REDUCING COVARIATE FACTORS OF GAIT RECOGNITION USING FEATURE SELECTION, DICTIONARY-BASED SPARSE CODING, AND DEEP LEARNING

Munif Alotaibi

Under the Supervision of Dr. Ausif Mahmood

DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

AND ENGINEERING

THE SCHOOL OF ENGINEERING

UNIVERSITY OF BRIDGEPORT

CONNECTICUT

May, 2017

REDUCING COVARIATE FACTORS OF GAIT RECOGNITION USING FEATURE SELECTION, DICTIONARY-BASED SPARSE CODING, AND DEEP LEARNING

REDUCING COVARIATE FACTORS OF GAIT RECOGNITION USING FEATURE SELECTION, DICTIONARY-BASED SPARSE CODING, AND DEEP LEARNING

Munif Alotaibi

Under the Supervision of Dr. Ausif Mahmood

Approvals

Committee Members

Name	Signature	Date
Dr. Ausif Mahmood	Ausif Mahmood	5-10-2017
Dr. Miad Faezipour	Miad Faezipour	May 10, 2017
Dr. Navarun Gupta		May 9, 2017
Dr. Xingguo Xiong	Xingguo Xiong	May, 9th, 2017
Dr. Saeid Moslehpour	Moslehpour	may 3, 2017

Ph.D. Program Coordinator

Dr. Khaled M. Elleithy		May 12, 2017

Chairman, Computer Science and Engineering Department

Dr. Ausif Mahmood	Ausif Mahmood	5-10-2017

Dean, School of Engineering

Dr. Tarek M. Sobh	T. Sobh	5/15/2017

Abstract

Human gait recognition is a behavioral biometrics method that aims to determine the identity of individuals through the manner and style of their distinctive walk. It is still a very challenging problem because natural human gait is affected by many covariate conditions such as changes in the clothing, variations in viewing angle, and changes in carrying condition. Although existing gait recognition methods perform well under a controlled environment where the gait is in normal condition with no covariate factors, the performance drastically decreases in practical conditions where it is susceptible to many covariate factors. In the first section of this dissertation, we analyze the most important features of gait under the carrying and clothing conditions. We find that the intra-class variations of the features that remain static during the gait cycle affect the recognition accuracy adversely. Thus, we introduce an effective and robust feature selection method based on the Gait Energy Image. The new gait representation is less sensitive to these covariate factors. We also propose an augmentation technique to overcome some of the problems associated with the intra-class gait fluctuations, as well as if the amount of the training data is relatively small. Finally, we use dictionary learning with sparse coding and Linear Discriminant Analysis (LDA) to seek the best discriminative data representation before feeding it to the Nearest Centroid classifier. When our method is applied on the large CASIA-B and OU-ISIR-B gait data sets, we are able to outperform existing gait methods.

In addition, we propose a different method using deep learning to cope with a large number of covariate factors. We solve various gait recognition problems that assume the training data consist of diverse covariate conditions. Recently, machine learning based techniques have produced promising results for challenging classification problems. Since a deep convolutional neural network (CNN) is one of the most advanced machine learning techniques with the ability to approximate complex non-linear functions, we develop a specialized deep CNN architecture for gait recognition. The proposed architecture is less sensitive to several cases of the common variations and occlusions that affect and degrade gait recognition performance. It can also handle relatively small data sets without using any augmentation or fine-tuning techniques. Our specialized deep CNN model outperforms the existing gait recognition techniques when tested on the CASIA-B large gait dataset.

ACKNOWLEDGEMENTS

My thanks are wholly devoted to God who has helped me all the way to complete this work successfully. I owe a debt of gratitude to my family for understanding and encouragement.

I would like to express special thanks to my supervisor Dr. Ausif Mahmood for his constant guidance, comments, and valuable time. Without his support, I would not have been able to finish this work.

My special appreciation goes also to Dr. Khaled Elleithy for his feedback and support and to my committee members Dr. Miad Faezipour, Dr. Navarun Gupta, Dr. Xingguo Xiong, and Dr. Saeid Moslehpour for their valuable time and suggestions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Biometrics refers to the use of the intrinsic physical or behavioral traits in order to identify humans. Human gait is known as a repetitive interaction between various patterns of human body parts. Consequently, gait recognition is the identification of human beings purely based on the distinctive style and manner of their walk. It is categorized as one of the future generation recognition technologies. Compared to the first generation of biometric methods (such as face recognition and fingerprinting), gait recognition has several advantages. For example, it can be done without consent, collaboration, or body contact from the target at low resolution and from a far distance. It is not easy to obscure or conceal someone's gait. Usually, criminals wear face masks, dark sun-glasses and gloves to invalidate face, eyes, and finger print recognition. In such scenarios, gait recognition is the only effective and useful identification method. However, some factors like posture, changes in the physical shape of the body, speed of walking, stimulants, walking surface, and the conditions and psychology of an individual can affect the gait. These factors make the gait recognition a challenging problem. Some of the above mentioned influential factors can even cause limitations to the human visual system that identifies known individuals from others. Also, it is worth mentioning that these limitations are common with other biometric methods such as face recognition.

1.1 Problem and scope

Many gait recognition techniques have been presented in the past decade. However, despite the wide variety of approaches, the ability to automatically recognize a person accurately and reliably does not meet the needs of the practical applications. One of the most challenging difficulties that face many gait recognition methods is the intra-class variations caused by covariate conditions that affect the gait adversely. Examples of these covariate conditions that commonly occur in live situations are changes in clothing condition, changes in carrying conditions (such as carrying a bag or a briefcase) and viewing angle variations. These covariate conditions create problems for practical gait recognition systems and significantly deteriorate their performance. They do not only alter and occlude the appearance of the body shape, but also affect the dynamic pattern of body movements. Whenever there are occluded pixels as a result of carrying a bag or wearing a baggy coat, it is normal for the accuracy of the recognition to be decreased. In addition, these covariates are very difficult to be automatically detected for removal or to be fully covered in training so that they may be part of matching process. For example, there are many types of bags and clothes, which can take many different forms, and can be carried or worn in different ways. These covariate factors interfere with body shape, causing pixel-related confusion between the motion of the covariate factor and the motion of the gait. Thus, it is difficult to accurately capture the style and motion of the gait under these conditions.

1.2 Motivation behind the research

Gait recognition is motivated by the fact that humans have the capability of recognizing one another from even displays of gait that are not clear. In addition, it is motivated by the study of biological motion, which proves that the human visual system can easily extract information about others from their walking style, such as gender, age, emotion, and weight, which leads to the ability to determine their identity.

1.3 Contributions

Recently, several feature selection techniques have been brought into gait recognition and have shown promising results. These stimulated us to investigate how the feature selection can be represented in an efficient way. Thus, we introduce a new feature selection technique that boosts the performance of gait recognition. Our proposed method concentrates on the features that are less affected by the changes in the covariate conditions and have important discriminant power. This is accomplished by developing an understanding of what are the most important features in the gait representation. We also use dictionary learning with sparse coding and Linear Discriminant Analysis (LDA) to find the best discriminative data representation before feeding it to the Nearest Centroid classifier. Later, we demonstrate that by using the proposed robust method, we can outperform the existing gait recognition methods since our approach is less sensitive to these covariate factors.

Also, one of the popular deep learning methods is based on Convolution Neural Networks (CNN). The Deep CNN is an advanced machine learning technique that has inspired many researchers because it has achieved the state of art results in several applications of pattern recognitions. Thus, we propose a gait recognition approach based on a specific CNN architecture in order to approximate complex non-linear functions from high-dimensional images. Typically, the convolutional neural networks consist of two different types of layers, which are subsampling and convolutional layers. Every layer of these consists of multiple feature maps. These feature maps can be subsampling or a convolutional feature map, depending on their type of the layer. In the subsampling layer, the feature map would usually do either maximum or average subsampling in order to perform down sampling on the data. In the convolutional layer, the feature map will convolve the image with specific filter (weights). Then, these layers are followed by the fully connected layers which are similar to the layers in a standard multilayer neural network. It is worth mentioning that the CNN is designed to recognize the visual patterns straightly from images with minimal preprocessing.

Typically, the convolutional neural networks consist of two different layers, i.e., convolution followed by subsampling. The output from a convolution layer or subsampling is referred to as a feature map. The feature map can thus be a convolutional or subsampling feature map. Each layer contains several feature maps. The sub-sampling layers would usually do either maximum or average subsampling in order to perform down sampling on the feature maps. In the convolutional layer, the feature maps are produced by convolving the image with specific learned filters (weights). In a deep convolution network, many of these layers are combined in a cascaded manner. The last stage is fed to a fully-connected neural network layer which performs the final classification. It is worth mentioning that the CNN is designed to recognize the visual patterns from images with minimal preprocessing.

In this dissertation, we develop the appropriate architecture for deep CNNs for gait recognition. Particularly, we examine how many fully-connected, subsampling, and convolutional layers are needed. Further, the optimal number of feature maps per layer, the optimal sizes of the feature maps, and the best type of input feature to be used with CNNs for gait recognition are determined empirically. Finally, we investigate the strength and effectiveness of our proposed work by applying our CNN model on the CASIA-B gait database and compare the results with the existing methods in gait recognition field.

CHAPTER 2: LITERATURE SURVEY

There have been several gait techniques proposed to eliminate the effect of the covariate conditions on performance. In general, these methods of gait recognition can be categorized into two types: model-based approaches and model-free approaches.

2.1 Model-based approaches

The model-based approaches normally concentrate on recovering the physical structures of the moving human body by using some extracted static parameters. Factors including stride length, stride speed, cadence, and size ratio of various body parts, such as hands and feet, are used to develop unique features for every individual. Some works also use dynamic features like joint trajectories. The model-based approaches, although known to be view and scale invariant, are still difficult to estimate accurately and are computationally expensive. They also require high-quality resolution silhouettes.

One of the early model-free approaches is proposed by Cunado, Nixon, et al. [1]. The proposed method extracts one of the moving legs from the gait video sequences. They then use the Fourier series to extract the gait features from the motion of the human leg and fed them to K-Nearest Neighbor for the classification.

Similarly, Yam, Nixon, et al. [2] extracted thigh and leg rotation from the gait videos filmed under different speeds (walking and running). They use the Fourier series to extract the gait features and the K-Nearest Neighbor is used for classification.

Yoo and Nixon [3] introduced the stick figure representation for gait recognition. In this model, there are six joints and eight sticks extracted from the human body to form the figure representation. Different static features are calculated from this figure. Then, the K-Nearest Neighbor is used for the classification.

After dividing the silhouette into three regions and labeling the parts of the body, Bobick and Johnson [4] calculate the distances between four body part locations (pelvis, left and right foot, head) to extract some static body parameters for gait recognition.

Zhang et al. [5] fit a five-link biped model to extract trajectory-based kinematic components. The extracted frequency components from these joint trajectories are fed to The Hidden Markov Models for classification.

Tafazzoli and Safabakhsh [6] also use the Fourier analysis to describe the motion patterns of a leg and arm. The k-Nearest Neighbor is then used for classification.

Kim and Paik [7] proposed a hierarchical active shape model that estimates and tracks the positions of several landmark points from the silhouette. Then, several static parameters are extracted as features to measure the similarities between the two gaits. Wang , Tan et al. [8] proposed a gait method based on some statistical descriptions of the human body.

2.2 Model-free approaches

The model-free approaches, which are sometimes referred to as appearance-based or holistic approaches, focus on motion information and body shape. They function directly on the extracted gait features that are represented in the silhouette.

The general framework of the model-free approaches usually involves several pre-processing steps such as background-foreground subtraction, alignment and normalization, feature extraction, and classification. Contrary to the model-based approaches, the model-free approaches do not require high-resolution images and are less computationally expensive. Hence, they are well proper for practical applications.

2.2.1 Gait representations

In an attempt to overcome the challenges of matching Sequences on a frame-by-frame basis, several gait representations have been proposed in the past. For example, Liu and Sarkar [9] proposed a straightforward and popular spatio-temporal gait representation, which is the average of silhouettes across one gait cycle. In their proposed work, similarity between a given probe (testing) sequence and the gallery (training) sequence is computed using the Euclidean distance. The averaged silhouette representation has the ability to describe the motion of human gait in one single image. It also can preserve both spatial and temporal information. Similarly, Han and Bhanu [10] proposed the gait energy image (GEI). GEI is also the normalized, averaged silhouette of one gait cycle. Later, Chunli and KeJun [11] proposed the Enhanced GEI, which is the subtraction between the GEI and the frames that used to compute the GEI. Then, 2DPCA is used to reduce the dimensionality of the representation.

Wang et al. [12] introduced a gait template, called Chrono-Gait Image (CGI). The CGI is generated by computing the contours that represent one gait cycle and encoding them using a multi-channel mapping function into one template. The final template is a single, colored image.

Liu et al. [13] introduced new model that combines both GEI and CGI representations. They compute the Histogram of Oriented Gradients (HOG) of both the GEI and the CGI templates to extract multiple HOG templates. Then, the HOG templates are used for classification.

Chen et al. [14] proposed frame difference energy image (FDEI) to overcome the problem of noise and incomplete frames. It is calculated by dividing the gait cycle into states, and then using the denoised averaged frame of each state to construct the FDEI representation. Roy et al. [15] introduced a new gait template over one gait cycle called Pose Energy Image (PEI). Frames of one gait are divided into a collection of key poses, then each is averaged to generate the gait features. Zhang et al. [16] proposed the active energy image (AEI). It is the averaged silhouette of the difference between the frames that represent one gait cycle. AEI focused more on the active regions of the silhouette. Given several sequences f1, f2,.., f_t, the difference between these frames is calculated as follows: $DF_t(x,y) = ||f_t(x,y) - f_{t-1}(x,y)||$. Then, the averaged result is calculated in equation 2.1:

$$AEI(x,y) = 1/k \sum_{t=1}^{k} DF(x,y,t) \tag{2.1}$$

After analyzing the gait variances using the Analysis of Variance (ANOVA) on three databases separately, Veres, Gordon et al. [17] concluded that the most important features in the averaged silhouettes are around the head and upper body, and there is little contribution from the dynamic part. They also concluded that both the parts of the body that remain static and the kinematics are required for more accurate classification.

Bashir, Tao et al. [18] proposed another gait representation named Gait Entropy Image (GEnI). It is generated by computing Shannon entropy for each pixel of the silhouettes that represent one gait cycle. Mathematically, it is calculated in equation 2.2:

$$GEnI(x,y) = -q \ \log_2 q - (1-q)\log_2(1-q) \tag{2.2}$$

where the q is the GEI value at the coordinate (x,y).

Lee, Tan et al. [19] introduced the Gait Probability Image (GPI) by calculating the binomial distribution of each pixel during the gait cycle.

Li, Xu et al. [20] introduced the Body-Part Segmentation method. The GEI is divided into six parts and the relevant silhouette parts are selected based on the number of foreground pixels, which should not exceed some intervals (two threshold values). These selected parts are then used in the classification process using the Nearest Neighbor classifier.

Another gait representation over one complete gait cycle used Pal and Pal Entropy. It was proposed by Jeevan, Jain et al. [21]. It is followed by the Principal Component Analysis (PCA) to reduce the dimensionality of the representation and Support Vector Machine (SVM) for classification.

Chai, et al. [22] proposed a method that based on the Perceptual Curve. The Perceptual Shape Descriptor of the silhouette is computed, and then the Neighbor classifier (NN) and the K-Nearest Neighbor classifier (K-NN) are used for classification.

Liu and Sarkar [23] have presented a method for gait recognition, by using population Hidden Markov Models (pHMMs) to come up with a dynamics-normalized gait signature for every gait cycle. They map any gait sequence stance onto one of the pHMM states. The stances of each pHMM state are then averaged to arrive at one, normalized signature. Then, they highlight the variances in stance shapes between different individuals and minimize the variations in stance for the same individual by using Linear Discriminant Analysis Space and Principal Component Analysis. The paper concluded that body-stance shape plays a more essential role than dynamics in gait recognition.

Huang and Boulgouris [24] proposed the Shifted Energy Image (SEI) and extracted another representation named Gait Structural Profile (GSP). Both SEI and GSP are combined to achieve a good performance where LDA is used for dimensionality reduction.

Guha and Ward [25] present the Differential Radon Transform (DiffRT), derived from the standard Radon Transform to capture the high-frequency components from the averaged silhouette. The extracted coefficients of the DiffRT are then used for the classification. Guan and Li et al. [26]

proposed a model based on the classifier ensemble model, with the majority voting to reduce the effect of covariate factors.

Whytock et al. [27] proposed three different gait representations named Gait Variance Image, Skeleton Energy Image, and Skeleton Variance Image. These representations aim to reduce the effect of covariate factors. They are generated via different steps, which consist of using the screened Poisson equation, boundary perturbation, and image kernels.

Some gait approaches combined both model-free and model-based approaches. For example, [28] proposed the spatiotemporal shape and dynamic motion (STS-DM), which is a model that is robust against many gait variations. The model consists of three phases: the use of the Fourier Descriptor Analysis, fitting ellipses to five contour segments, and the use of Dynamic Time Warping to analyze the rotations of body parts. However, the recognition performance of the model is sensitive to segmentation shortcoming and depends on the preprocessing step. In a slightly similar way, [29] proposed a model that uses the Procrustes Shape Analysis (PSA) and the Elliptic Fourier Descriptors (EFDs), which are then applied to silhouette contours. The classification results by PSA and EFDs are combined to improve the recognition accuracy.

2.2.2 Optical Flow

There are several methods that use the texture features of the optical flow of gait. For example, by using the silhouettes that represent one walking cycle, the optical flow fields for every pixel are computed to form new gait representation that can capture both the motion direction and motion intensity [30]. Likewise, Lam, Cheung et al. [31] proposed a different gait representation called a gait flow image (GFI) by using an optical flow field for gait recognition. GFI is the optical flow lengths observed over a complete gait cycle of the silhouette contour. Instead of using the silhouette, the Local Binary Pattern (LBP) [32] is employed to describe the texture features of the optical flow of gait. It is then followed by the hidden Markov model (HMM), which is used for classification [33]. Also, Kusakunniran [34] has proposed a framework that does not rely directly on the silhouettes. It relies on Space-Time Interest Points (STIPs), which signify the movements of the body as detected from video frames. Then both the Histograms of Oriented Gradient (HOG) and the Histogram of Optical Flow (HOF) are employed to calculate a descriptor for each STIP. After that, the Bag-of-Words method is applied on every set of STIP descriptors in order to generate the gait features. Finally, the K-Nearest Neighbor (KNN) is used to classify the data.

2.2.3 Feature selection

Recently, several feature selection techniques have been brought into gait recognition and have shown promising results. For example, Bashir et al. [35] proposed a feature selection mask to be applied on the training and testing data. In his method, the GEI is segmented into upper and lower parts and the pixel thresholds are used to reduce the effects of the covariate conditions. This approach has yielded good recognition results.

Based on the Random Forest feature rank algorithm, Dupuis et al. [36] also proposed a new feature selection method on the GEI that focuses on the upper and lower regions.

Rida et al. [37] proposed a new feature selection based on the group Lasso for the horizontal motion using GEI as the gait representation.

Whytock et al. [38] proposed three different, bolt-on modules to detect and remove the covariate factors using different gait representations. Covariate factors are detected by subtracting the training and testing data. Then, different removal techniques based on the pixel thresholds are used to remove the detected covariate factors.

Iwashita et al. [39] also proposed a different method that divided the GEI into several regions. The matching weight for each region is computed based on the comparison between the regions of the probe and the regions of the gallery. Then, the regions with high matching weights (less affected by covariate factors) are used in the classification process.

Likewise, Islam et al. [40] also proposed a different method to detect covariate factors by subtracting the average of all GEIs with these factors in the dataset from the averaged GEIs without these covariate factors. Then, some pixel points from the boundary of the Covariate factor are used to detect the presence of the covariate factors. After that, the GEI is divided into seven parts, and the parts that contain the covariate factor are discarded in classification.

Shaikh et al. [41] presented a new method of gait recognition using only a portion of the gait silhouette by eliminating the redundant information from the silhouette. They focused on one moving part of the body, which is the swinging hands, to extract the gait signature. Both Principal Component Analysis (PCA), followed by the Multiple Discriminant Analysis (MDA), are used to reduce the dimensionality. Then, the minimum distance classifier based on Euclidean distance is used for classification.

2.2.4 Subspace learning methods

The subspace learning approaches are the most widely-used methods for gait recognition. There have been many subspace learning approaches that started to consider learning the features from an object by considering the representation of higher-order tensors.

For example, Xu et al. [42] introduced the Concurrent Subspaces Analysis (CSA), which can extract the features directly from the 2D data. Yan et al. [43] proposed the Discriminant Analysis with Tensor Representation (DATER) as an alternative to LDA. Xu et al. [44] used both CSA and DATER on GEIs for gait recognition.

Tao et al. [45] proposed a general tensor discriminant analysis algorithm to preserve discriminative features, where the input data is obtained by convolving the Gabor functions with the

GEI.

The matrix-based sparse bilinear discriminant analysis (SBDA) is proposed by Lai et al. [46] as a sparse learning method effective for gait recognition and used with on GEIs. It is derived from both the matrix dimensionality reduction algorithm and the sparse subspace algorithm.

Also, Locality Preserving Projections(LPP) [47] and Local Fisher Discriminant Analysis (LEFDA) [48] are employed to generate the gait features as in [49].

As an alternative to calculating the distance between the probe image and the gallery image and then classifying it using the Nearest Neighbor classifier, Huang, et al. [50] proposed the integer programming problem to compute the image-to-class distance. Each class consists of several GEIs of the same subject.

2.3 Cross-view gait recognition

The aim of the cross-view gait methods is to reduce the influence of viewpoint variations. The cross-view gait methods can be categorized into three types: features-invariant model, 3D construction model, and View Transformation Model (VTM).

In the features-invariant model, one-view gait is to be estimated from any other view gait, usually based on handcrafted features. However, if the view angle change is large, the accuracy will decrease dramatically. Thus, it is difficult to recognize gait under literal view from the frontal view. The second type which is a 3D construction model [51] [52], requires a cooperating, multi-camera setup, and is very costly.

The third type, which is the View Transformation Model, is where one-view gait is estimated via the projection of a training set that consists of multi-view gaits. It projects a large, dimensional, multi-view gait data into a lower-dimensional feature space that has sufficient discriminative capability. The performance of these methods enhances as more training sequences are given from multiple views.

Early work of this type was proposed by Bashir et al. [53], where the Canonical Correlation Analysis (CCA) is used to find the correlation of multi-view gaits.

Lee and Elgammal [54] used the multilinear generative model to compute the gait into aligned gait cycles. Then, they used the higher-order singular value decomposition to learn the features of view-invariant gait.

Liu et al. [55] proposed the multi-view gait recognition method to extract a linear subspace of many sequences that are taken from different viewpoints. In this model, Multiview Subspace Representation is introduced and used to extract the bases of the linear subspace and to handle the intra-class variation. Then, a new learning-based method named Marginal Canonical Correlation Analysis (MCCA) is introduced. MCCA has a better projection that can find the discriminative information and maximize the interclass variations. Finally, the Nearest Neighbor is used for the classification.

Kusakunniran [56] proposed View Transformation Model based on a sparse regression process to find the correlation of multi-view gaits.

Hu et al. [49] presented a View-Invariant Discriminative Projection (ViDP) technique to compute the unitary projection for gait features from multiple views. The result of ViDP method was encouraging compared to other multi-view gait methods.

Most recently, Wu et al. [57] used a CNN with several different architectures for cross-view gait recognition. Yu et al. [58] used the auto-encoder to find the invariant gait features, and then used both the PCA and the K-Nearest Neighbor (KNN) to classify the data.

Zeng and Wang [59] introduced a new method, for gait recognition via the deterministic learning theory using a Radial Basis Function (RBF) neural network to eliminate the effects of changes in angle view. In this method, gait features that reflect the view variations were extracted.

The extracted gait feature depends on the width of the outer contour and silhouette area. The gait silhouette was divided into four equal sub-regions from top to bottom; the holistic width features of these sub-regions are calculated, and combined to derive gait signatures. Then, the RBF neural networks are used to approximate and classify the unknown gait. However, this method is not effective against clothing and carrying condition changes, since the extracted features heavily depend on width of the outer contour and silhouette area.

2.4 Deep Convolutional Neural Networks

CNN has been proven to be an efficient technique in several fields of pattern recognition. For example, GoogLeNet, which was proposed by Szegedy et al. [60], is an efficient deep neural network architecture for Large-Scale Visual Recognition. It is a large network that consists of around 27 layers, uses max and average pooling, a dropout method, and a soft-max classifier.

Ijjina et al. [61] proposed a method to recognize periodic human actions by using a 2D image of the height of a person's feet above the ground as features and an input to the CNN. They used a CNN architecture consisting of two subsampling layers and two convolution layers, followed by the fully-connected layers. Each of the first two layers has ten feature maps while each of the last two layers consists of 20 feature maps.

A large and deep CNN that has five convolutional layers and three fully-connected layers was introduced by Krizhevsky et al. [62]. The network was trained using 1.2 million high-resolution images from the ImageNet dataset, where there are 1000 different classes. The proposed network achieved a promising result.

Karpathy et al. [63] have proposed the Multiresolution, CNN architecture that aims to speed up the training time, and is suitable for largescale video classifications. One million YouTube videos belonging to 487 classes of sports were used to train this CNN network. The Multiresolution has two separate streams of processing: a context stream that functions on a low-resolution image, and a fovea stream that functions on the middle portion of the high-resolution frame. The result of the Multiresolution was encouraging.

2.5 Existing gait databases

In this section, the existing gait databases used in gait recognition are analyzed and compared.

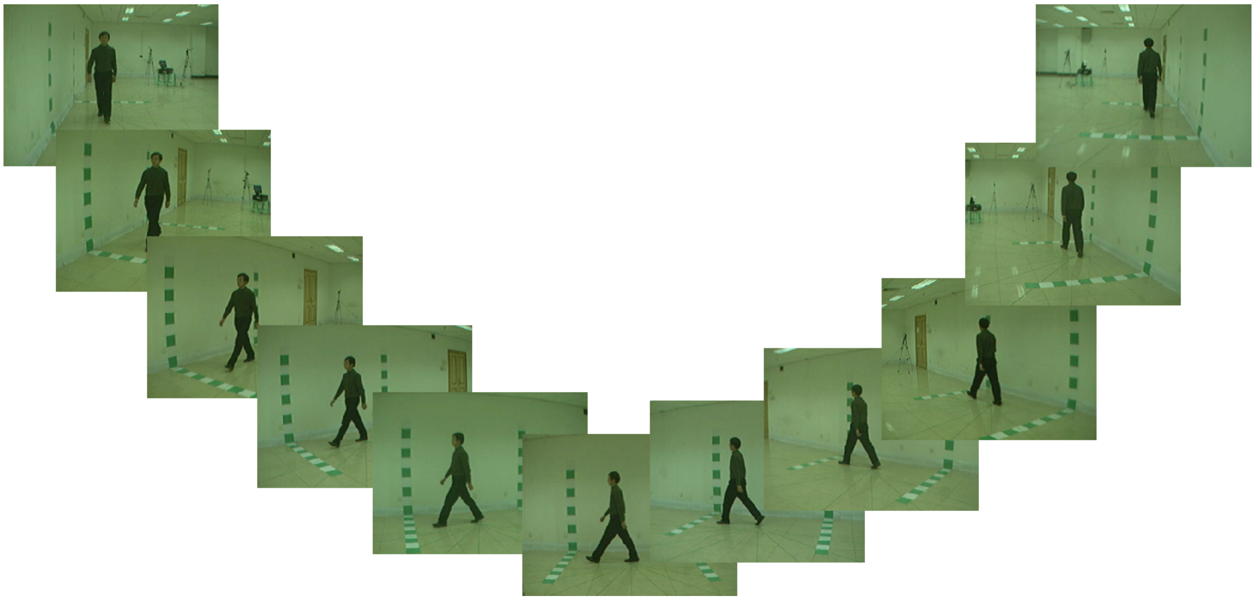


Figure 2.1: CASIA-B data set[2]

2.5.1 NLPR gait database

The Chinese National Laboratory of Pattern Recognition presented three public data-sets namely: CASIA-A [65], CASIA-B [64] and CASIA-C [66].

CASIA- A covers the gait data of 20 subjects. Each subject has 12 image sequences, captured from three different directions: front view, side-view, and 45° view.

[2] Image source: CASIA Gait Database, The Center for Biometrics and Security Research (CBSR), image published in [64]

CASIA- B 2.1 is a large multi-view gait dataset that covers the data for 124 subjects. The gait data is captured from 11 viewing angles in different carrying conditions (bg), normal walking conditions (nm), and clothing conditions (cl). CASIA-B will be described in more details in the next section.

CASIA- C is recorded by an infrared (thermal) camera outdoors at night. It covers the gait data for 153 subjects under different speeds and carrying conditions.

2.5.2 USF database

The USF [67] is a large gait dataset that covers the data of 122 subjects. The gait sequences are filmed outdoors under several walking variations: two viewpoints, surface, shoes, carrying condition, and time.

2.5.3 Southampton database

The Southampton [68] has two gait datasets. The Soton Small dataset covers the data of 12 subjects, captured inside track, with a chroma-key green screen backdrop, under several walking variations, which are: footwear, clothes and carrying bags, and different speeds.

The Soton Large database covers the data of 115 subjects, captured outdoors on an inside track and on a treadmill under six different views.

2.5.4 OU-ISIR

The OU-ISIR [69] has several gait datasets. OU-ISIR-A covers the gait data of 12 subjects. It is captured under nine different speeds.

OU-ISIR-B covers the gait data of 68 subjects. It is captured indoors under different clothing conditions.

OU-ISIR-D covers the gait data of 185 subjects from a side view with different gait fluctuations within the gate cycle period. “OU-ISIR Gait Large Population Dataset” covers gait data of 4007 subjects. It is captured under four different viewing angles (namely: 55°, 65°, 75 °, 85°).

2.5.5 Other gait databases

There are other gait datasets including the HID-UMD gait databases [70], the MIT Artificial Intelligence Lab data set (MIT AI) [71], and the CMU Mobo data-set [72]. More details about these existing data-sets are shown in Table 2.1.

Dataset	**Number of Subjects**	**Number of Sequences**	**Variations**	**Scene**
CASIA A	20	240	3 viewing angles	Outdoor
CASIA B	124	13,640	clothing and carrying conditions, 11 viewing angles	Indoor
CASIA C	153	1,530	Speed, carrying condition	Outdoor, thermal camera (at night)
USF Gait	122	1,870	2 viewpoints, surface, footwear, carrying condition, time	Outdoor
OU-ISIR A	34	-	9 speeds	Indoor (Treadmill)
OU-ISIR B	68	2746	clothing condition	Indoor(Treadmill)
OU-ISIR D	185	-	gait fluctuations	Indoor(Treadmill)
OU-ISIR Gait Large Population Dataset	4007	-	Different age ranges, 4 viewing angles	Indoor
Soton Small	12	-	Carrying condition, clothing, footwear (shoe), viewing angles	Indoor
Soton Large	115	2,128	6 different viewing angles	Indoor (Treadmill and track), outdoor
HID-UMD 1	25	100	4 viewing angles	Outdoor
HID-UMD 2	55	220	2 viewing angles	Outdoor
CMU-Mobo	25	600	surface, 6 viewing angles, speed, carrying condition	Indoor(Treadmill)
MIT AI	24	194	View, time	Indoor

Table 2.1: The existing gait datasets

CHAPTER 3: DATA SET AND GAIT REPRESENTATION PLAN

In this paper, we used the CASIA-B dataset and OU-ISIR Gait Database- Treadmill Dataset B. We have conducted several different experiments on CASIA-B dataset set using our proposed methods. Also, it should be noted that our work falls into the category of model-free approaches.

3.1 Data sets

The CASIA-B database [64] is the most recognized in gait recognition research and has various covariate conditions. It is used here to evaluate and compare our methods with others. The CASIA-B is a large, multi-view gait database covering the data of 124 subjects. Every subject has ten sequences. Among these ten sequences, six sequences are taken under normal condition (NM), two sequences are taken under a carrying condition (BG), and two sequences are taken under a clothing condition (CL). Each of these ten sequences is captured from 11 different viewing-angles with a step of 18°, namely: 0°, 18°, 36°, 54°, 72°, 90°, 108°, 126°, 144°, 162°, and 180°. Some samples of gait silhouettes from CASIA-B data set are shown in Figure 3.2.

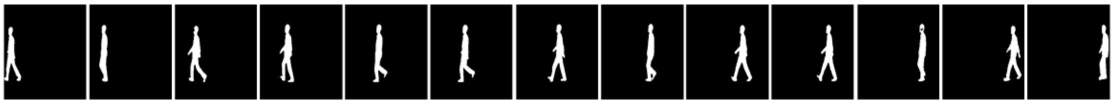


Figure 3.2: Samples of silhouettes from the CASIA-B data set

We also used the OU-ISIR Gait Database- Treadmill Dataset B [69]. The OU-ISIR-B covers the gait data of 68 subjects. It is captured indoors with different combinations of clothes (such as different short and long pants, shirts, hats or caps, skirts, parkas, scarves, different coats, and jackets) and was taken from one side view. Some samples of silhouettes from OU-ISIR Gait Database- Treadmill Dataset B are shown in Figure 3.3.

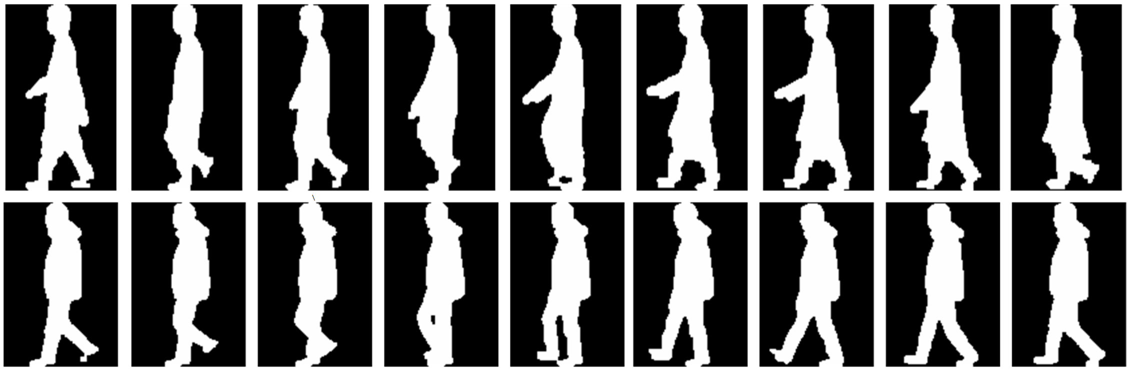


Figure 3.3: Samples of silhouettes from the OU-ISIR Gait Database- Treadmill Dataset B

3.2 Gait representation

Our feature selection method is designed for use on GEI, a very popular 2D gait representation. We selected the GEI since it has proven more effective when compared to other gait representations. A recent experimental study by Iwama et al. [73] demonstrated that GEI is more effective representation for gait recognition on their proposed large gait database under normal conditions when compared to other gait representations. However, the GEIs are still sensitive to covariate conditions. The GEI is generated by the following Equation 3.3:

$$GEI(x,y) = 1/k \sum_{t=1}^{k} Fr(x,y,t) \tag{3.3}$$

where k is the total number of the silhouettes Fr that represent one complete gait cycle (left-to-left heel striking the ground). (x,y) are the coordinates of the silhouette, and t is the silhouette index. Several GEI from the benchmark CASIA-B [64] [74] are shown in Figure 3.4. We re-sized each one to (175 * 175).

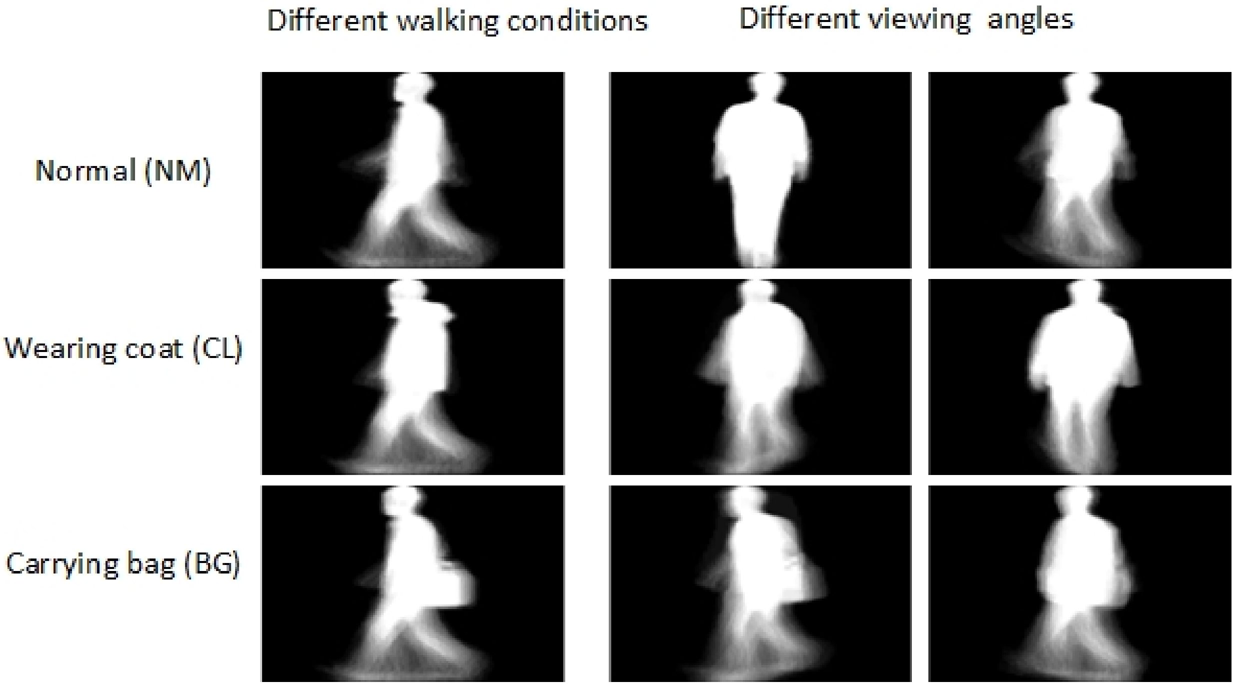


Figure 3.4: Samples of GEI from CASIA-B data set

We implemented our deep CNN model using Keras [75], a deep learning library written in Python. We also used the scikit-learn [76] , an open source Python library. We used (Intel ^{R} Xeon(R) CPU E3-1271 v3 with 3.60GHz × 8) and used Ubuntu Linux operating system.

CHAPTER 4: FEATURE SELECTION, AUGMENTATION AND SPARSE CODING

4.1 Our feature selection method

The feature selection for classification has gained enormous interest among the computer vision and pattern analysis community. Besides removing the irrelevant features, there are many other potential benefits, such as reducing the computation cost, storage requirements and the training time. The features in gait representation can be divided into three types: discriminant features, redundant features, and irrelevant features. Irrelevant features include objects such as bag or briefcase, to name a few. We argue that irrelevant features are unrelated and should not be a part of the classification process.

4.1.1 The evaluation of the most important features of gait

It is well known that not all features contribute equally in the classification process. In fact, some of the features may mislead the classifier and negatively impact the performance. Thus, before discarding any information from the GEI, we investigate which gait features in the GEI are considered to be redundant and what features are the most imperative. Ensemble algorithms are widely used to rank the importance of features in many pattern analysis tasks. Therefore, we used the Extra Tree algorithm [77], an ensemble algorithm, to estimate what features are more discriminative based on the Gini Impurity criterion. We did two experiments using two different databases namely, CASIA-B database and OU-ISIR-B database. In the first experiment, we used the gait data for all 124 subjects from the CASIA-B database. For every subject, we selected two sequences under normal condition where there is no bag or clothing covariate. The 50 most discriminative features (with gait under normal condition without covariate factors) are shown in the heat-map in figure 4.5. The hotter the value of the pixel, the more important it is. It indicated that the static features of GEI are very useful for recognition if the gait sequences are under normal conditions.

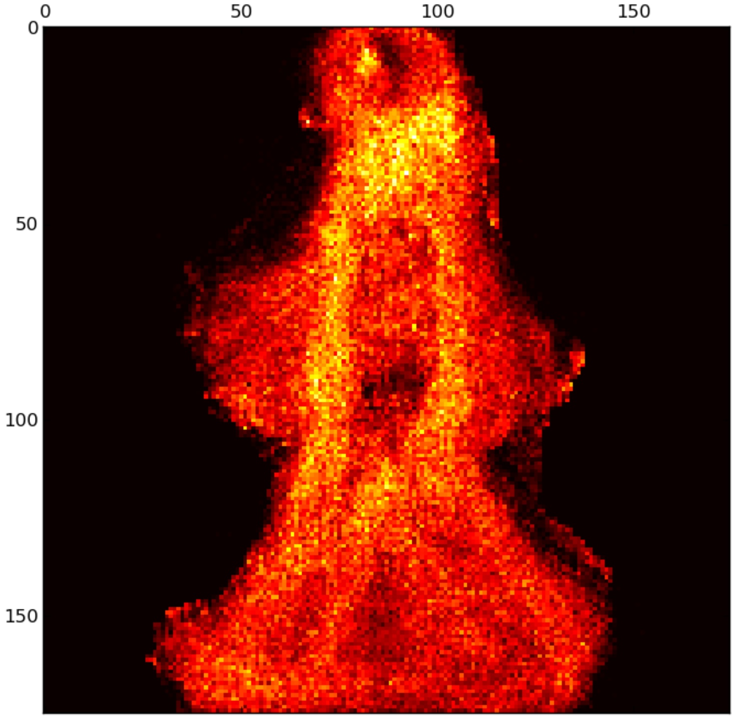


Figure 4.5: The most discriminative features of GEI under normal conditions (CASIA-B Dataset)

However, if there are different covariate factors in the dataset, the most discriminative features are different because of the intra-class variations caused by these covariate factors. In the second experiment, we used the gait data in the CASIS-B database for all 124 subjects and the gait data in the OU-ISIR-B database for all 48 subjects as in the (probe set). In this analysis, we included the carrying and clothing conditions in the training.

For every subject in CASIS-B database, we selected one sequence under bag condition, one sequence under clothing condition, and one sequence under normal condition.

For OU-ISIR-Treadmill dataset B [69], we used the probe set (with clothes variation up to 32 combinations) for all 48 subjects. Figure 4.6 and Figure 4.7 show the 50 most discriminative features when the gait is affected by these different covariate factors in both databases.

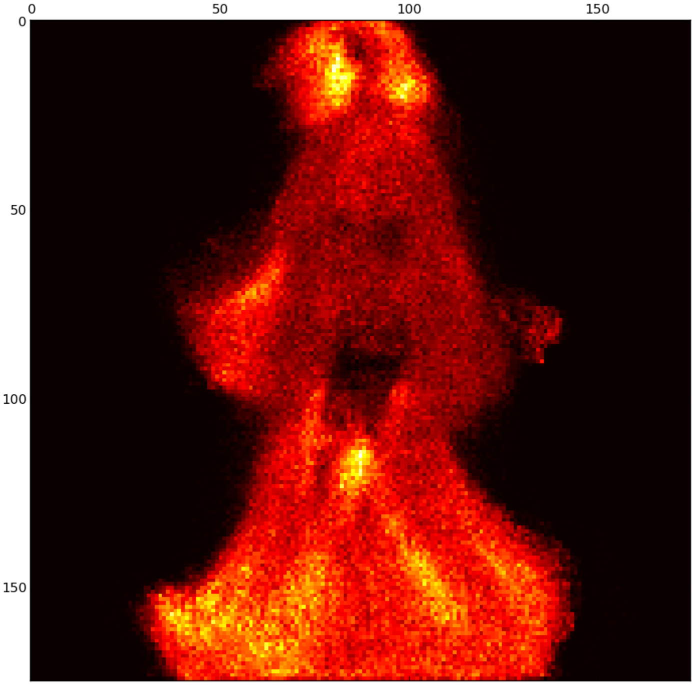


Figure 4.6: The most discriminative features of GEI under covariate factors (CASIA-B Dataset)

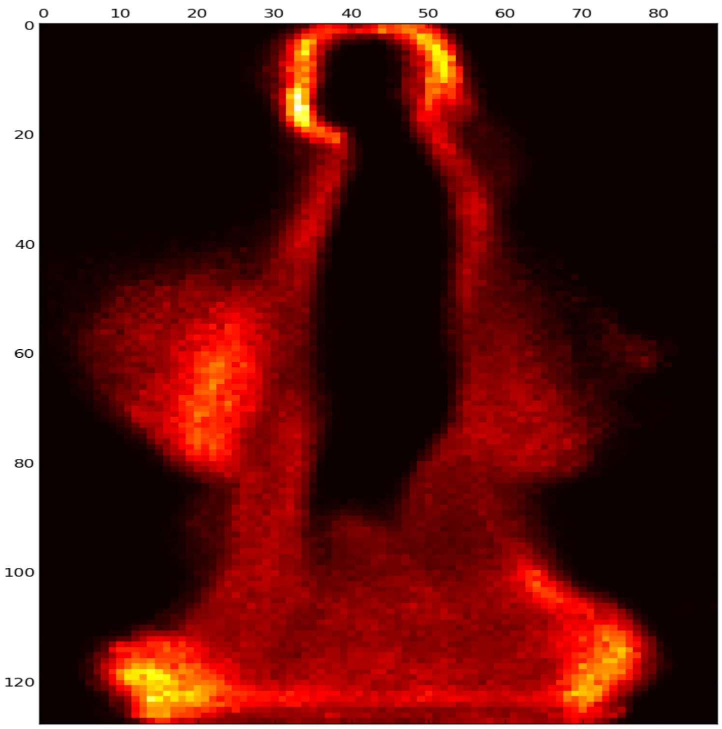


Figure 4.7: The most discriminative features of GEI under covariate factors (OU-ISIR Gait Database- Treadmill dataset B)

During the gait cycle, the head tilts to the right and left side of the body, causing a consistent, relative motion. Also, both limbs swing during the gait. Figures 4.6 and 4.7 show that part of the body that remains static during a gait cycle (particularly the trunk of the human body) has less important discriminative features. Conversely, the dynamic areas of the body that undergo consistent relative motion (such as arms, legs, and head) contain more important discriminative features. After these analyses, we conclude that the hot pixels as shown in figures 4.6 and 4.7 contain the most discriminative features that are less sensitive to covariate factors.

4.1.2 The removal of the irrelevant features

To overcome the problems caused by covariate conditions, we selected the most discriminant and relevant regions in GEI as highlighted in figure 4.6. Thus, we use the following algorithm to remove the irrelevant features and focus only on the most discriminant regions.

```
Data: GEI is the input
Result: SG: the new representation after applying the feature selection method
Looping through the GIE;
for $< y = 0; y < W; y++ >$ do
    for $< x = 0; x < H; x++ >$ do
        if $(x > (H * \beta))$ AND $(x < (H * \beta'))$ then
            if $(GEI[x,y] < T)$ AND $(y < W * \Omega)$ AND $(x < H * \eta)$ AND $(x > H * \eta')$ then
                Extract the middle region;
                $SG[x,y] = GEI[x,y]$;
            else
                $SG[x,y] = 0$;
            end
        else
            Extract the upper and bottom regions of the GEI ;
            $SG[x,y] = GEI[x,y]$;
        end
    end
end
```

Algorithm 1: Our feature selection method

where GEI is the averaged silhouette that has W x H size. x and y are the coordinates of the silhouette. β, β', Ω, η, and η' are the parameters that identify the regions of interest. T is a pixel-threshold value, to eliminate the static features that have low motion. Mathematically, our feature selection method can be represented in three steps. We first extract the upper and bottom regions of the GEI as shown in equation 4.4:

$$SG_A(x,y) = \begin{cases} SG_A(x,y) = GEI(x,y) & \text{if } (H*\beta) < y \; OR \; y > (H*\beta\text{'}) \\ SG_A(x,y) = 0 & \text{if otherwise} \end{cases} \tag{4.4}$$

Then we obtain the second region of interest as shown in equation 4.5:

$$SG_B(x,y) = \begin{cases} SG_B(x,y) = GEI(x,y) & \text{if } ((GEI(x,y) < T) AND (H*\eta < x \; OR \; x > H*\eta\text{'}) \\ & AND \; (y < W*\Omega)) \\ SG_B(x,y) = 0 & \text{if otherwise} \end{cases} \tag{4.5}$$

After that we combine both of them in 4.6:

$$SG(x,y) = SG_A(x,y) + SG_B(x,y) \tag{4.6}$$

The proposed method has six parameters that are decisive for the reliability of our feature selection. The objective of the threshold value T is to remove the static features and maintain the dynamic features. The optimal values for these parameters determined empirically are listed in table 4.2:

parameters	β	β'	Ω	η	η'	T
value	0.08	0.68	0.40	0.51	0.11	105

Table 4.2: The optimal values for the parameters of our feature selection method

Figure 4.8 shows an example of the GEI after applying the proposed feature selection method. The result is a new gait representation that reflects the unique characteristics of the gait that is invariant to appearance variations.

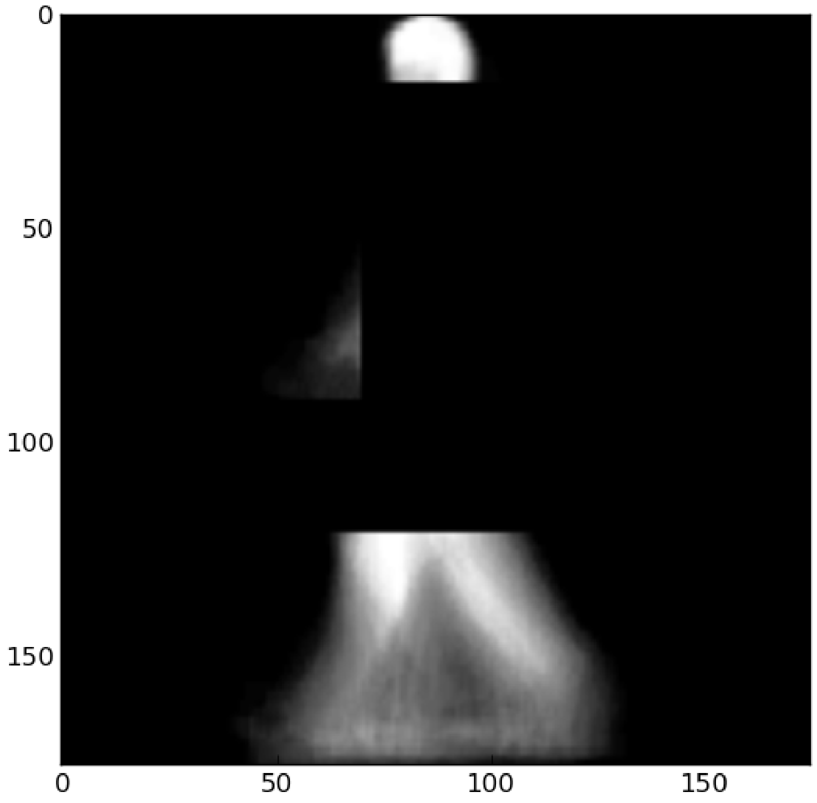


Figure 4.8: The gait representation (SG) after applying our feature selection to GEI

Since the data of GEI silhouettes are normalized, aligned, and centered, the proposed method is not affected by age, weight, gender, or even image post-resizing. Our method is applied to both the gallery data (before the training) and on the probe data (before classification). However, even if we only apply the proposed feature selection method on the training data, we still get a promising result which will be demonstrated later. Compared to the related works, our method does not depend highly on the extraction of the covariate conditions from testing data. We assume the varieties of the conditions in testing data and training data are unknown, boundless, and cannot be fully estimated or covered.

4.2 The augmentation of the gait representations

The GEI still exhibits poor performance in some situations even after applying our feature selection. The recognition is difficult due to natural body rotations, particularly when there is a carrying condition. The center of gravity becomes leaned and shifted as shown in figure 4.9.

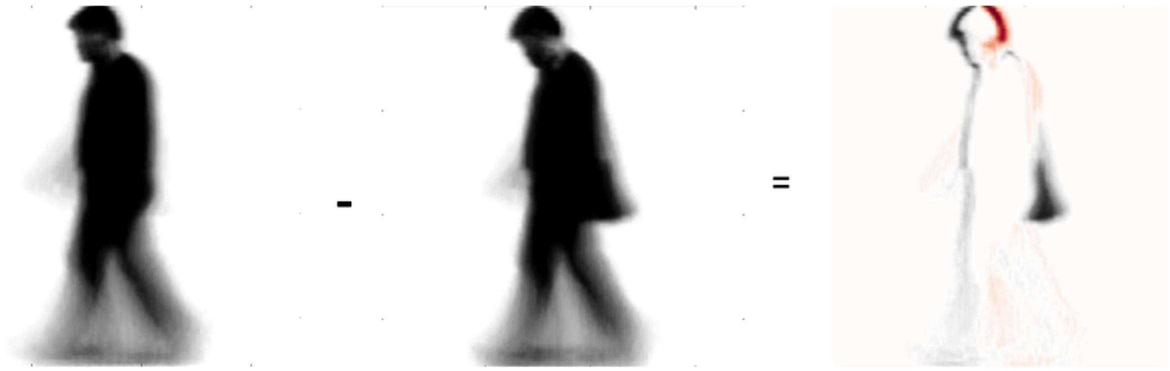


Figure 4.9: The effects of the carrying covariate factor on gait

Another reason for poor performance is the small amount of available training data. To overcome this problem, we augmented each one of the four training samples by duplicating it. Then, we reconstructed these new, augmented samples by randomly rotating each one by one or two degrees to the left and right. We also duplicated the training data again and shifted each one of them randomly left, right, up, or down by one pixel. This augmentation step overcomes some of the translation and fluctuation caused by the load carrying. It also increases the wealth of the feature set.

4.3 Classification

Feeding a large number of features directly into the classifier is not practical. These features are expensive to store, slow down the computations, and cause overfitting issues. In algorithms like K-Nearest Neighbors or Nearest Centroid classifier (which is used later), distances in high dimensions are distorted. In this paper, different from the existing gait recognition methods, we use online dictionary learning to decompose our data and the sparse representation vector (code) as our discriminant features.

The dictionary learning framework leads to the state-of-art results for numerous image processing and machine learning tasks. Compared to the subspace methods such as PCA, sparse representation is less sensitive to occlusions and noises. Thus, we combine the discrimination power of dictionary-based sparse coding and Linear Discriminant Analysis (LDA) to enhance the performance of gait recognition.

There are two types of the dictionaries: the non-adaptive dictionary (e.g., Fourier, Curvelet, Gabor, and Wavelet) and the adaptive dictionary (uses the training data). We used the adaptive dictionary method where the dictionary learns from the training data.

In sparse coding, the objective is to represent a given data as a linear combination of some given dictionary $D = [d1, d2, .., dk]$, which is a set of atoms, in such a way that only few components of the representation's coefficients are equal to non-zero and the rest are zeros. The sparse coding algorithm is used to find the sparsest representation such that $X \simeq \alpha D$. X is decomposed as a linear combination of only a few atoms of D.

In our work, we used the Online Mini-batch dictionary learning (as illustrated) in [78] which is much faster. It is iterative online algorithm, based on stochastic approximations and suitable for a large data set. It consists of two procedures: dictionary initialization and dictionary update. It uses the block-coordinate descent with warm restarts to update the dictionary and the least angle regression (LARS) to compute the sparse coding during dictionary updating phase. In every iteration during the dictionary learning phase, more than one sample (mini-batch) is processed allowing for faster convergence.

4.3.1 Learning the sparse dictionary

We used the original training data without augmentation to initialize the dictionary. Thus, the dimensionality of the data is 496×30625, where each one of the 496 gait representations (SG) is reshaped as a vector of size 30625. Then, the data are fed to Mini-batch dictionary to find the dictionary D. We used LARS method to compute the sparse coding at the dictionary learning phase.

The number of iterations is set to 27. The number of samples in each mini-batch is 124, and the number of dictionary atoms to be extracted is 116. On average, the training time for the dictionary learning took around 5.38 minutes.

The output obtained is a dictionary matrix that has a size of 116×30625. Some samples of the generated dictionary D are shown in figure 4.10 after being normalized and reshaped to a 2D image.

Then, using the precomputed dictionary, we compute the sparse codes based on the soft thresholding technique for the augmented data. The augmented data contains 1984 sequences. The threshold value, below which coefficients will be set to zero, is set to 1.

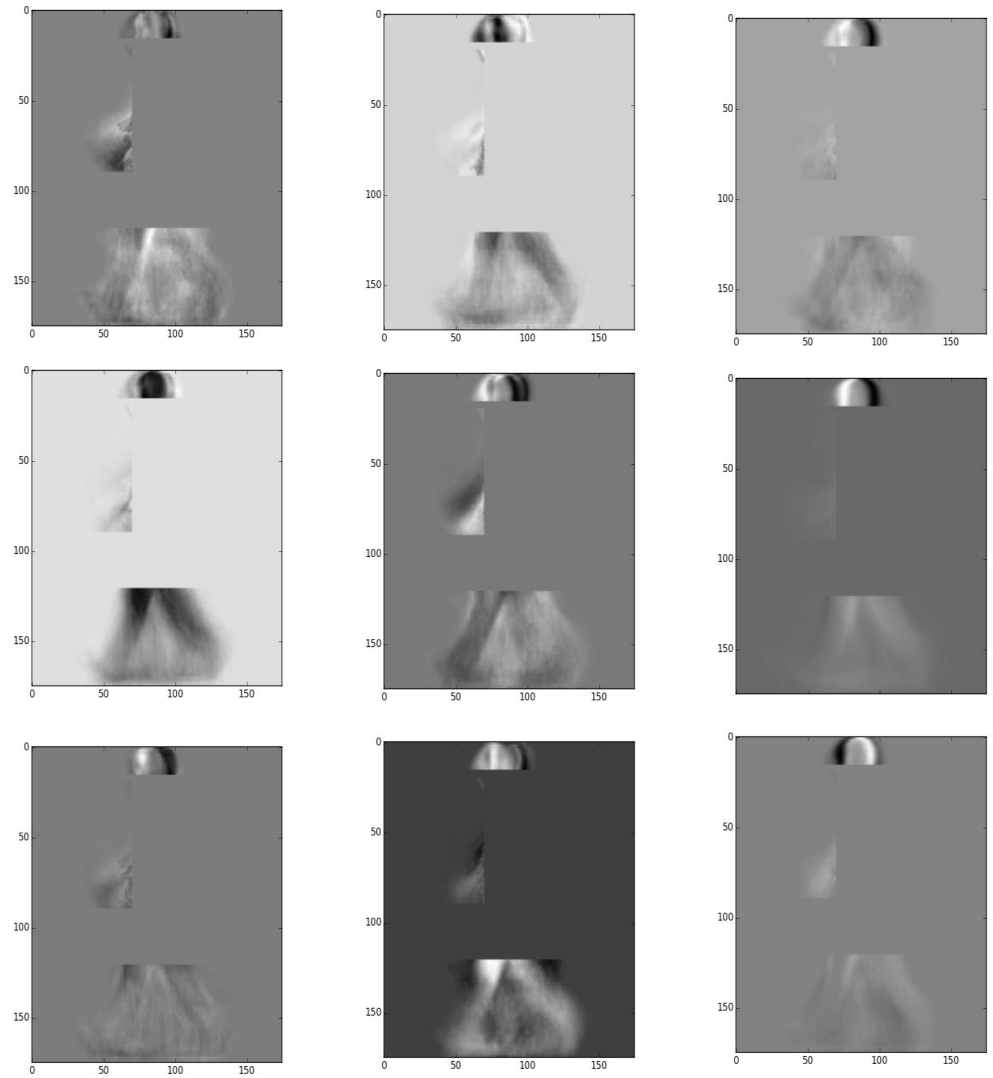


Figure 4.10: Some samples (atoms) of the generated dictionary D

An example of four sparse codes that belong to two different classes is shown in Figure 4.11. Each one of these codes contains 116 components. The codes that belong to the same class are almost identical.

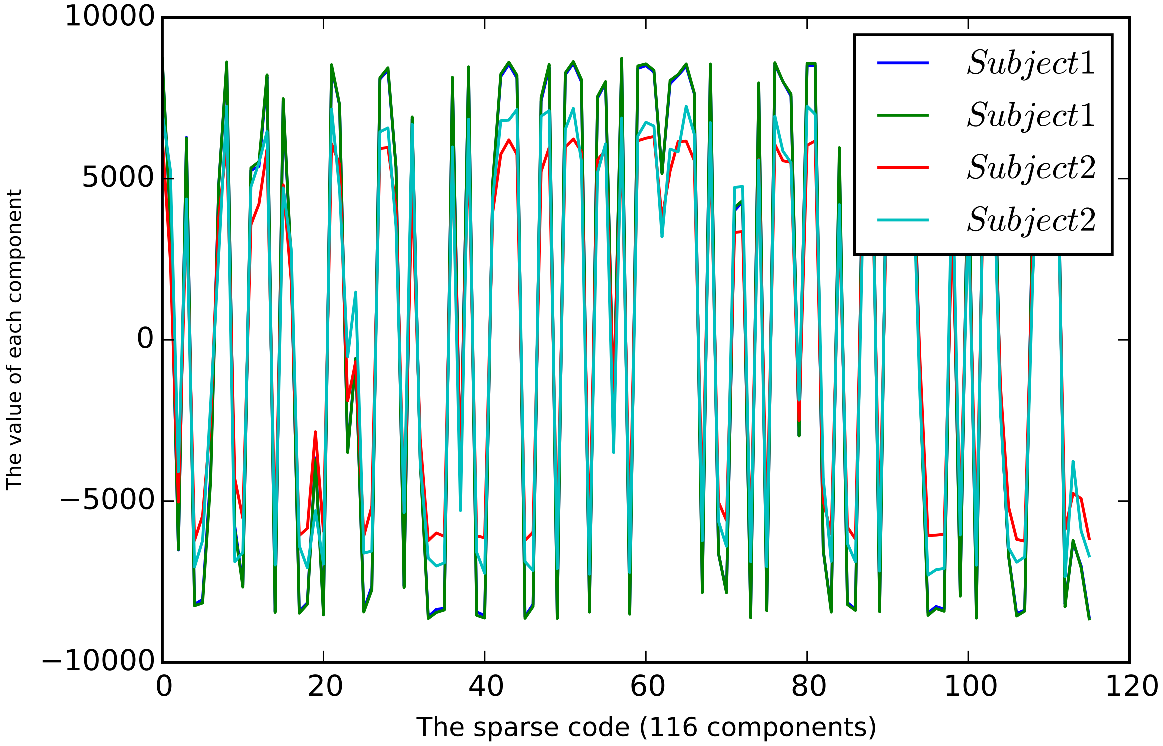


Figure 4.11: Four examples of the obtained sparse codes (for 2 different subjects)

After that, the sparse codes for the augmented data, which has a size of 1984 x 116, are fed to the Linear Discriminant Analysis to get better representation. The LDA is an efficient dimensionality reduction technique. It allows us to find the best components that maximize the separation between the classes. We only used the LDA to project the data to seek the best representation that maximizes the separability of the classes. Figure 4.12 shows the effects of the LDA projection on four of the obtained sparse codes.

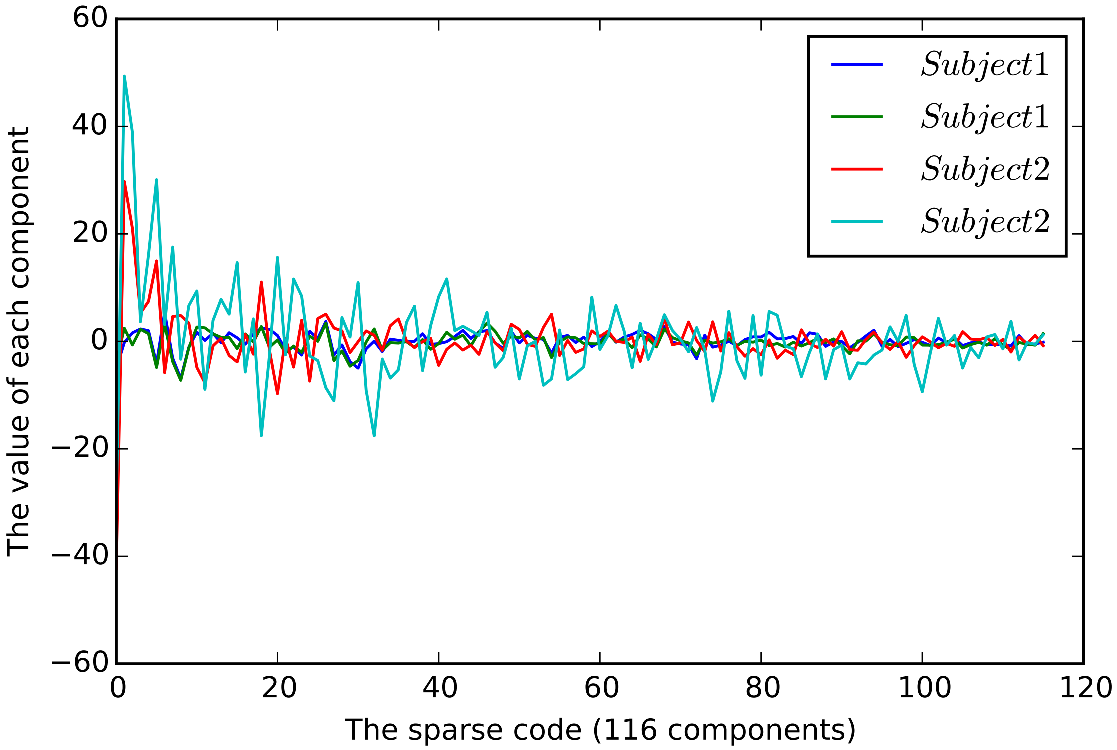


Figure 4.12: Four sparse codes after the LDA projection

Figure 4.13 illustrates the steps at training phase.

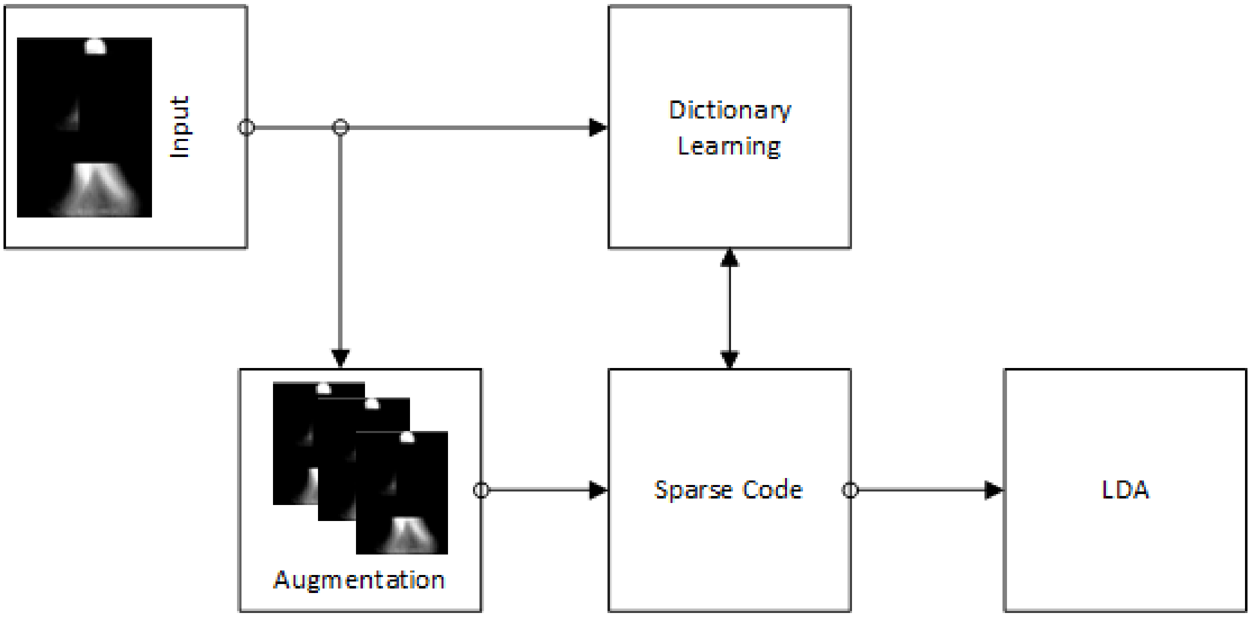


Figure 4.13: Our feature extraction pipeline at the training phase

4.3.2 The classifier

We used the Nearest Centroid classifier [79] to classify the output data of the LDA. Unlike the K- Nearest Neighbors, (which is computed based on the majority vote of the nearest neighbors of the probe sample), the Nearest Centroid algorithm computes a standardized centroid for each class. When there is a new probe sample, it is assigned to the class whose mean (centroid) is the closest. Figure 4.14 shows an example where the probe sample is assigned to class B although the probe sample is closer to sample 1 of class A. We find the Nearest Centroid classifier performs better than other classifiers. Also, contrary to other classification algorithms, there are no parameters to choose in the Nearest Centroid classifier.

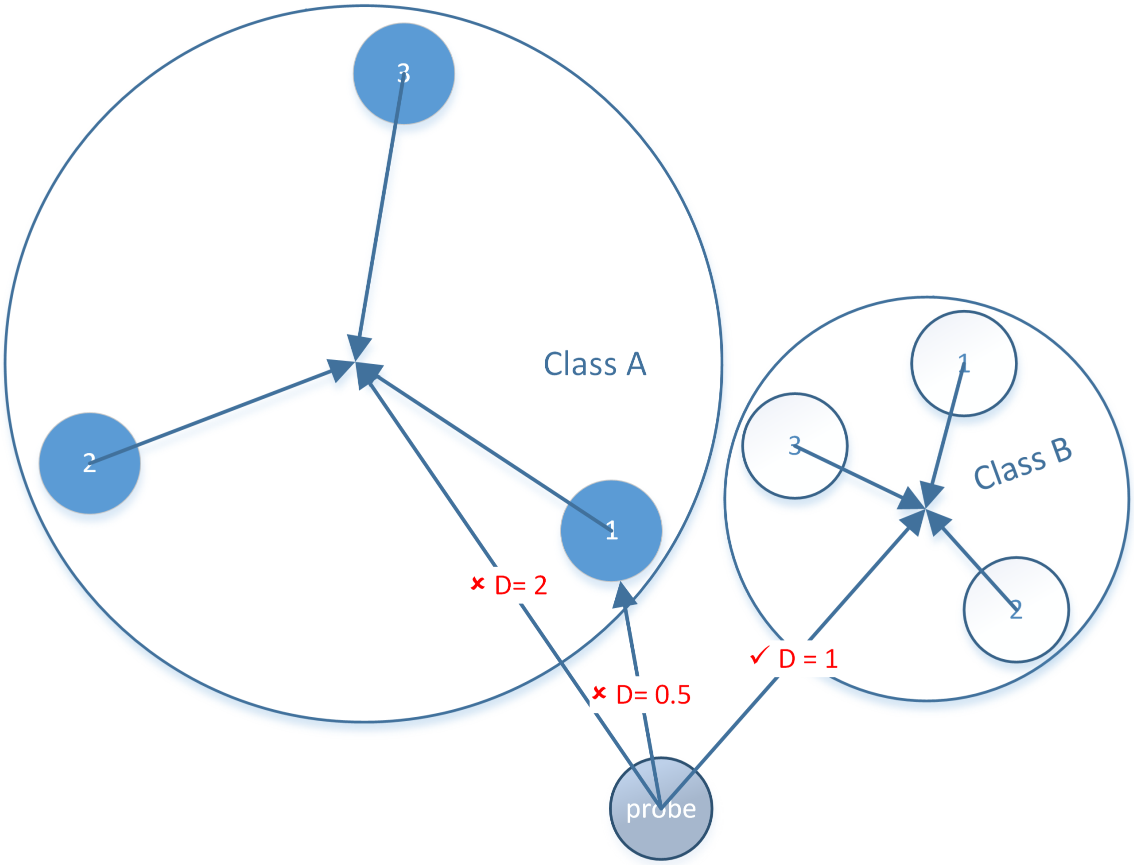


Figure 4.14: The Nearest Centroid classifier

At the testing stage, the input data (image vector) are transformed via the sparse coding algorithm to get the sparse code. The sparse code is a vector of size 116. Then, the sparse code is projected using the LDA and fed to the Nearest Centroid classifier (NC) classifier. Figure 4.15 illustrates our feature extraction pipeline method followed by the Nearest Centroid classifier at the testing phase.

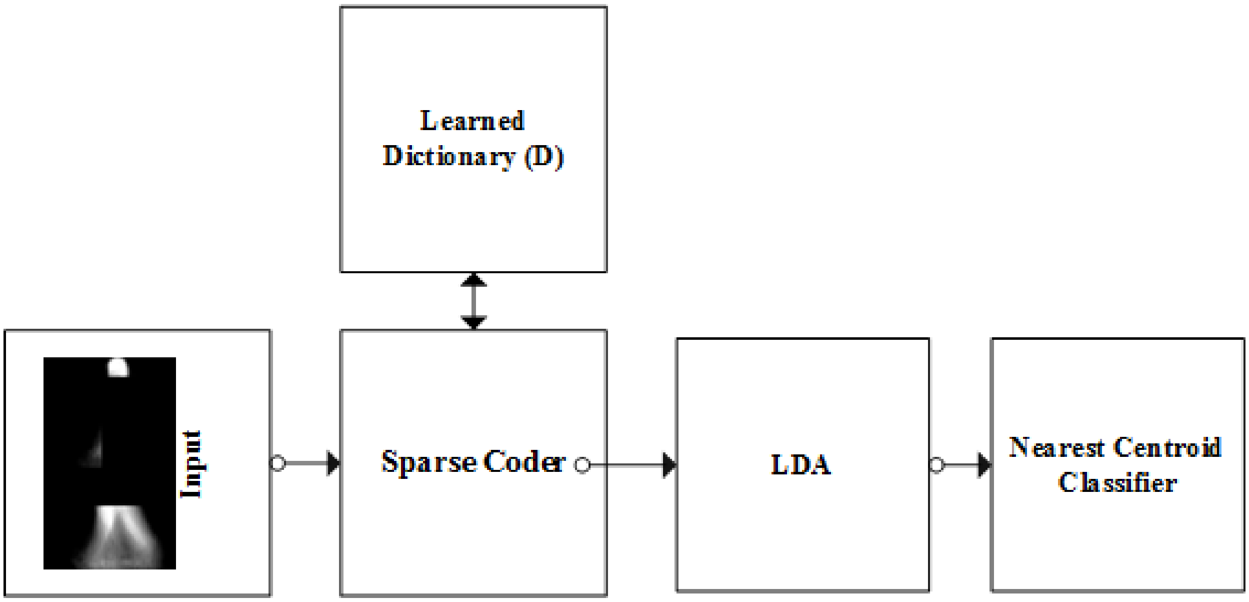


Figure 4.15: Our sparse model at the testing stage

4.4 Our experimental results

4.4.1 First experiment: gait under different covariate Conditions

The benchmark experiment of CASIA-B under angle 90° is used to evaluate our model and compare it with the recent works of gait recognition. In this experiment, the first four sequences of the normal condition set (NM) namely (Sequences: “nm-01”, “nm-02”, “nm-03”, and “nm-04”) are used as the gallery set for training. The two remaining sequences of NM (sequences: “nm-05”, and “nm-06”), the two sequences of BG (sequences: “bg-01” and “bg-02”) and the two sequences of CL (Sequences: “cl-01” and “cl-02”) are used as a probe set for testing. We carried out our experiments under 90° view angles. Table 4.3 shows the obtained results. On average, the experimental results show that our proposed model has the best recognition accuracy. It also shows that under the clothing condition (CL), our method achieves the best result.

CASIA-B	NM	BG	CL	Average
Baseline TM [64]	97.6	52.0	32.7	60.8
CGI [12]	88.1	43.7	43.0	58.2
GEI [10]	**100.0**	53.2	22.2	58.5
GEnI [80]	**100.0**	78.3	44.0	74.1
MII + MDIs [18]	97.5	83.6	48.8	76.6
SEIS [24]	99.0	72.0	64.0	78.3
$P_{RW}GEI$ [81]	98.4	93.1	44.4	78.6
Part-based Selection [20]	99.2	75.8	80.6	85.2
AEI + 2DLPP [16]	98.4	91.9	72.2	87.5
Matching Weight [39]	97.7	91.9	78.0	89.2
SGEI + GEI [82]	98.2	80.7	83.9	87.6
GEI + Group Lasso [37]	98.4	75.9	91.9	88.7
GPPE [21]	93.36	56.12	22.44	57.3
GEI + MPOC [83]	93.60	81.70	68.80	81.40
STIPs + BoW [84]	94.5	60.9	58.5	71.3
GEI + bolt-on module [38]	98.4	77.4	93.1	89.7
GVI + bolt-on module [38]	95.6	85.9	71.4	84.3
SVIM + bolt-on module [38]	98.0	**96.8**	73.0	89.2
M_G [35]	**100.0**	91.0	80.6	90.5
Our proposed method	98.4	86.7	**94.8**	**93.3**

Table 4.3: Comparison with other methods

Carrying a bag has more influence on the recognition accuracy. Although the bag does not sharply affect the body shape (it only affects a small area of silhouette), it affects the dynamic movement of the body. On the other hand, the clothes have less influence on the dynamic movement and more influence on the shape-appearance, i.e., occlude more pixels of the silhouette. The results of CGI [12] and GEI [10] are as reported in [38]. It should be noted that the methods in [16] and in [39] have used a different way other than the baseline to calculate the CCR. Instead of using the first four sequences of NM for the training, they evenly split the six normal sequences of each subject into training and testing and use the 2-fold cross validation method to calculate the

CCR. The average computation time for one Sequence at the testing process is around 0.2 second.

4.4.2 Second experiment: the effect of the augmentation and feature selection

In the previous experiment, we assumed that the proposed feature selection is applied to both the testing and training set. However, we found that if the proposed feature selection method is applied only on the training data, we still get a promising result, as shown in Table 4.4. Also, even without the without the augmentation step, we still can achieve the best result in the CL case as well as the best average of the three cases.

CASIA-B	NM	BG	CL	Average
Our method without Augmentation	98.8	79.1	93.9	90.6
Testing Data without feature selection	96.8	83.1	85.5	88.5

Table 4.4: The effect of the augmentation and feature selection on the probe set

4.4.3 Third experiment: gait under the viewing angle variations

In the third experiment, we compared our result under the viewing angle variations. We only applied the proposed feature selection to the training data, leaving the probe set as it was. We used the same precomputed dictionary from the previous experiment. The first four sequences of the set (NM) namely (Sequences: “nm-01”, “nm-02”, “nm-03”, and “nm-04”) under angle 90° were used as the gallery set for training. The rest of the data as indicated in table 4.5 (11 different viewing angles, each under 3 different conditions) were used as the probe. We compared our result with the baseline algorithm [64]. As indicated in the table, the left side shows the results of our method and the right side shows the results of the baseline method. Averaging all the cases, our method has better performance.

	Our method				Baseline			
Viewing angle	NM	CL	BG	AVG	NM	CL	BG	AVG
0°	1.6	0.8	2.1	1.5	0.4	0.4	1.2	0.7
18°	1.2	0.8	1.6	1.2	2.4	2.8	2.4	2.5
36°	2.4	1.6	2.8	2.3	4.8	5.2	4.0	4.7
54°	11.7	8.1	8.5	9.4	17.7	8.5	6.0	10.7
72°	86.7	71.8	70.9	76.5	82.3	42.3	20.6	48.4
90°	96.8	85.5	83.1	88.5	97.6	52.0	32.7	60.7
108°	62.1	37.9	39.1	46.4	82.3	31.9	16.5	43.6
126°	4.838	3.3	8.1	5.4	15.3	9.7	6.0	10.3
144°	4.1	2.9	4.9	3.9	5.2	6.0	3.6	4.9
162°	1.6	0.8	2.4	1.6	3.6	3.2	3.2	3.3
180°	0.4	0.8	0.5	0.6	1.2	2.0	0.8	1.3
	Total average = 21.6				Total average =17.4			

Table 4.5: The results of our method and the baseline method under different viewing-angles

4.4.4 Fourth experiment: using OU-ISIR gait database- treadmill dataset B

In this experiment, we have applied our approach to OU-ISIR-Treadmill Dataset B. The database has pre-defined the probe and gallery sets. In the gallery set, each subject has one sequence. From each sequence, we generated three GEIs. Our proposed method achieved a promising result, as shown in Table 4.6.

Table 4.6: The result of our method using OU-ISIR-B dataset

Method	Recognition Accuracy
Baseline TM [64]	51.2
M_G [35]	32.7
Our proposed method	53.7

4.4.5 Discussion

From the experiments carried out, we can conclude that the proposed model is more robust against the covariate factors. Our feature selection extracts the most discriminative features and avoids the irrelevant features. Then, the augmentation technique overcomes the issues of intra-class gait fluctuations and the small size of the training data. Thus, by feeding the augmented data to the model of the dictionary learning with LDA, we generate multiple sparse codes of the same signal, where each one can capture unique features.

CHAPTER 5: THE DEEP CONVOLUTIONAL NEURAL NETWORK

5.1 Our Deep CNN Architecture

In our deep CNN architecture, we have eight layers: four convolutional layers and four subsampling (pooling) layers. Each of these layers has eight feature maps. There are eight convolutional filters that are randomly initialized in every convolutional layer, and eight subsampling maps in each subsampling layer. These layers are trained using the backpropagation learning algorithm. Root Mean Square Propagation (RMSProp), which is an improved version of stochastic gradient descent with an adaptive learning rate, is used as our optimization algorithm to minimize the cost function. Thus, the weights of the network are updated every iteration using a small number of input batches ranging from 4 - 25. We prefer to use the RMSProp optimizer because it is simply much faster. The input data is normalized by dividing each pixel value by 255. An initial learning rate value of 0.001 is used for all layers in our model. We use the Gait Energy Image (GEI) [10] with the size of (140 * 140) as the gait feature descriptor and input to the CNN. We conduct all the next experiments on the CASIA-B data set [64].

5.1.1 The convolutional method

The weights of the convolutional filters are randomly initialized in a uniform distribution using the Xavier 5.7 uniform variance scaling method [85].

$$Var(W) = \sqrt(6/(Fan^{in} + Fan^{out})) \tag{5.7}$$

A convolutional filter of the size 5x5 is applied with a stride of 1. The output feature maps are added with the bias terms, where each feature map has one bias term, and then the result is transformed by the nonlinear activation function. Each feature map in the convolutional layer is calculated as shown in Equation 5.8, below:

$$FM^i = Tanh(W^i \otimes FM^{i-1} + \beta^i) \tag{5.8}$$

where $\otimes$ denotes the convolution operation and FM^{i-1} is the feature map from the previous layer. On the first layer, the FM^{i-1} represents the raw pixels of GEI. Each feature map has a bias term β. The bias terms are initialized to the value of zero. We use the Hyper Tan function as our activation function which is defined as per Equation 5.9:

$$Tanh(x) = \frac{e^x - e^{(-x)}}{e^x + e^{(-x)}} \tag{5.9}$$

where x is the result of the convolutional operation added to the bias term of that feature map as shown in Equation 5.8. We have tried three other activation functions: the sigmoid function, the rectified linear unit (ReLU), and the Leaky ReLU. In our case, these did not produce better results. In the first convolution layer, each one of the eight units produces a 136 x 136 output. In the third layer, each one of them produces a 64 x 64 output feature map. In the fifth layer they produce a 28 x 28 output feature map. In the last convolution layer, each one of the eight filters produces a $10x10$ output feature map. The outputs of the convolutional layer are fed directly to the eight subsampling units in the pooling layer.

5.1.2 The pooling method

Each pooling layer, in our CNN model, outputs eight pooled feature maps that summarize the output values of the neighboring groups of neurons of each kernel map. It also assists in reducing spectral variance in the input data and produces translation-invariant features. This advantage is very valuable in gait recognition, since the shape of the body in gait recognition is a non-rigid shape that can undergo many fluctuations.

In our model, the pooling units perform the max pooling, where the pooling factor is C = 2. Therefore, it down-samples the data by using the max pooling filter of the size 2 x 2 applied with a stride of 2. Hence, the pooling windows in this model are non-overlapping. The operation of the pooling layer is defined as follow:

$$FM^i = MaxP(FM^{i-1}) \tag{5.10}$$

where *MaxP* denotes the max pooling operation. In the first subsampling layer, each one of the 8 pooling filters produces a 68 x 68 output. In the fourth layer, each one of the pooling filters produces a 32 x 32 output. In the six layer, each one of them produces a 14 x 14 output. In the last pooling layer, each one of the 8 pooling filters produces a 5 x 5 output.

In the fully-connected part, we have only two layers (an input layer and an output layer), where the soft-max is our classifier. We do not have any hidden layers. The input layer has 200 neurons that come from the last pooling layer (5 x 5 x 8).

5.1.3 The connections between the layers

In our CNN model, each map FM^i in layer l is only connected to one feature map FM^i from the previous layer $l-1$. This greatly reduces the computation cost, speeds up the training time, and reduces the number of parameters. Figure 5.16 shows an example of the one-to-one connections or the single connection between the kernels of three layers.

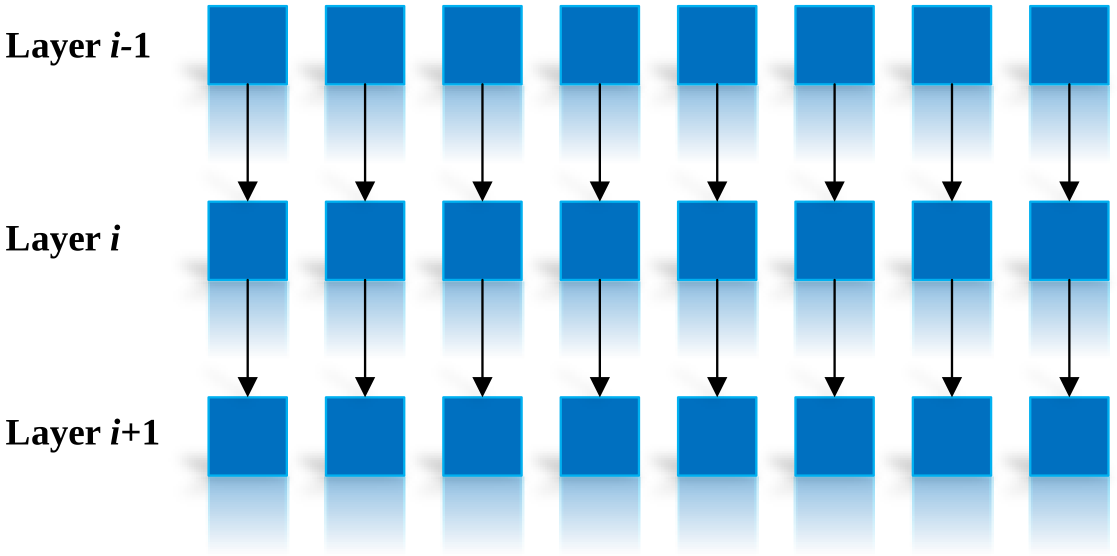


Figure 5.16: Illustration of the one-to-one connections between layers

Usually, the deep CNN models, such as GoogLeNet [60], consist of millions of parameters and are trained on large datasets. However, in gait recognition, the dataset is relatively small and cannot afford to train all of these parameters. Hence the over-fitting issues may occur. Table 5.7 shows the comparison between the total number of parameters in our model with the standard CNN model where each feature map is fully connected to all previous maps.

We are referring to this standard model as a typical CNN model. Both models are using the same settings, where there are 8 layers and each layer has 8 maps. Also, in both models, we use a softmax classifier that has 200 inputs to predict 124 classes.

Layer	Size of feature map	Number of parameters	
		Our model	standard model
Convolution	(136, 136)	208	208
Max Pooling	(68, 68)	0	0
Convolution	(64, 64)	208	1608
Max Pooling	(32, 32)	0	0
Convolution	(28, 28)	208	1608
Max Pooling	(14, 14)	0	0
Convolution	(10, 10)	208	1608
Max Pooling	(5, 5)	0	0
Fully Connected	SoftMax (124)	24924	63612
Sum of trainable parameters		20,932	68,644

Table 5.7: Comparison between the number of parameters in our model with the standard CNN model

The fully proposed architecture of our CNN is shown in Figure 5.17.

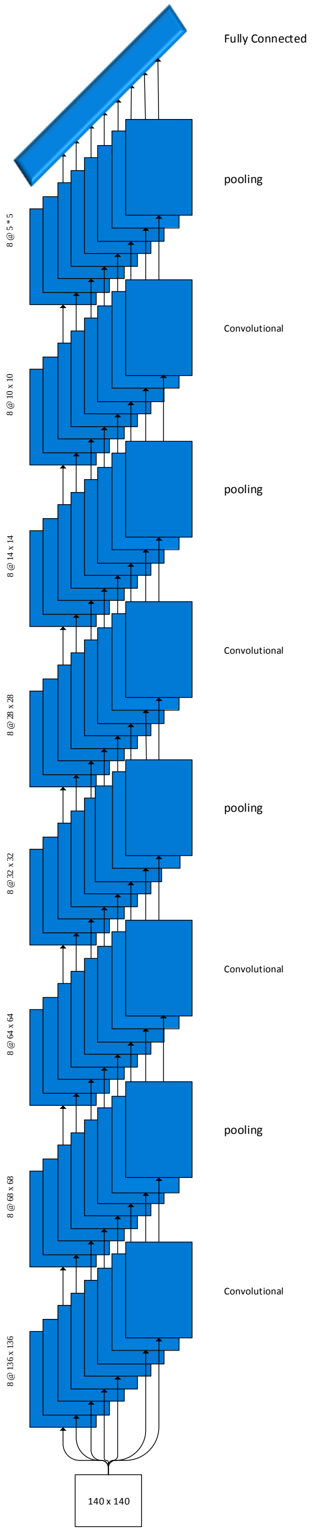
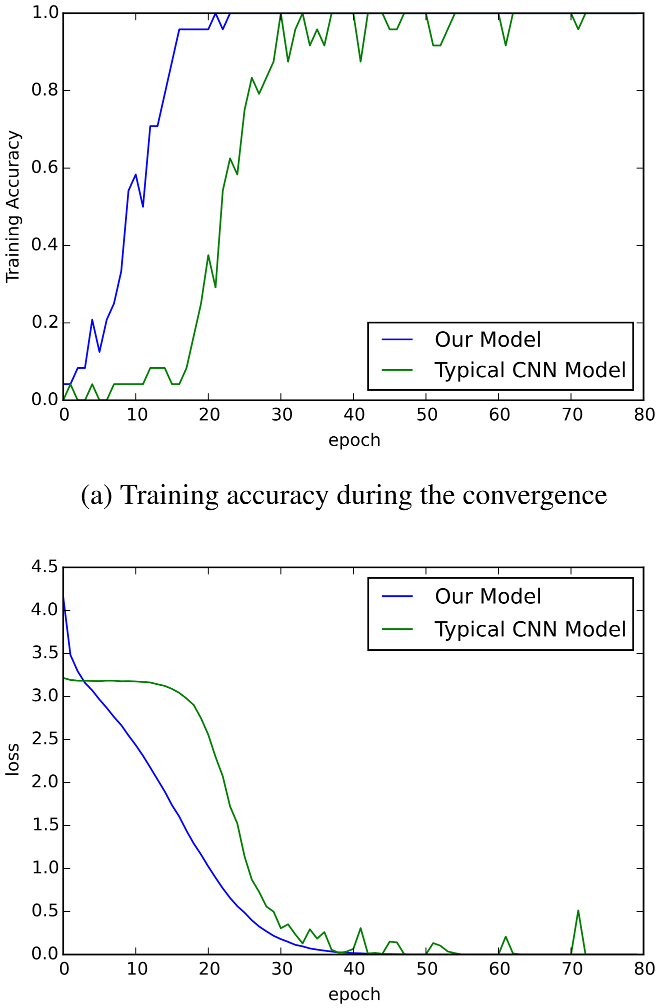


Figure 5.17: Illustration of our proposed CNN architecture

To further illustrate the advantage of our model, we trained both models by using the data of the first 24 subjects, where each subject has only one sample (namely: nm-01 under 90°) in the gallery. Both models are trained for 80 epochs using the RMSprop optimizer with a batch size of 2. We use ReLU as the activation function and Softmax for the classification.

The training accuracy and the loss during the convergence of both models to the local minimum are shown in Figures 5.18a and 5.18b. Our model converged efficiently with approximately 20 epochs, where the typical CNN took an estimated 75 epochs.



(a) Training accuracy during the convergence

(b) Loss during the convergence

Figure 5.18: Comparison between our model the typical CNN model on small dataset

Our model converged efficiently because it has fewer parameters that need to be trained. Our CNN model doesn't have fully connected layers other than the classifier. In addition, the connections between the features maps are based on one-to-one connections.

5.2 Open-set gait recognition

The existing methods of gait recognition, as in the literature, are designed to deal with closed set recognition. These methods are trained to recognize gait from known people who are registered and labeled during the training phase. However, the objective of open set recognition is to deal with unseen classes (claimed identities) who have not been registered during the training phase. For example, if the system was trained with only the data of 24 subjects, the appearance of a new subject at the testing phase will mislead the deep learning algorithm and be recognized as one of the 24 subjects.

This outcome is due to the closed nature of the deep learning networks which has been addressed by several recent deep learning works [86]. Thus, during the training of our CNN, we added one more class (background class) to detect any unknown subject and to immunize the CNN from being fooled by impostors.

We use the Gait Energy Image (GEI) [10] as the gait feature descriptor and input to the CNN. The silhouettes are normalized to a smaller fixed size (140 * 140) to speed up and accelerate the training process of the CNN. Additionally, we resized the images to smaller images to overcome the issue of a low resolution sequences and thus to simulate the real scenario (where the distance between the camera and the target can be far).

We conducted an experiment to show the robustness of our method. In our experiment, we divide the data in terms of the subjects as follows:

- **Closed Set**: This data set contains the gait data of the first 24 subjects [nm-01 - nm-04] under “90°”. They are annotated with labels from 0 – 23. These are well recognizable subjects.

- **Joint Set**: This data set contains the gait data of the next 25 subject with sequences [nm-01 - nm-04] under “90°” as one joint gait data to train our CNN to detect any unknown subject or potential impostors. These gait data are assigned during the training to a single joint class with the label 24.

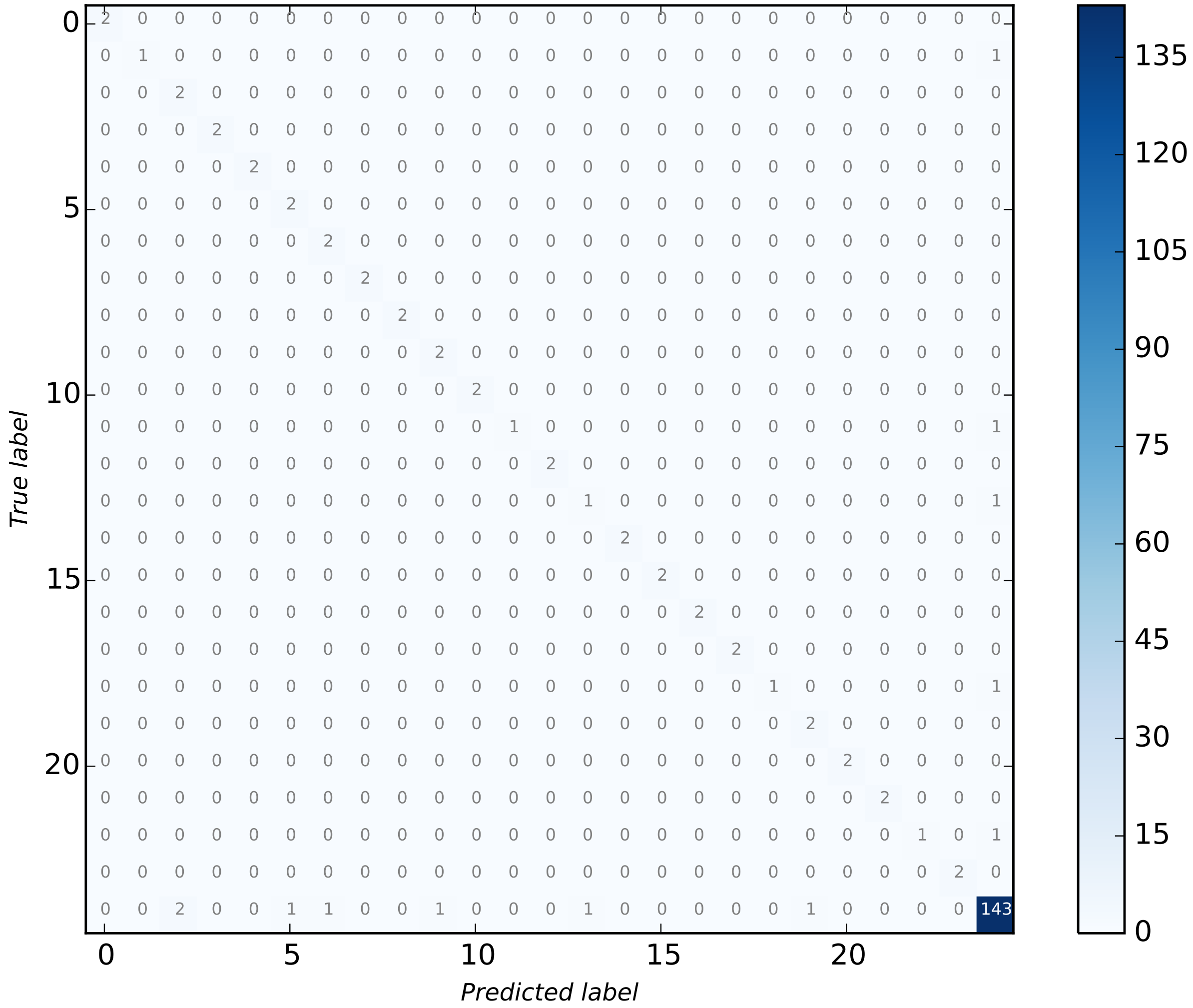
- **Open Set**: This data set is mutually exclusive and doesn’t overlap with the closed set or the joint set. It doesn’t not appear in training and is kept only for testing. This data set contains the remaining data for the next 75 subjects. It is used to test the robustness of our method against any unknown classes.

At the training phase, we trained our deep CNN with both the closed set with sequences [nm-01 - nm-04] and the joint set with sequences [nm-01 - nm-04].

During the testing phase, the data for the first 24 subjects (Closed Set) with testing sequences [nm-05 - nm-06] and the data of the final 75 subjects (Open Set) with sequences [nm-05 - nm-06] are used as the probe set. We set the batch size to 4, and the number of epochs to 30. We were able to obtain a recognition accuracy of 93.93%.

The confusion matrix is shown in Figure5.19. It shows that 143 out of 148 were detected as impostors.

This experiment demonstrates that the above method ideally addresses the open set gait recognition using deep CNN and that the method can detect unknown classes.



2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
0	0	2	0	0	1	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	143

Figure 5.19: Confusion Matrix for the Open-set gait recognition experiment with an accuracy of 93.9

5.2.1 Enrollment of a new subject

We analyzed the computation time for enrolling a new subject using our CNN model. First, we trained our CNN model with the gait data of 24 subjects. We used the training sequences [nm-01 - nm-04] under 90 degree. Then, we conducted two different transfer learning methods. In the first method, we added a new subject to the data, dropped the weights of the softmax layer, and re-trained the entire model with the data of 25 subjects. We fine-tuned the weights of the pre-trained network with a new softmax layer that has 25 outputs. This method is denoted as ”Fine-tuned CNN.’ In the second method, which denoted as ”Re-learn softmax only’, we froze the weights of the convolutional layers, and relearned only the weights of the softmax layer. We stopped the iteration when the training accuracy reached 100%. Table 5.8 shows the computation time of training the model from scratch with the data of 24 subjects as well as the computation time of training the model using both transfer learning methods.

Method	Subjects	computation time
Training of our CNN Model	24 subjects	124.82 seconds
Fine-tuned CNN	+1 subject (25 subjects)	42.41 seconds
Re-learn softmax only	+1 subject (25 subjects)	22.12 seconds

Table 5.8: Computation time

The fully connected part, which contains the softmax, has the largest number of weights when compared to the number of weights in the convolutional layers.

5.3 Experimental results and comparison with other gait methods

In order to compare the result of our method with our other methods, we have conducted three different experiments on the CASIA-B data set [64] using our proposed CNN model. For each subject, there are ten video sequences. From these ten sequences, two sequences are captured in different carrying conditions (bg), six sequences in normal walking conditions (nm), and two with different clothing conditions (cl). For each subject, the ten video sequences were captured from 11 views, namely: “0°”, “18°”, “36°”, “54°”, “72°”, “90°”, “108°”, “126°”, “144°”, “162°”, and “180°”.

5.3.1 First experiment: Gallery and probe under identical view with similar clothing and carrying Conditions

In this experiment, our work is compared to the results of other existing approaches on the CASIA-B dataset as reported in literature [34] focusing only on three cases of small variations (same covariate conditions) of $[gallery - probe]$. These cases are $[nm - nm]$, $[bg - bg]$, and $[cl - cl]$. These cases are under one viewing angle which is 90°. In the cases of clothes and carrying conditions, not only the shape of the human body gets affected, but also the dynamics of human walking get affected. In this experiment, we used the dataset with all 124 subjects. In $[nm - nm]$ the first four of six normal sequences are taken as Gallery and the last two sequences as the probe set. Then, the 2-fold cross validation is used to calculate our recognition rate for the last two cases where there are two sequences in each case. The performance is promising, specifically in the case of $[bg - bg]$ and the case of $[cl - cl]$. Table 5.9 shows the obtained results which are reported in terms of the correct classification rates (CCRs).

Methods	[nm-nm]	[bg-bg]	[cl - cl]
LF+AVG [33]	71.4	63.1	60.7
LF+DWT [33]	61.9	17.9	0.0
LF+oHMM [33]	63.8	31.8	21.4
LF+iHMM [33]	94.0	64.2	57.1
GEI + PCA + LDA [10]	90.5	3.6	3.6
GPPE [21]	93.4	62.2	55.1
GEnl [80]	92.3	65.3	55.1
STIPS [34]	95.4	73.0	70.6
Ours (deep CNN)	98.3	83.87	89.12

Table 5.9: Comparison with other methods under similar covariate factors by accuracy

The average computation time for the three cases in this experiment is indicated in table 5.10.

Training	Testing (1 sequence)
9.11 min.	0.02 sec.

Table 5.10: Computation time

In the next experiment, we consider relatively large variation cases.

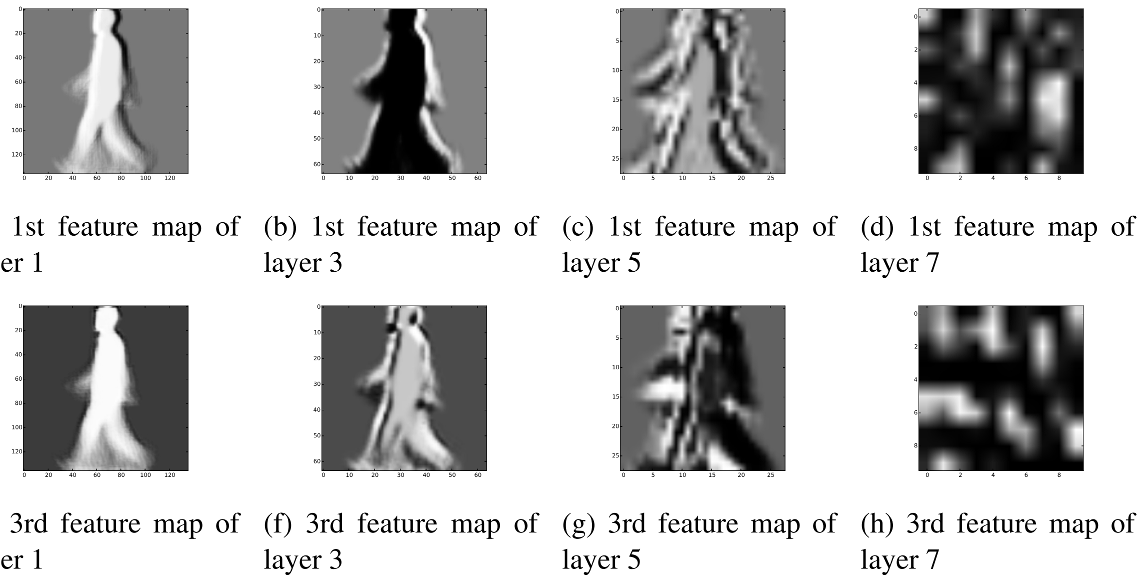
5.3.2 Second experiment: Gait recognition without Subject Cooperation

We used the experiment that was designed in [36] for gait recognition without subject cooperation under angles of 90°. In this experiment, the gallery consists of three sequences under three covariate conditions (one under normal condition, one under carrying condition, and one under clothing condition). The remaining seven sequences are used as the probe. In the probe set, there are five normal sequences (SetBX), one sequence under carrying condition (SetAX) and one clothing covariate sequence (SetCX). Then, we used the 24-fold cross-validation to cover all possible scenarios. Our results are compared with the results that have been reported in literature [36] as shown in Table 5.11.

Methods	Recognition Accuracy		
	SetAX	SetBX	SetCX
GEI + CDA [10]	44.90(± 22.6)	37.37(± 18.52)	41.23(± 20.4)
M_G + ACDA [35]	23.13(± 1.81)	18.94(± 3.96)	20.26(± 4.1)
Masked GEI + CDA[36]	56.80 (± 1.98)	63.27(± 2.77)	61.86(± 3.21)
Ours (deep CNN)	86.22(± 1.8)	82.92 (± 2.7)	87.80 (± 2.4)

Table 5.11: Comparison with other methods using ”uncooperative subject” experiment by accuracies

The advantage of this experiment is that it can cope with a large number of mixed covariate factors. In a deep CNN, under large covariate factors in the training, the discriminative features of gait are extracted in an automatic way, where there is no handed allocation of the discriminative features. Thus, it makes it more accurate because it avoids the loss of essential features. Some internal features maps of subject 1 in layers 1,3,5, and 7 are shown in 5.20



(a) 1st feature map of layer 1

(b) 1st feature map of layer 3

(c) 1st feature map of layer 5

(d) 1st feature map of layer 7

(e) 3rd feature map of layer 1

(f) 3rd feature map of layer 3

(g) 3rd feature map of layer 5

(h) 3rd feature map of layer 7

Figure 5.20: Display of some output feature maps

We also normalized the values of the learned conventional filters into gray-scale values and then visualized the first two learned filters of each conventional layer as shown in Figure 5.21

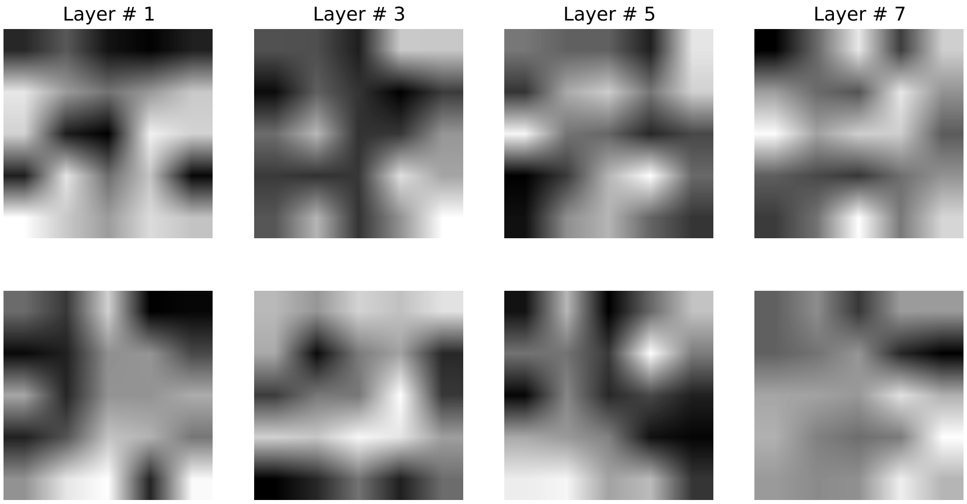


Figure 5.21: The first two learned filters of each convolution layer

5.3.3 Third experiment: Cross-view gait recognition

In the third experiment, we used the cross-view gait recognition. As in the previous works in this category, the CASIA-B data set is divided into 24 subjects for training and 100 subjects for performance evaluation. The data of the 100 subjects is divided into the gallery set [nm-1 - nm-4] and the probe set [nm-5 , nm-6]. We have 2 sections as indicated in Table 5.12. In both sections, we worked with 3 probe-viewing angles that included 45°, 90°, and 126°. In the first section on the left, the gallery consists of all viewing angles from 0° to 180°, excluding the probe angle in every case. Here, the gallery has ten different viewing angles. In the right section, the gallery consists of all viewing angles from 36° to 144°, excluding the probe angle in every case. On the right section, the gallery has six different viewing angles.

Gallery	(0° - 180°)			(36° - 144°)		
probe angle	54°	90°	126°	54°	90°	126°
Methods						
PCA	14.8	23.7	18.5	21.5	38.5	27.8
LDA	15.0	23.5	17.7	21.8	38.0	26.7
LPP	15.4	23.8	18.7	23.3	38.5	28.2
LFDA	51.0	45.9	48.6	72.2	70.2	70.8
SVR [87]	23	29	34	35	44	45
ViDP [49]	59.1	50.2	57.5	83.5	76.7	80.7
Matching-CNN [57]	77.8	64.9	76.1	90.8	85.8	90.4
Auto-encoder [58]	63.3	62.1	66.3	-	-	-

Table 5.12: Comparison between existing methods under the cross-view gait experiment

In this experiment, we trained our network using the training set, that consisted of only the data of 24 subjects. Then, we froze the weights of the first layer and re-trained the rest of the layers of our the network using the gallery set. Thus, the first layer is excluded from the training. There are 208 parameters trained using the training set and 20,724 parameters fine tuned using the gallery set. In other words, the first layer is pre-trained using gait data that consists of 11 view angles, while the last three layers are trained with data under 10 view angles. We trained our network with all of the view angles in the gallery set at the same time. In this case, we assumed that we had access to ten view angles on the left section or to six view angles on the right section (excluding the probe view angle). We used one training (projection using) for every probe case.

Gallery	(0° - 180°)			(36° - 144°)		
probe angle	54°	90°	126°	54°	90°	126°
Our Deep CNN	71.5	69.0	80.0	69.0	81.5	86.50

Table 5.13: The results of our method under the cross-view gait (one to multi-view) experiment by accuracies

Our model was able to learn the discriminative features. The results in Table 5.13, shows that our proposed deep CNN model has promising results when the gallery consists of gaits from multiple views. The results of using the following subspace learning methods: PCA, LDA, Locality Preserving Projections(LPP) [47] and Local Fisher Discriminant Analysis (LEFDA) [48] are as reported in [49]. There is another scenario used by the previous VTM-based methods where the gallery has one changing view angle and the probe has one fixed view angle. In this case, the gallery is divided into 10 gallery sets, where each contains the gait data from one view angle. We tried to train our CNN separately on every one of these 10 sets. However, our CNN model cannot learn very well under this scenario because the size of the gallery sets are very small and don't contain sufficient data. The average computation time is indicated in Table 5.14

Training	Testing (1 sequence)
61.94 mins.	0.1 secs.

Table 5.14: Average computation time

5.3.4 Discussion

- The small variations in the translations and fluctuations that accrue in the gait decrease the recognition performance; however, the sub-sampling methods in the CNN, which is very well known to be translation-invariant, help to overcome these variations. This can be clearly seen in the case of $[bg - bg]$.

- In some cases, the results of subspace methods seem to be poor. These typical dimensionality reduction methods do not take advantage of the higher-order tensors. Thus, they suffer from the lack of training data, curse of dimensionality, and other problems that decrease the recognition performance.

- Our CNN model can extract the discriminative features of gait especially if the available gallery set is large and consist of diverse covariate conditions. However, the recognition performance of the CNN model decreases when the gallery set does not cover the covariate conditions. For example, if the gallery covers only the gait under normal condition and the probe under clothing condition, the recognition performance decreases.

- In deep CNN, the discriminative features of gait are extracted in an automatic way, where there is no handed allocation of the discriminative features. Thus, this makes it more accurate, since it avoids the loss of essential features.

5.3.5 Statistical Power Analysis

Power analysis can be used to calculate the required minimum sample size. We use the power analysis to estimate that our sufficient sample size of the data can attain adequate power. We used the result of our first experiment to conduct the power analysis. The power analysis also allow us to examine the alternative hypothesis. There are two types of statistical hypotheses: the null hypothesis (H0) and the alternative hypothesis (H1). The result of the exam will be: Accept H0 or Reject H0. The statistical Hypothesis for our work is that our model performs better than the best previous method under the case of normal condition. There are several types of tests, we used the two-tailed test.

The sample size (S) is calculated as in 5.11:

$$S = \frac{N * p(1-p)}{[N - 1 * (\frac{d^2}{z^2})] + p(1-p)} \tag{5.11}$$

where d is the effect size, P is the required power (commonly: 80%). z is the t-value of $\alpha/2$. N is the total number of the samples. In our case $N = 248$ $P = 80\%$, $d = 5\%$, z =1.96. The values of P, z and d as well as the defined equations are based on Cohen's standard [88]. The minimum sample size in our case is computed as in equation 5.12

$$S = \frac{248 * 0.8(1-0.8)}{[247 * (\frac{0.05^2}{1.96^2})] + 0.8(1-0.8)} = \frac{39.68}{247 * 0.00065077051 + 0.16} \approx 124 \tag{5.12}$$

According to the result, the sample size of 124 is adequate. Also, it concludes that the sample size we used is higher than the required minimum sample size. The decision of the hypothesis is as follow:

- Mean $\overline{X} = \frac{\sum x}{S}$ = 0.79
- Variance σ^2 = 0.10690
- Standard Deviation (σ) = $\sqrt{\sigma^2}$ = 0.32697
- Significance (α) = 0.05

- Level of confidence (1- α) = 95%

- Critical t = 1.653

- Standard Error (SE) = $\sigma/\sqrt{S}$ = $0.32697/\sqrt{124}$ = 0.02936

- Hypothesis (Ho) = 50%

- t value = $\frac{\overline{X}-Ho}{SE}$= $\frac{0.79-0.050}{0.02936}$ = 25.2019

The H0 is rejected when the t value $>$ Critical t. In our case, the t value (25.202) $>$ critical t (1.653). Thus, it means Reject H0 and Accept H1. Furthermore, the power analysis shows that a sample set of 124 is sufficient and can prove that our proposed method is efficient.

6. CONCLUSIONS

In this dissertation, we have investigated the most discriminative features in the human gait that are less sensitive to covariate conditions. We propose a new method to extract these features to significantly enhance gait recognition performance. We also propose a simple augmentation technique to overcome some of the problems associated with the intra-class gait fluctuations, as well as the small size of the training data. In addition, we use dictionary learning with sparse coding and LDA to seek the best discriminative data representation before feeding it to the Centroid Nearest classifier. Clearly, we have demonstrated the benefits of our method using CASIA-B data set and OU-ISIR Gait dataset B. The results show that our method is able to outperform the existing methods of gait recognition by achieving the highest average result. Also, the proposed method has achieved promising results in the experiments when the viewing angle changes.

Different from the above model, we also developed a specialized deep CNN model, which consists of many layers, for human gait recognition. The advantage of the deep CNN is its ability to extract discriminative features and better classification, especially if the available training dataset is large. We empirically determined the best architecture of the deep CNN for gait recognition. The obtained results on the CASIA-B databases demonstrate better recognition accuracy than existing approaches in several cases. The proposed CNN is capable of overcoming many problems associated with gait recognition especially when covariate factors are involved, and hence leads to better gait recognition performance.

REFERENCES

[1] D. Cunado, M. S. Nixon, and J. N. Carter, “Automatic extraction and description of human gait models for recognition purposes,” *Computer Vision and Image Understanding*, vol. 90, no. 1, 2003, pp. 1–41.

[2] 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, 2004, pp. 1057–1072.

[3] J.-H. Yoo and M. S. Nixon, “Automated markerless analysis of human gait motion for recognition and classification,” *Etri Journal*, vol. 33, no. 2, 2011, pp. 259–266.

[4] A. E. Bobick and A. Y. Johnson, “Gait recognition using static, activity-specific parameters,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, IEEE, vol. 1, 2001, pp. I–423.

[5] R. Zhang, C. Vogler, and D. Metaxas, “Human gait recognition at sagittal plane,” *Image and vision computing*, vol. 25, no. 3, 2007, pp. 321–330.

[6] F. Tafazzoli and R. Safabakhsh, “Model-based human gait recognition using leg and arm movements,” *Engineering applications of artificial intelligence*, vol. 23, no. 8, 2010, pp. 1237–1246.

[7] D. Kim and J. Paik, “Gait recognition using active shape model and motion prediction,” *IET Computer Vision*, vol. 4, no. 1, 2010, pp. 25–36.

[8] L. Wang, T. Tan, W. Hu, and H. Ning, "Automatic gait recognition based on statistical shape analysis," *IEEE transactions on image processing*, vol. 12, no. 9, 2003, pp. 1120–1131.

[9] Z. Liu and S. Sarkar, "Simplest representation yet for gait recognition: Averaged silhouette," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, IEEE, vol. 4, 2004, pp. 211–214.

[10] J. Han and B. Bhanu, "Individual recognition using gait energy image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 2, 2006, pp. 316–322.

[11] L. Chunli and W. KeJun, "A behavior classification based on enhanced gait energy image," in *Networking and Digital Society (ICNDS), 2010 2nd International Conference on*, IEEE, vol. 2, 2010, pp. 589–592.

[12] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan, "Human identification using temporal information preserving gait template," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, 2012, pp. 2164–2176.

[13] Y. Liu, J. Zhang, C. Wang, and L. Wang, "Multiple hog templates for gait recognition," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, IEEE, 2012, pp. 2930–2933.

[14] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian, "Frame difference energy image for gait recognition with incomplete silhouettes," *Pattern Recognition Letters*, vol. 30, no. 11, 2009, pp. 977–984.

[15] A. Roy, S. Sural, and J. Mukherjee, "Gait recognition using pose kinematics and pose energy image," *Signal Processing*, vol. 92, no. 3, 2012, pp. 780–792.

[16] E. Zhang, Y. Zhao, and W. Xiong, "Active energy image plus 2dlpp for gait recognition," *Signal Processing*, vol. 90, no. 7, 2010, pp. 2295–2302.

[17] G. V. Veres, L. Gordon, J. N. Carter, and M. S. Nixon, “What image information is important in silhouette-based gait recognition?” In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, IEEE, vol. 2, 2004, pp. II–776.

[18] B. Khalid, X. Tao, and G. Shaogang, “Gait recognition using gait entropy image,” in *3rd International Conference on Crime Detection and Prevention (ICDP 2009)*.

[19] C. P. Lee, A. W. Tan, and S. C. Tan, “Gait probability image: An information-theoretic model of gait representation,” *Journal of Visual Communication and Image Representation*, vol. 25, no. 6, 2014, pp. 1489–1492.

[20] N. Li, Y. Xu, and X.-K. Yang, “Part-based human gait identification under clothing and carrying condition variations,” in *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, IEEE, vol. 1, 2010, pp. 268–273.

[21] M. Jeevan, N. Jain, M. Hanmandlu, and G. Chetty, “Gait recognition based on gait pal and pal entropy image,” in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, IEEE, 2013, pp. 4195–4199.

[22] Y. Chai, Q. Wang, R. Zhao, and C. Wu, “A new automatic gait recognition method based on the perceptual curve,” in *TENCON 2005 2005 IEEE Region 10*, IEEE, 2005, pp. 1–5.

[23] Z. Liu and S. Sarkar, “Improved gait recognition by gait dynamics normalization,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, 2006, pp. 863–876.

[24] X. Huang and N. V. Boulgouris, “Gait recognition with shifted energy image and structural feature extraction,” *Image Processing, IEEE Transactions on*, vol. 21, no. 4, 2012, pp. 2256–2268.

[25] T. Guha and R. Ward, “Differential radon transform for gait recognition,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, IEEE, 2010, pp. 834–837.

[26] Y. Guan, C.-T. Li, and F. Roli, "On reducing the effect of covariate factors in gait recognition: A classifier ensemble method," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 7, 2015, pp. 1521–1528.

[27] T. Whytock, A. Belyaev, and N. M. Robertson, "Dynamic distance-based shape features for gait recognition," *Journal of Mathematical Imaging and Vision*, vol. 50, no. 3, 2014, pp. 314–326.

[28] S. D. Choudhury and T. Tjahjadi, "Gait recognition based on shape and motion analysis of silhouette contours," *Computer Vision and Image Understanding*, vol. 117, no. 12, 2013, pp. 1770–1785.

[29] S. Choudhury and T. Tjahjadi, "Silhouette-based gait recognition using procrustes shape analysis and elliptic fourier descriptors," *Pattern Recognition*, vol. 45, no. 9, 2012, pp. 3414–3426.

[30] K. Bashir, T. Xiang, S. Gong, and Q. Mary, "Gait representation using flow fields.," in *BMVC*, 2009, pp. 1–11.

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

[32] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision &amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, IEEE, vol. 1, 1994, pp. 582–585.

[33] M. Hu, Y. Wang, Z. Zhang, D. Zhang, and J. J. Little, "Incremental learning for video-based gait recognition with lbp flow," *Cybernetics, IEEE Transactions on*, vol. 43, no. 1, 2013, pp. 77–89.

[34] W. Kusakunniran, "Recognizing gaits on spatio-temporal feature domain," *Information Forensics and Security, IEEE Transactions on*, vol. 9, no. 9, 2014, pp. 1416–1423.

[35] K. Bashir, T. Xiang, and S. Gong, “Feature selection for gait recognition without subject cooperation.,” in *BMVC*, Citeseer, 2008, pp. 1–10.

[36] Y. Dupuis, X. Savatier, and P. Vasseur, “Feature subset selection applied to model-free gait recognition,” *Image and vision computing*, vol. 31, no. 8, 2013, pp. 580–591.

[37] I. Rida, X. Jiang, and G. L. Marcialis, “Human body part selection by group lasso of motion for model-free gait recognition,” 2016.

[38] T. Whytock, A. Belyaev, and N. M. Robertson, “On covariate factor detection and removal for robust gait recognition,” *Machine Vision and Applications*, vol. 26, no. 5, 2015, pp. 661–674.

[39] Y. Iwashita, K. Uchino, and R. Kurazume, “Gait-based person identification robust to changes in appearance,” *Sensors*, vol. 13, no. 6, 2013, pp. 7884–7901.

[40] M. Shariful Islam, A. Matin, J. Paul, M. Rokanujjaman, and M. Altab Hossain, “A new effective part selection approach for part-based gait recognition,” in *Computer and Information Technology (ICCIT), 2013 16th International Conference on*, IEEE, 2014, pp. 181–184.

[41] S. H. Shaikh, K. Saeed, and N. Chaki, “Gait recognition using partial silhouette-based approach,” in *Signal Processing and Integrated Networks (SPIN), 2014 International Conference on*, IEEE, 2014, pp. 101–106.

[42] D. Xu, S. Yan, L. Zhang, S. Lin, H.-J. Zhang, and T. S. Huang, “Reconstruction and recognition of tensor-based objects with concurrent subspaces analysis,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 1, 2008, pp. 36–47.

[43] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, “Discriminant analysis with tensor representation,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, IEEE, vol. 1, 2005, pp. 526–532.

[44] D. Xu, S. Yan, D. Tao, L. Zhang, X. Li, and H.-J. Zhang, "Human gait recognition with matrix representation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, no. 7, 2006, pp. 896–903.

[45] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 10, 2007, pp. 1700–1715.

[46] Z. Lai, Y. Xu, Z. Jin, and D. Zhang, "Human gait recognition via sparse discriminant projection learning," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24, no. 10, 2014, pp. 1651–1662.

[47] X. He and P. Niyogi, "Locality preserving projections," in *NIPS*, vol. 16, 2003.

[48] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of machine learning research*, vol. 8, no. May, 2007, pp. 1027–1061.

[49] M. Hu, Y. Wang, Z. Zhang, J. J. Little, and D. Huang, "View-invariant discriminative projection for multi-view gait-based human identification," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 12, 2013, pp. 2034–2045.

[50] Y. Huang, D. Xu, and T.-J. Cham, "Face and human gait recognition using image-to-class distance," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 3, 2010, pp. 431–438.

[51] G. Zhao, G. Liu, H. Li, and M. Pietikäinen, "3d gait recognition using multiple cameras," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, IEEE, 2006, pp. 529–534.

[52] R. Bodor, A. Drenner, D. Fehr, O. Masoud, and N. Papanikolopoulos, "View-independent human motion classification using image-based reconstruction," *Image and Vision Computing*, vol. 27, no. 8, 2009, pp. 1194–1206.

[53] K. Bashir, T. Xiang, and S. Gong, "Cross view gait recognition using correlation strength.," in *BMVC*, 2010, pp. 1–11.

[54] C.-S. Lee and A. Elgammal, "Towards scalable view-invariant gait recognition: Multilinear analysis for gait," in *Audio-and Video-Based Biometric Person Authentication*, Springer, 2005, pp. 395–405.

[55] N. Liu, J. Lu, G. Yang, and Y.-P. Tan, "Robust gait recognition via discriminative set matching," *Journal of Visual Communication and Image Representation*, vol. 24, no. 4, 2013, pp. 439–447.

[56] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition under various viewing angles based on correlated motion regression," *IEEE transactions on circuits and systems for video technology*, vol. 22, no. 6, 2012, pp. 966–980.

[57] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, Feb. 2017, pp. 209–226.

[58] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, 2017.

[59] W. Zeng and C. Wang, "View-invariant gait recognition via deterministic learning," in *Neural Networks (IJCNN), 2014 International Joint Conference on*, IEEE, 2014, pp. 3465–3472.

[60] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[61] E. P. Ijjina and C. K. Mohan, "One-shot periodic activity recognition using convolutional neural networks," in *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*, IEEE, 2014, pp. 388–391.

[62] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

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

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

[65] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 12, 2003, pp. 1505–1518.

[66] D. Tan, K. Huang, S. Yu, and T. Tan, "Efficient night gait recognition based on template matching," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, IEEE, vol. 3, pp. 1000–1003.

[67] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 2, 2005, pp. 162–177.

[68] J. D. Shutler, M. G. Grant, M. S. Nixon, and J. N. Carter, "On a large sequence-based human gait database," in *Applications and Science in Soft Computing*, Springer, 2004, pp. 339–346.

[69] Y. Makihara, H. Mannami, A. Tsuji, M. A. Hossain, K. Sugiura, A. Mori, and Y. Yagi, "The ou-isir gait database comprising the treadmill dataset," *IPSJ Transactions on Computer Vision and Applications*, vol. 4, 2012, pp. 53–62.

[70] A. Kale, A. Sundaresan, A. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Krüger, and R. Chellappa, "Identification of humans using gait," *Image Processing, IEEE Transactions on*, vol. 13, no. 9, 2004, pp. 1163–1173.

[71] L. Lee and W. E. L. Grimson, "Gait analysis for recognition and classification," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, IEEE, 2002, pp. 148–155.

[72] R. Gross and J. Shi, “The cmu motion of body (mobo) database,” 2001.

[73] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, “The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition,” *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 5, 2012, pp. 1511–1521.

[74] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, “Robust view transformation model for gait recognition,” in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, IEEE, 2011, pp. 2073–2076.

[75] F. Chollet, *Keras*, `https://github.com/fchollet/keras`, 2015.

[76] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–2830.

[77] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine learning*, vol. 63, no. 1, 2006, pp. 3–42.

[78] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online dictionary learning for sparse coding,” in *Proceedings of the 26th annual international conference on machine learning*, ACM, 2009, pp. 689–696.

[79] A. Andoni and P. Indyk, “Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions,” in *Foundations of Computer Science, 2006. FOCS’06. 47th Annual IEEE Symposium on*, IEEE, 2006, pp. 459–468.

[80] K. Bashir, T. Xiang, and S. Gong, “Gait recognition without subject cooperation,” *Pattern Recognition Letters*, vol. 31, no. 13, 2010, pp. 2052–2060.

[81] P. Yogarajah, J. V. Condell, and G. Prasad, “P rw gei: Poisson random walk based gait recognition,” in *Image and Signal Processing and Analysis (ISPA), 2011 7th International Symposium on*, IEEE, 2011, pp. 662–667.

[82] X. Li and Y. Chen, “Gait recognition based on structural gait energy image,” *Journal of Computational Information Systems*, vol. 9, no. 1, 2013, pp. 121–126.

[83] I. Rida, S. Almaadeed, and A. Bouridane, "Gait recognition based on modified phase-only correlation," *Signal, Image and Video Processing*, vol. 10, no. 3, 2016, pp. 463–470.

[84] W. Kusakunniran, "Attribute-based learning for gait recognition using spatio-temporal interest points," *Image and Vision Computing*, vol. 32, no. 12, 2014, pp. 1117–1126.

[85] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks.," in *Aistats*, vol. 9, 2010, pp. 249–256.

[86] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1563–1572.

[87] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, IEEE, 2010, pp. 974–981.

[88] J. Cohen, *Statistical power analysis for the behavioral sciences 2nd ed*. Routledge, 1988, p. 400.