA A dataset contains total 20 subjects walking in an outdoor environment where CASIA B dataset includes total 124 subjects walking in an

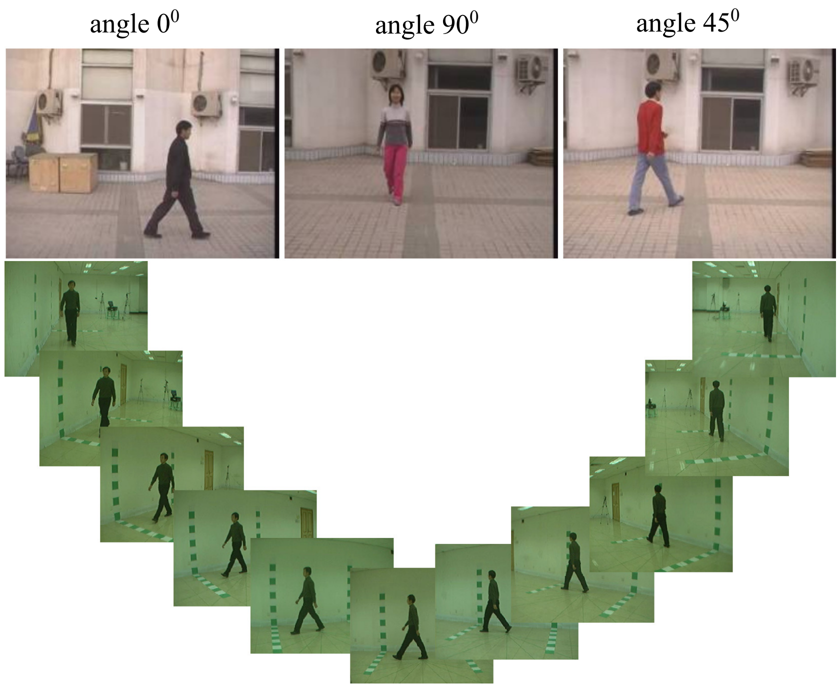


Figure 4.1: Sample video frames of CASIA A and CASIA B dataset. In top, some of the sample images from CASIA A dataset are shown where subjects walking along straight line in 3 different view angle, and in bottom, CASIA B dataset is shown with its 11 viewing angle.

indoor environment. In CASIA A gait dataset, each subject walks along a straight line in 3 different view angles lateral (0°), oblique (45°) and frontal (90°). For each viewing angle every subject has four gait sequences of which two of them have same walking direction while the other two have opposite direction. In CASIA B dataset, there are 10 walking sequences of each subject captured from 11 view angles: 6 sequences for normal walking ('nm'), 2 sequences for walking in a coat ('cl') and 2 sequences for walking with bag ('bg') on shoulder. Hence, this dataset separately considered three variations in people walking namely viewing angle, clothing and carrying conditions. The view angle set of the camera is ranging from 0° to 180°. Figure 4.1 illustrates some of the sample video frames of CASIA dataset.

4.2 Single-view Gait Recognition

4.2.1 Experimental Evaluation on CASIA-A dataset

Since, CASIA A dataset contains only 20 subjects each of which have only four gait sequence in three different angles, we trained three model for each of the gait angle with 20 output neurons in the final softmax layer of our proposed temporal network. To evaluate the performance of our proposed method on CASIA A dataset, we used leave-one-out cross

Table 4.1: Comparison among different state-of-the-art gait recognition methods without view variation on all three view angles of CASIA-A dataset. It has been seen that, proposed method achieves best performance in correct class recognition rate on view angle $45°$. Overall, it gets higher average recognition rates **98.3%** and outperforms other state-of-the-art methods by a large margin.

Methods	0°	45°	90°	Mean
Wang [?]	88.75	87.50	90.00	88.75
Goffredo [?]	100.0	97.50	91.00	96.16
Liu [?]	85.00	87.50	95.00	89.17
Lima [?]	92.50	97.50	98.75	96.25
Kusakunniran [?]	100	100	98.75	99.58
Proposed	**100.0**	**100.0**	**100.0**	**100.0**

validation rule, i.e., one sequence was set for testing and the remainder was set for training the network for each view angle. We compare our results with four other prevailing state-of-the-art gait recognition methods including Wang [?], Goffredo [?], Liu [?], Lima [?], Kusakunniran [?] (see Figure 4.2). Table 4.1 illustrates that the proposed method have achieved higher average correct class recognition rates (CCR) 100.0% compared to other methods.

4.2.2 Experimental Evaluation on CASIA-B Dataset

Experimental Setup

We designed two experimental setups (A, B) in CASIA B dataset for evaluation. Experiment setup A is for evaluating the performance of the proposed method in single-view gait recognition. To investigate the robustness of view variation, comparison results of the proposed approach against other state-of-the-art methods in different view variations have been reported. Experiment setup B is designed for evaluating the cross-view recognition performance.

For setup A, as demonstrated in Table IV, we divided the dataset into two groups where the first group which consists of 62 subjects was used to train the network. The second group contains rest of the subjects for evaluating the performance of the model. For setup B, the ratio between train and evaluation set was 24 to 100. In the evaluation set for both setup, 4 normal walking sequences of each subject are put into gallery set and rest 6 walking sequences consist three probe set (*ProbeNM*, *ProbeBG*, *ProbeCL*). *ProbeNM* consists of 2 other normal walking sequences where *ProbeBG* and *ProbeCL* consists of two subjects carrying bag and wearing coat respectively.

We divide CASIA-B dataset into two group where first group consists of 24 subjects and

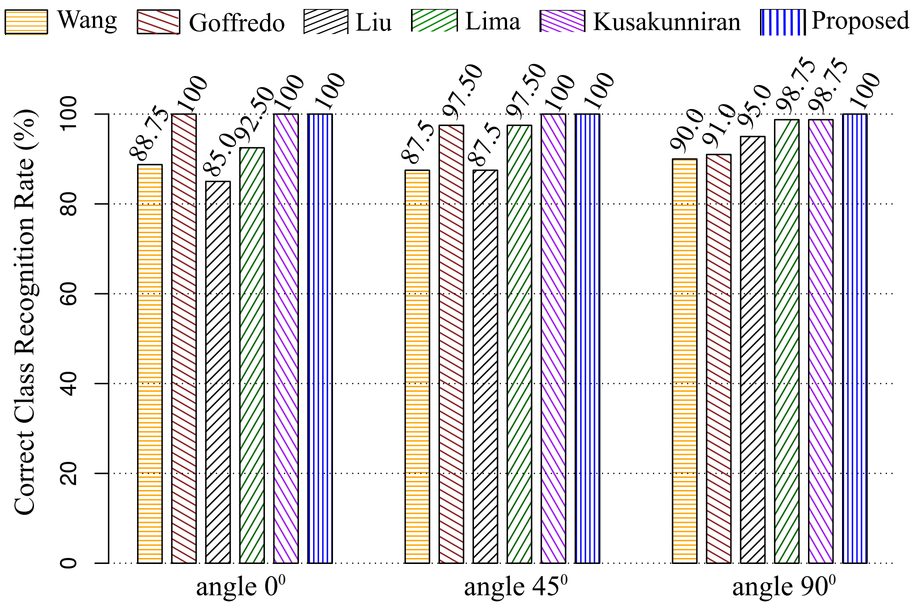


Figure 4.2: Comparison in correct class recognition rate (CCR) at different viewing angles among proposed method with other prevailing gait recognition methods proposed in literature on CASIA A dataset. Our method achieves 100% class recognition rate on all of the view angles which proved the efficacy of the proposed method.

Table 4.2: Experimental setup for the CASIA B dataset. The dataset is divided into two different setups to organize two different types of experiment. the evaluation is subdivided into a gallery set and a probe set. Gallery set consists of the first 4 normal walking sequences of each subject and the probe set contains rest of the walking sequences

Setup	**Training set**		**Evaluation set**		**Sequences**	
	ID	Total	ID	Total	Gallery	Probe
A	01 - 62	62	63 - 124	62	$nm01 - nm04$	$nm05 - nm06$ $bg01 - bg02$ $cl01 - cl02$
B	01 - 74	74	75 - 124	50		

is used to train the network. The second group contains rest 100 subjects that are used to evaluate the performance of the model. In evaluation set, 4 normal walking sequences of each subject are put into gallery set and rest 6 walking sequences consist three probe set (*ProbeNM*, *ProbeBG*, *ProbeCL*). *ProbeNM* consists of 2 other normal walking sequences where 2 subjects carrying bag are kept in *ProbeBG* and remaining 2 subjects wearing coat are kept in *ProbeCL*. Table 4.2 shows this experimental setup.

Results on Single-View Gait Recognition of CASIA B Dataset without View Variation

Experimental results of single-view gait recognition on all the three probe set of CASIA B dataset without view variation is illustrated in Table 4.3. We achieved higher average recognition rate 97.80% and 82.82% on the probe set of (*ProbeBG*) and (*ProbeCL*) respectively.

Table 4.3: Correct class recognition rate (CCR) of proposed method in all three probe sets of CASIA B dataset. Here, column represents a specific view of gallery and probe set. It has been observed that the probe set of normal walking (*ProbeNM*) achieves **99.41%** average recognition rate while the ProbeBG and ProbeCL set achieve **97.80%** and **82.82%** average recognition rates respectively.

Probe Type	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	M
ProbeNM	100	100	100	100	100	98.39	100	100	100	98.39	96.77	**99**
ProbeBG	100	100	100	100	98.39	96.77	96.77	98.39	98.39	95.16	91.93	**97**
ProbeCL	81.52	82.11	83.58	85.48	84.46	83.72	83.28	84.16	83.58	80.65	78.45	**82**

Table 4.4: Comparison between the proposed method and other state-of-the-art gait recognition methods in CASIA B dataset without view variation. It has been observed that the proposed method outperforms other methods in all three probe set of CASIA B dataset. As the proposed method doesn't depend on any body point higher than knee, it shows the robustness towards these covariate factors. It also achieves higher average correct class recognition rate (CCR) **93.34%** by outperforming other methods at a significant margin.

Methods	*ProbeNM*	*ProbeBG*	*ProbeCL*	Average
Liao *et al.* [24]	96.92	85.78	68.11	83.60
Yu *et al.* [**?**]	97.58	72.14	45.45	71.72
Yu *et al.* [22]	98.24	76.25	42.89	72.46
Liao *et al.* [24]	96.63	71.26	54.18	74.02
Proposed	**99.41**	**97.80**	**82.82**	**93.34**

This performance proves the robustness of our proposed method towards both carrying and clothing covariate conditions. We also achieved higher average class recognition rate *99.41%* on normal walking condition.

Comparison on Single-View Gait Recognition of CASIA B Dataset with State-of-the-art Methods without View Variation

We compare our experimental results with other state-of-the-art methods such as GaitGANv2 [22], PTSN [23], PoseGait [24], Yu *et al.* [**?**] as shown in Figure 4.3. The experimental setup for all these methods were set A (see Table 4.2). Table 4.4 reports that CCR of the proposed method outperforms all other methods in all three covariate conditions of CASIA-B dataset; our method achieved average CCR of **93.34%** with improvement of approx. **10%** from PTSN.

Results on Single-View Gait Recognition of CASIA B Dataset with View Variation

The performance of the proposed method on single-view gait recognition with view variation is demonstrated on Table 4.6. Here, for a specific gallery (θ_g) angle the average CCR

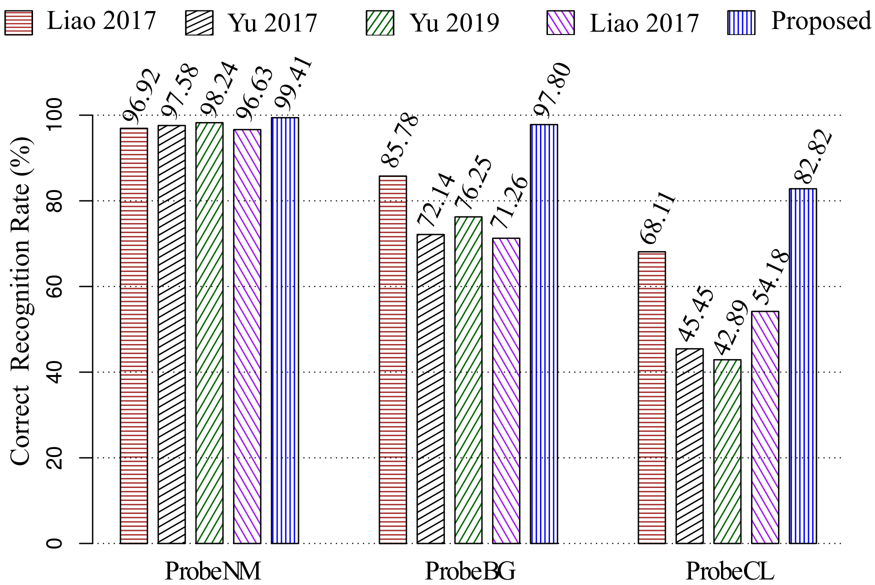


Figure 4.3: Correct class recognition rates (%) of the proposed method with other state-of-the-art methods on all three probe set of CASIA-B dataset without view variation. Proposed method demonstrates better performance compared to other by achieving 89.64% and 96.45% in two covariate conditions of CASIA-B dataset *ProbeCL*, and *ProbeBG* respectively. The result proves the robustness of proposed pose-based temporal network against carrying and clothing conditions variations.

Table 4.5: The average recognition rates for all three probe sets of CASIA B dataset. Each row represents the average value of all eleven probe angles at a specific gallery angle (θ_g) in all three probe sets.

Probe Type	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	M
ProbeNM	61.73	63.64	67.30	68.33	68.33	66.42	64.22	62.02	58.80	56.45	52.35	**6**
ProbeBG	45.01	47.80	48.97	50.15	50.44	49.12	48.39	47.07	47.51	44.13	40.91	**4**
ProbeCL	32.40	32.99	34.46	37.24	39.0	36.36	34.75	32.40	31.82	29.77	26.83	**3**

(%) of all eleven probe angles has been reported; our method achieved average CCR of 62.69%, 47.23%, and 33.46% for Probe NM, probe BG, and probe CL respectively.

Comparison on Single-View Gait Recognition of CASIA B Dataset with State-of-the-art Methods with View Variation

To better illustrate the robustness to view variation of our gait recognition method, the proposed method has been compared to three other state-of-the-art methods such as GaitGANv2 [22], PoseGait [24], Yu *et al.* [**?**]. It is observed from Figure 4.4 and Table 4.6 comparison that proposed method outperforms other in covariate variation and achieves comparable performance in normal walking.

Since, to recognize gait, we consider features based on effective body joints, hence our method doesn't get affected by the variation in covariate conditions compared to other

Table 4.6: Comparison among different state-of-the-art methods for gait recognition with view variation in all three probe sets of CASIA B dataset. Each row represents the average value of all the gallery view's average recognition rate. It has been seen that, similar to first experiment, the proposed method achieves higher performance in two different probe set (*ProbeBG*, *ProbeCL*) and comparable performance in normal walking with to other prevailing methods.

Methods	*ProbeNM*	*ProbeBG*	*ProbeCL*
Yu *et al.* [?]	62.82	40.38	26.05
Yu *et al.* [22]	66.34	46.17	25.91
Liao *et al.* [24]	63.78	42.52	31.98
Proposed	**62.69**	**47.23**	**33.46**

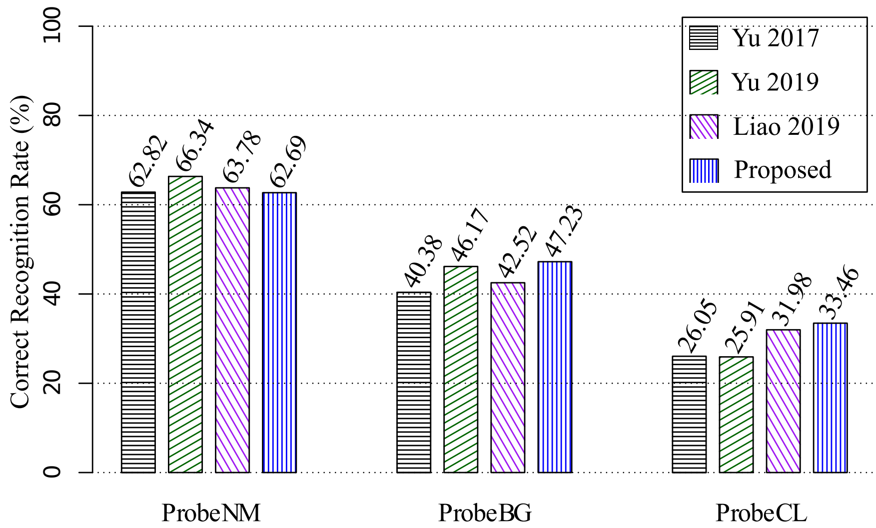


Figure 4.4: Comparison with different state-of-the-art methods for gait recognition with view variation in all three probe set of CASIA B dataset. Here, the value reported for each algorithm is the average of all the gallery view's average CCR. Proposed method outperforms other state-of-the-art methods achieving 47.23% and 33.46% in two covariate conditions *ProbeBG*, and *ProbeCL* respectively.

appearance-based method or other model-based methods which consider ineffective features to build their gait descriptor. Thats why our method is proven to be less sensitive to view change and performs better in carrying-bag and clothing condition.

4.3 Cross-View Gait Recognition

The gait recognition scheme in which gallery and probe set are getting matched from two different views is commonly known as cross-view gait recognition.

Table 4.7: Comparison of our proposed method with the previous best results of cross-view gait recognition at different probe angles of CASIA B dataset by CCR (%). The network was trained according to experimental setup B to have the same setup with other methods.

Gallery View	18°	36°	18°	36°	72°	90°	54°	72°	108°	126°	90°	108°	14
CNN	95.0	73.5	**91.5**	98.5	98.5	**93.0**	–	**99.5**	**99.5**	–	**92.0**	**99.0**	97
CMCC	85.0	47.0	65.0	97.0	95.0	63.0	66.0	96.0	95.0	68.0	78.0	98.0	**98**
GEI-SVR	84.0	45.0	64.0	95.0	93.0	59.0	63.0	95.0	95.0	65.0	78.0	98.0	**98**
Proposed	**97.0**	**80.0**	83.0	**100.0**	**100.0**	83.0	**84.0**	96.0	95.0	**71.0**	76.0	92.0	96

4.3.1 Comparison with the State-of-the-art Methods of CASIA B Dataset on Cross-View Gait Recognition

To show the effectiveness of our method in cross-view recognition, we make the comparisons between the proposed method and three other state-of-the-art methods including CNN [17], CMCC [?], and GEI-SVR [?] with the same experimental setup. The probe angles were selected $0°, 54°, 90°$, and $126°$ for comparison.

Although, the proposed method contains only one model to handle any view angle variation, it achieves comparable performance with other prevailing state-of-the-art methods proposed in literature which were specially designed and trained for cross-view gait recognition. From Table [?], it is seen that CNN [17] achieves the highest recognition rates when the view variation is large due to the use of supervised information of all gallery angles during training.

The comparison in Table IX also illustrates that the proposed method performs better when the view variation is small. The reason for not achieving better performance at large view variation is because it was trained with only one viewing angle.

4.4 Multi-View Gait Recognition

In multi-view gait recognition, multiple views of gallery gaits are combined to recognize probe set for an unknown gait view. In our work, for multi-view gait recognition, we trained a two-stage network in which we initially identify the walking direction of a gait video using a 3D-CNN network.

Table 4.8: Comparison with other state-of-the-art methods on all three probe set of CASIA-B dataset in multi-view gait recognition. From the comparison, it is been observed that proposed two-stage network achieves higher average recognition rates in 8 of 11 different probe angles.

	Methods	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°
Normal	Dupuis [?]	97.2	99.6	97.2	96.3	98.8	98.4	97.1	97.6	97.14	93.0	96.0
	VI-MGR [?]	100.0	99.0	100.0	99.0	100.0	100.0	99.0	99.0	100.0	100.0	99.0
	Isaac [?]	98.5	99.0	99.0	97.0	97.5	96.0	95.0	97.5	94.0	93.9	99.0
	Proposed	100.0	100.0	100.0	100.0	100.0	98.4	100.0	100.0	100.0	98.4	96.8
Bag	Dupuis [?]	73.2	74.1	74.7	76.3	78.5	75.8	76.3	76.7	73.4	73.2	74.6
	VI-MGR [?]	93.0	89.0	89.0	90.0	77.0	80.0	82.0	84.0	92.0	93.0	89.0
	Isaac [?]	95.0	98.5	96.5	96.0	97.5	93.5	93.5	94.0	92.5	91.3	94.4
	Proposed	100	100	100	100	98.39	96.77	96.77	98.39	98.39	95.16	91.93
Coat	Dupuis [?]	81.64	87.39	86.29	84.34	89.96	91.86	89.50	85.04	72.24	78.40	82.70
	VI-MGR [?]	67.0	56.0	70.0	80.0	71.0	75.0	77.0	75.0	65.0	64.0	66.0
	Isaac [?]	97.0	99.5	97.5	94.0	88.0	90.5	89.5	94.5	92.0	91.3	94.0
	Proposed	81.52	82.11	83.58	85.48	84.46	83.72	83.28	84.16	83.58	80.65	78.45

Table 4.9: Correct walking direction identification rate (%) of proposed 3D-CNN network on all three probe set of CASIA-B dataset. The network achieved **100%** identification accuracy in all of the 11 view angles.

View angle	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°
Rate(%)	100	100	100	100	100	100	100	100	100	100	100

4.4.1 Comparison with the state-of-the-art methods on multi-view gait recognition

We tested the proposed 3D-CNN network with all three probe set of CASIA-B dataset and have achieved **100%** identification accuracy in all viewpoint angles proving the fact that our 3D-CNN is efficient in classifying walking direction from gait videos. Table 4.9 illustrates our test result.

To evaluate the performance of the proposed two-stage network, we compare it with the recent state-of-the-art multi-view gait recognition methods such as Dupuis *et al.* [**?**], Isaac *et al.* [**?**], and VI-MGR [**?**] on all three probe set of CASIA-B dataset. The comparison, as illustrated in Table 4.8 and Fig. 4.5, shows that the proposed method exceeds the previous best in result all three probe set by a significant margin. It outperforms other in **8** of 11 total probe angles.

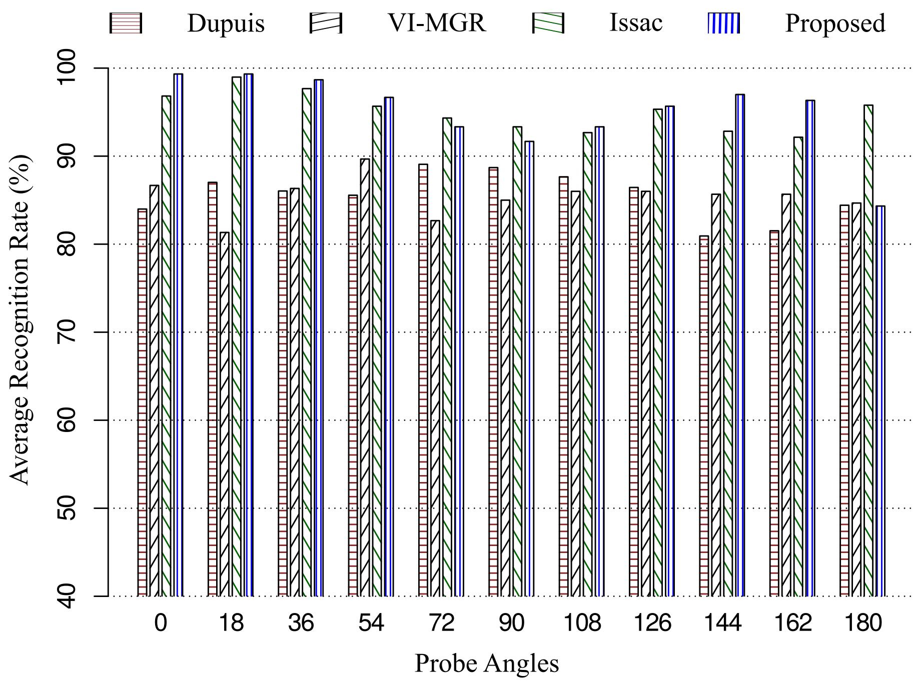


Figure 4.5: Average recognition rates(%) of the proposed method compared to the other state-of-the-art methods in multi-view gait recognition. Proposed method achieves higher average recognition accuracy on 8 of total 11 probe angles of CASIA-B dataset compared other methods in literature.

Chapter 5

Conclusion

5.1 Summary of Our Work

The main task of our thesis is to develop an effective algorithm for the Planted Motif Search problem. Our proposed algorithm qPMS-Sigma tries to find the (l, d)-motifs for n given sequences. It is an exact version of the PMS algorithm. qPMS-Sigma is based on the previous PMS algorithms qPMSPrune, qPMS7, TraverStringRef and PMS8. In our proposed algorithm we introduce clever techniques to compress the input sequences and thus space complexity is improved. We also include a faster comparison technique of the l-mers by adding bitwise comparison techniques. As we know the PMS is a NP-Hard problem, it is exponential in terms of time. So, parallel implementation techniques are proposed at the end of our work.

5.2 Future Prospects of Our Work

Though in our thesis we have given some ideas about parallel implementation of our algorithm, it is not implemented and tested on multiprocessor system. All comparison of our algorithm with the stated algorithms are done in single processor system. To understand the real speedup and slackness we need to experiment in a system involving a network of nodes having multiple processors. Our work is the extension of PMS8 codes which involves openMPI libraries. In future we want to continue our work using OpenMPI project which is a open source Message Passing Interface. Apart from the parallelism, more advanced pruning conditions can be used to reduce the search space. Different randomized approaches can also be included for improving the runtime, though it will make the exact version of our algorithm approximate.

References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans. on Pattern Anal. Mach. Intell.*, July 2019.

[2] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. on Signal Proces.*, vol. 45, no. 11, pp. 2673–2681, 1997.

[3] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. on Machine Learning*. Lille, France, 2015, pp. 448–456.

[4] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE Int. Conf. on Computer Vision*. Santiago, Chile, 2015, pp. 4489–4497.

[6] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *18th Int. Conf. on Pattern Recognition*. Hong Kong, China, 2006, pp. 441–444.

[7] I. Rida, N. Almaadeed, and S. Almaadeed, "Robust gait recognition: a comprehensive survey," *IET Biometrics*, vol. 8, no. 1, pp. 14–28, January 2019.

[8] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, February 2006.

[9] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *3rd Int. Conf. on Imaging for Crime Detection and Prevention*. London, UK, 2009.

[10] T. H. W. Lam, K. H. Cheung, and J. N. K. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recognition*, vol. 44, no. 4, pp. 973 – 987, April 2011.

[11] X. Huang and N. V. Boulgouris, "Gait recognition with shifted energy image and structural feature extraction," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2256 – 2268, April 2012.

[12] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: regarding gait as a set for cross-view gait recognition," in *The Thirty-Third AAAI Conference on Artificial Intelligence*. July, 2019, pp. 973 – 987.

[13] C. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognition*, vol. 37, no. 5, pp. 1057 – 1072, May 2004.

[14] G. Ariyanto and M. S. Nixon, "Model-based 3d gait biometrics," in *Int. Joint Conf. on Biometrics*. Washington DC, USA, 2011, pp. 1 – 7.

[15] F. Tafazzoli and R. Safabakhsh, "Model-based human gait recognition using leg and arm movements," *Engineering Appl. of Art. Intell.*, vol. 23, no. 8, pp. 1237 – 1246, December 2010.

[16] Y. Feng, Y. Li, and J. Luo, "Learning effective gait features using lstm," in *23rd Int. Conf. on Pattern Recognition*. Cancun, Mexico, 2016, pp. 325–330.

[17] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep cnns," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209 – 226, February 2017.

[18] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," in *Int. Conf. on Biometrics (ICB)*. Halmstad, Sweden, 2016.

[19] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3d convolutional neural networks," in *IEEE Int. Conf. on Image Processing*. Phoenix, AZ, USA, 2016, pp. 4165–4169.

[20] C. Zhang, W. Liu, H. Ma, and H. Fu, "Siamese neural network based gait recognition for human identification," in *IEEE Int. Conf. On Acoustics, Speech and Signal Processing*. Shanghai, China, 2016, pp. 2832 – 2836.

[21] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh, "Gaitgan: invariant gait feature extraction using generative adversarial networks," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*. Honolulu, HI, USA, 2017, pp. 532 – 539.

[22] S. Yu, R. Liao, W. An, H. Chen, E. B. Garcia, and a. Huang, Y, "Gaitganv2: Invariant gait feature extraction using generative adversarial networks," *Pattern Recognition*, vol. 87, no. 11, pp. 179 – 189, March 2019.

[23] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations," in *Chinese Conf. on Biometric Recognition*, 2017, pp. 474 – 483.

[24] R. Liao, S. Yu, W. An, H. Chen, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, February 2019.

[25] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," vol. 1, no. 2, 2017, pp. 4263 – 4270.

[26] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*. Boston, MA, USA, 2015, pp. 1110 –1118.

[27] D. Cunado, M. S. Nixon, and J. N. Carter, "Using gait as a biometric, via phase-weighted magnitude spectra," in *Int. Conf. on Audio-and Video-Based Biometric Person Authentication*. Berlin, Heidelberg, 1997, pp. 93–102.

[28] L. Wang, H. Ning, T. Tan, and W. Hu, "Fusion of static and dynamic body biometrics for gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 149–158, March 2004.

[29] R. Araujo, G. Graa, and V. Andersson, "Towards skeleton biometric identification using the microsoft kinect sensor," in *ACM Symposium on Applied Computing*. Coimbra, Portugal, 2013, pp. 21–26.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd Int. Conf. on Learning Representations*. San, Diego, 2015.

[31] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conf. on Computer Vision*, 2016, pp. 499 – 515.

[32] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition*. Columbus, USA, 2014, pp. 1725 – 1732.

Appendix A

Codes

A.1 Compressed String

Here is a sample C++ structure for compressing a 1000 character long DNA string

```
//letterToInt[i] means a unique id for each character of the
    alphabet;
//Here,
//letterToInt['A']=0;
//letterToInt['B']=1;
//letterToInt['C']=2;
//letterToInt['D']=3;

struct CompressedlMer
{
    unsigned int **cs;//[2][32];
    int l;
    CompressedlMer(char *string, int start, int l)    {//
       constructor
        int group = 0;
        int intPerGroup = ceil(l/sizeof(unsigned int));
        int temp = l;
        while(temp)    {
            group++;
            temp = temp/2;
        }
        cs = new unsigned int*[group];
```

```
        for(int i = 0; i < group; i++)
            cs[i] = new unsigned int[intPerGroup];

        for(int i = 0; i < l; i++)    {
            int j = letterToInt[string[start+i]];
            int positionMask = 1<<(i%sizeof(unsigned int));
            int intNo = i/sizeof(unsigned int);
            int grp = 0;
            while(j>0)    {
                if(j&1) cs[grp][intNo] = cs[grp][intNo] | positionMask;
                j = j/2;
                grp++;
            }
        }
    }

    ~CompressedlMer()    {//destructor
        for(int i = 0,j=1; j < l; i++,j*=2)    {
            delete [] cs[i];
        }
        delete [] cs;
    }
};

int hammingDist(const CompressedlMer &sa, const CompressedlMer &sb)    {
    int intPerGroup = ceil(sa.l/sizeof(unsigned int));
    int hamDist = 0;
    for(int i = 0; i < intPerGroup; i++)    {
        unsigned int flag = ~(0);    //all bits are set to 1
        for(int group = 0, j = 1; j < sa.l; group++, j*=2)    {
            flag = flag | (sa.cs[group][i]^sb.cs[group][i]);
        }
        hamDist = hamDist + __builtin_popcount(flag);
    }
    return hamDist;
}
```

Generated using Undegraduate Thesis LaTeX Template, Version 1.2. Institute of Information and Communication Technology, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.

This thesis was generated on Sunday 19th January, 2020 at 9:32pm.