Multi-Biometric Person Identification System Using Face and Gait Fusion: A Deep Learning Approach

A Dissertation Submitted in Fulfillment of the Requirement for the Award of the Degree

of

Master of Engineering

in

Electronics and Communication Engineering

Submitted by

Ankit Sharma

Roll No: 801661003

Supervisor

Dr. Neeru Jindal

Assistant Professor, ECED

ELECTRONICS AND COMMUNICATION ENGINEERING DEPARTMENT

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY (A DEEMED TO BE UNIVERSITY), PATIALA, PUNJAB

JULY, 2018

DECLARATION

I, Ankit hereby declare that the work presented in this thesis entitled "**Multi-Biometric Person Identification System Using Face and Gait Fusion: A Deep Learning Approach**" in fulfillment of the requirement for the award of degree of Master of Engineering (ECE) submitted at Electronics and Communication Engineering Department, T.I.E.T., Patiala is an authentic record of work carried out under supervision of **Dr. Neeru Jindal** (Assistant Professor), Electronics and Communication Department, T.I.E.T., Patiala from July 2017 to July 2018.

The matter presented in this thesis has not been submitted either in part or full to any other university or institute for the award of any other degree.

Date: _____

Ankit Sharma

Roll no: 801661003

It is certified that the above statement made by the candidate is correct to the best of my knowledge and belief.

Date: 9/7/18

Dr. Neeru Jindal

Assistant Professor, ECED

T.I.E.T. (A Deemed to be University)

Patiala, Punjab

ACKNOWLEDGEMENT

First of all, I am very thankful to the all almighty GOD for honoring me with great knowledge, intelligence, well-being, cognizance and the brainpower to conduct this research successfully.

I convey my sincere thanks to my supervisor **Dr. Neeru Jindal**, Assistant Professor, ECED, T.I.E.T., Patiala who has permitted me to work under her guidance and supported me throughout the thesis research. I take this opportunity to convey my deepest gratitude to her for her valuable advice, comfort encouragement, keen interest, constructive criticism, scholarly guidance & wholehearted support.

I express my sincere thanks to **Dr. Alpana Agarwal**, Head of the Department, and **Dr. Amit Mishra**, Program Coordinator, ECE Department, T.I.E.T., Patiala for providing me a great learning environment and infrastructure in ECED.

I am also thankful to my friends and my family who were always with me in my good and bad times to encourage and support me.

In the end, I would like to express my deepest gratitude to all those people who directly or indirectly supported me throughout my research process.

Ankit Sharma

ABSTRACT

Automatic authentication of people has always been a challenging task especially when it has to deal with the large datasets and the robustness against the factors affecting recognition such as pose variation, subject to camera angle, illumination, poor quality data and occlusion etc. Neural networks in particular have become very popular nowadays because of two things i.e. faster computers and large datasets. So, deep learning and neural networks prove to be a great remedy for above problems. Hence, we have designed an algorithm to identify people at a distance by fusing their gait and face biometrics using a 13-Layer Deep Convolutional Neural Network (DCNN). We utilized the concept of Gait Energy Images (GEIs) to represent the characteristics of human gait. The GEIs and face sequences from same individual are firstly resized into vectors after some sort of pre-processing. Then, both the vectors are fused and the output is fed to the DCNN for feature extraction and classification. Our proposed DCNN is composed of three triplets of Convolution, ReLU and Max-Pooling Layers followed by a Fully Connected Layer, a SoftMax Regression Layer and a Classification Layer. The proposed work is tested upon three publicly available databases i.e. CASIA Gait B, ORL Face, and FEI Face Datasets. A maximum accuracy of 98.75 % is achieved when ORL Face Database is fused with CASIA Gait B Database and 97.50 % accuracy is achieved when FEI Face Dataset is fused with CASIA Gait B Database.

We have also tested our model with three noise attacks to both face and gait test images i.e. Salt and Pepper, Gaussian and Speckle Noises. We utilized the median filter to de-noise the images affected with Salt and Pepper noise and mean filter to de-noise the images affected with Gaussian Noise and Speckle Noise both. A recognition accuracy of 97.5 %, 93.75 % and 95 % for the first experiment and 97.18 %, 95.97 % and 95.56 % for the second experiment is achieved in presence of Salt and Pepper, Gaussian and Speckle Noise respectively. We were also able to reduce the computational time as our model took only 3.5

minutes to train the network for 1st experiment and around 11 minutes for 2nd experiment. To further increase the recognition accuracy in future, we will try to combine our proposed model with some of the existing popular feature fusion algorithms such as Canonical Correlation Analysis (CCA) and Discriminant Correlation Analysis (DCA).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
ATM	Automated Teller Machine
CASIA	Chinese Academy of Sciences Institute of Automation
CCA	Canonical Correlation Analysis
CCD	Charge Coupled Device
CNN	Convolutional Neural Network
CV	Computer Vision
DCA	Discriminant Correlation Analysis
DCNN	Deep Convolutional Neural Network
DNA	Deoxyribonucleic Acid
FRS	Face Recognition System
GEI	Gait Energy Image
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
ID	Identification
K-NN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
MDA	Multiple Discriminant Analysis
MHI	Motion History Image
MOS	Mean Opinion Score
NMF	Non-negative Matrix Factorization
ORL	Operations Research Laboratory
PCA	Principal Component Analysis
PGM	Portable Gray Map
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristics
SOM	Self-Organizing Map
SVM	Support Vector Machine
VGG	Visual Geometry Group

CHAPTER 1
INTRODUCTION

1.1 PREFACE

The motivation for carrying out this research work initially originated because of my affectionate interest for image processing, biometric security and developing better methods of Person Identification and biometric fusion for biometric security. The need for a solid individual recognition system in mechanized access control has brought about an expanded enthusiasm for biometrics [1]. With the growth of the digital age, the issue of biometric security is becoming the centre of attraction for a lot of researchers. Person identification is one of the regions from Computer Vision (CV) that has drawn more enthusiasm for long. In spite of the fact that a considerable measure of work has just been directed in the field of individual identification, yet a less work has been done in this field by joining biometric combination with deep learning. Deep learning is the most demanding concept in almost each and every task in today's environment but only a few researchers have utilized the deep learning algorithms for fusion of multimodal biometrics for the purpose of person identification. My passion towards image processing and my concerns for biometric security and automatic person authentication encouraged me to carry out my thesis work in this field. Henceforth, we have melded the face and gait biometric qualities for individual ID utilizing a 13-layer deep convolution neural system.

1.2 BIOMETRICS RECOGNITION

The science and innovation of measuring and examining natural information for validation or ID intention is known as Biometrics. A framework that effectively recognizes and extracts the organic information out of an individual's body is termed as a biometric framework. Biometric Systems are mechanized strategies for confirming or perceiving the personality of a person on the premise of some physical highlights or appearances, similar to a unique finger impression or face example or a few characteristics of conduct, such as penmanship or keystroke designs. In the present howdy tech world, there is a regularly developing need to verify and distinguish individuals for security purposes. The biometric recognizable proof technique comprises of three operations, initially the biometric information is extracted from

the body of an individual and then the features are extracted from this biometric data which is stored in a template as feature sets. In the end, the set of extracted features stored in the template are matched against the set of features already stored in the database as training data by means of a matcher or a classifier. The outcomes decide whether the individual is recognized or not. We can utilize both types of features i.e. behavioural and physiological characteristics for identification using a biometric system [2]. The biometric information which can be extracted from the individual's body parts directly is termed as the physiological biometrics. The examples of physiological biometrics are such as DNA, saliva, hand and palm geometry, iris and fingerprints etc. And the biometric information which does not directly relate to an individual's body and is related to actions or gestures is termed as the behavioural biometrics. For examples, the gait biometrics is a behavioural biometrics which we have utilised in our proposed work to fuse with the face biometrics. Figure 1.1 below shows the block diagram of a basic biometric system in which the biometric data is first of all extracted from individual's body using some kind of sensors and then this information goes through pre-processing phase. After that, the feature extractor extracts the features from biometric data and stores them into a template as feature sets which are later matched with the already stored feature sets using some matching classifiers. Powerful techniques in the extraction of features and classification strategies for the separated features are the key factors in some genuine pattern recognition and classification assignments [3].

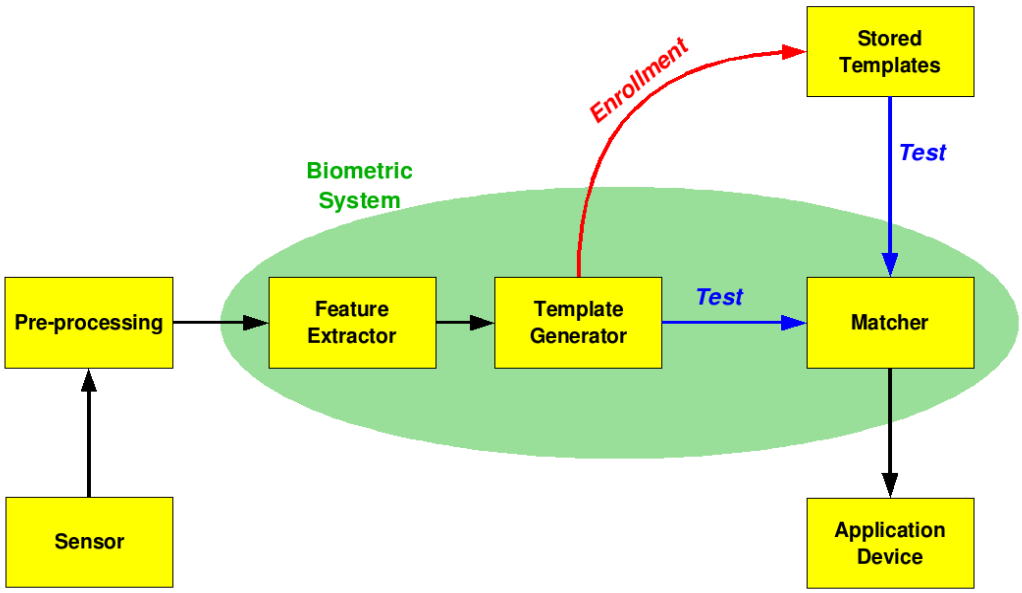


Fig. 1.1 Basic Biometric System [4]

1.2.1 Types of Biometrics

Biometrics classification can also be done according to the fact that how is the involvement of an individual in the identification process. On this basis, the classification of the biometrics can be done into two categories that are Soft biometrics and Hard Biometrics.

1.2.1.1 Hard Biometrics

This kind of biometrics permit the acknowledgment of individuals through their own morphological highlights like face, fingerprints, iris, palm print; organic highlights like DNA, spit; and behavioural highlights like sign, hand composing, voice and so forth. This sort of biometrics requires the individual's co-operation for the recognizable proof reason. It implies the acknowledgment of individual is impossible at a separation without the client's co-operation. For example: Fingerprint, Face, Palm print, Iris texture. Palm veins, Retina, voice/speech, Hand geometry, Signature/handwriting, Odour/scent, DNA and Saliva etc. [5].

1.2.1.2 Soft Biometrics

Soft biometrics is the kind of biometrics which permits the acknowledgment of individuals at a separation without the client's co-operation. They take care of the issue of extra obtaining expense and time of hard biometrics. A remote people acknowledgment framework can be composed utilizing soft face modalities like skin shading, hair shading, facial estimations, eye shading, age, sexual orientation and so forth. Expansion of soft biometrics to a video reconnaissance framework expands the execution of acknowledgment assignment and increment the exactness of the framework. Some of the examples are: Skin colour, Iris colour, Eye colour, Shape and size of head, Presence of beard/moustache, Birth marks/scars/tattoos, Height/weight, Gait biometrics, Accessories like glasses, hat etc., Ethnicity, Clothing, Age, Gender, Eye brows and Hair colour etc. [5].

1.2.2 Operational Modes

On the basis of type of applications, there are basically two modes on which a biometric identification framework work. First one is the enrolment mode and the second one is either an identification mode or verification mode. In the enrolment mode, the biometric data is first

of all extracted from an individual's body using some kind of sensors and then after some pre-processing, the feature extractor extracts the features from biometric data and stores them into a template as feature sets for later use. The stored biometric information in the template database is later used for matching purpose in either identification mode or verification mode [6]. The basic block diagram for understanding the operation of enrolment mode in a biometric system is given in figure 1.2 below.

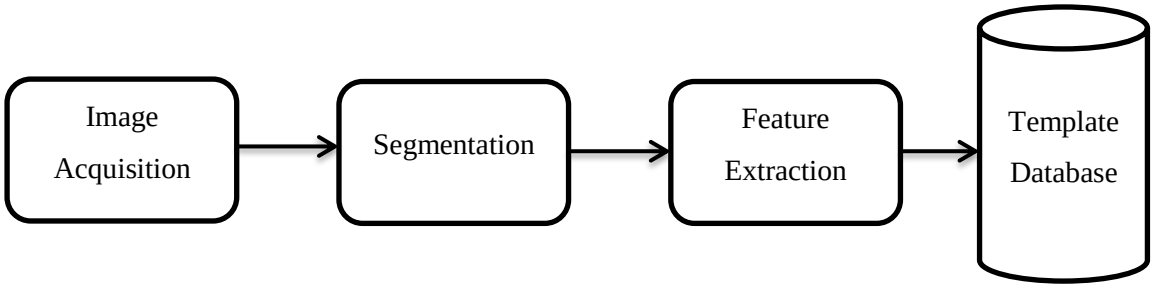


Fig. 1.2 The Enrolment Mode in biometric systems

In identification mode, the right ID of an obscure individual chosen from template of enrolled characters. This type of mode is also known as the 'one to many matching' operational mode as here, the biometric framework is requested to finish a correlation between the individual's features and all the features already present in the template of enrolled features. The framework can choose either the most relevant match or it may select all the conceivable matches, and arrange them according to comparability. It is further classified as positive and negative identification modes. The positive mode has a tendency to decide whether a given individual is truly present in a particular template. The negative mode decides whether a given individual isn't present in a particular template. The figure 1.3 below shows the basic operation of the identification mode.

In verification mode, the biometric system verifies whether an individual is really the same that is present in the already stored template. This type of mode is also known as the 'one to one matching' operational mode as here the framework needs to finish a correlation between the individual's features and just a single feature set present in the template of enrolled features. This type of strategy is utilizes if the objective is to anchor and limit particular access with particular clients. The figure 1.4 below shows the basic operation of the verification mode.

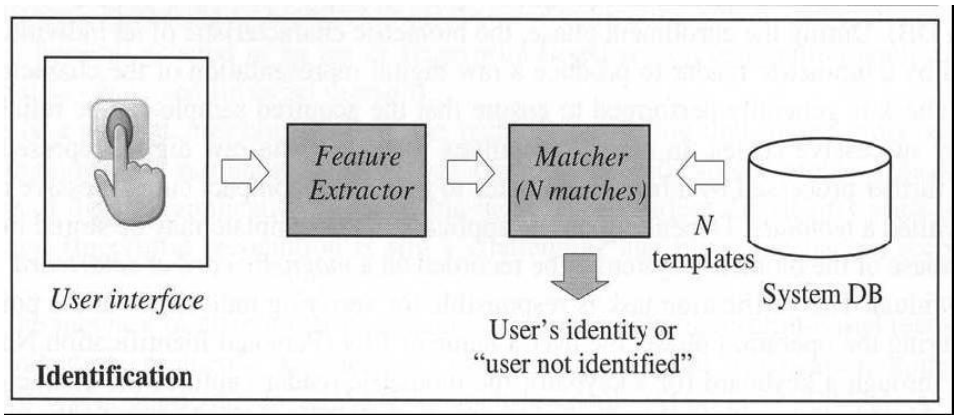


Fig. 1.3 The Identification Mode in biometric systems [7]

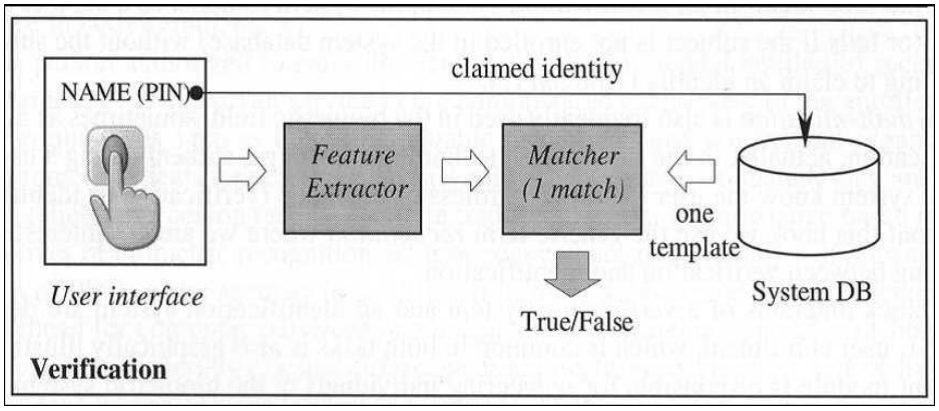


Fig. 1.4 The Verification Mode in biometric systems [7]

1.3 APPROACHES FOR MULTI-BIOMETRIC FUSION

A framework which depends upon the availability of numerous bits of biometric traits for individual recognition is known as multi-biometric framework. Past experiments demonstrate that single biometric frameworks contain numerous inconveniences with respect to execution and precision. Multi-biometric frameworks perform superior to single biometric frameworks and are more prominent than them. They can conquer the limitations of unimodal biometric systems and provide better arrangement precision. Multi-biometric framework can join any number of biometric traits for fusion and are also robust to the common spoofing attacks. In multi-biometric frameworks if any one of the techniques fails to do its job, it is not going to influence truly the individual recognition because we can utilize other available techniques. Subsequently, the spoof attacks can be limited radically in this way enhancing the proficiency

of the general framework. There are a variety of approaches for fusion of multiple biometric traits for individual recognition; some of them are described as under [8].

1.3.1 Sensor Level Fusion

This type of approach can't combine multiple biometric traits as a result of incongruence of information. There are two ways in which sensor level fusion can be carried out; the first one is the where the numerous samples of a single biometrics are obtained and combined utilizing a single sensor and the second one is where the numerous samples of a single biometrics are obtained and combined utilizing numerous sensors.

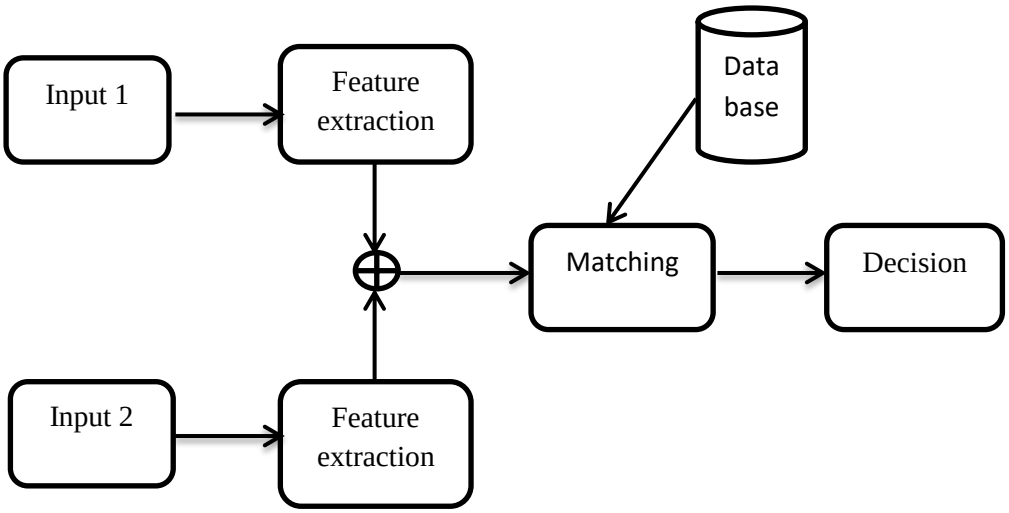


Fig. 1.5 Feature Level Fusion

1.3.2 Feature level Fusion

Pre-processing of the inputs originating from various biometric channels is carried out initially which is followed by extraction of features independently utilizing a particular technique and then we join these vectors to generate a composite set of feature vectors. Information from a sensor processes an element vector. As we know that biometric attribute's highlights are distinct to one another, it is sensible to connect two vectors to form the new one. Various algorithms are utilized for the reduction of feature dimensions to get the important feature sets from bigger ones. The figure 1.5 above shows the block diagram for feature level fusion approach.

1.3.3 Matching Score Level Fusion

Here in this type of fusion method, each of the frameworks gives a matching score demonstrating the feature set's vicinity with the feature sets stored in template database. These scores are joined to affirm the identity. Instead of consolidating the features, they are processed independently and then the match score is discovered at that point on the basis of exactness of each of the modality's match score that is later utilized to classify the results. It is further classified into three classifications i.e. Classifier based fusion, Transformation based fusion and Density based fusion. The figure 1.6 below shows the block diagram for score level fusion approach.

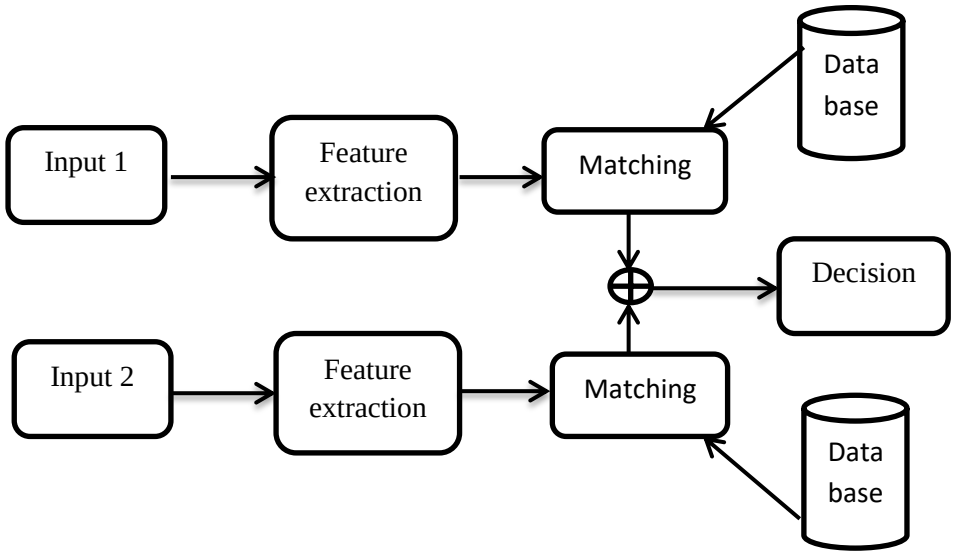


Fig. 1.6 Matching Score Level Fusion

1.3.4 Decision level fusion

In this kind of fusion approach, each of the sensor catches different modality information and then the set of features are separately divided in 2 classes i.e. either acknowledge or dismiss. The final choice is made by utilizing the majority vote scheme. This type of fusion approach utilizes "OR" and "AND" decisions rules combined with majority voting technique for making final decision. The figure 1.7 below shows the block diagram for decision level fusion approach.

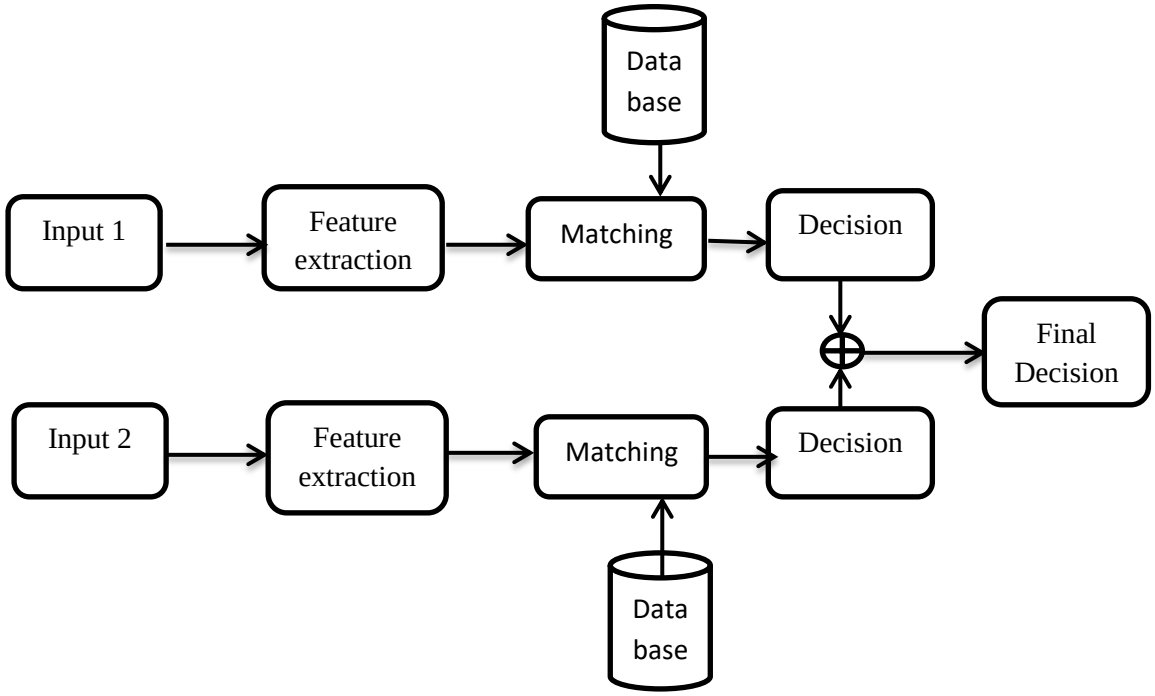


Fig. 1.7 Decision Level Fusion

1.3.5 Rank Level Fusion

This type of fusion approach is utilized to enhance the execution of the multi-biometric framework, in which rank are acquired out of at least two recognition outcomes such that the recognition results can be enhanced. The rank level fusion approach has been utilized in a limited way by the research scholars.

1.4 DEEP NEURAL NETWORKS

There are like a million of machine learning algorithms out there, but neural networks in particular have become very popular recently because of two things i.e. faster computers and large datasets. They have produced some amazing breakthroughs from image recognition to generating rap songs. A neural system is an inter-connected gathering of basic handling components or hubs, functioning of which is completely depended on the term neuron. This system directly relates to biology. In any case, comparative idea can be actualized with software vision too. Human mind comprises of around 100 billion neurons. All of them link to each other with the help of electrical pulses, which may be termed as the driving forces for them. The inter-neuron associations are interceded by electro-chemical intersections known

as synapses situated on branches of the dendrites. Every neuron normally gets a huge number of associations from different neurons and is subsequently always getting a large number of approaching pulses, which in the long run receive at the cell body. After that, they all go through some process where if some of them surpass some limit esteems then the neuron will fire. It is transferred to different neurons by means of branch fibre known as axon.

There are mainly three steps involved in machine learning i.e. build it, train it and test it. Once we build our model we can train it against our input and output data to make it better and better. If we utilise a neural network for making predictions, that contains not only one or two deep layers but has many deep layers, , it is known as deep learning [9]. Hence deep learning is a part of machine learning that is outperforming most of the other approaches, for a variety of purposes.

1.4.1 Types of Neural Networks

In 1943, two early computer scientists named Mc Culloch and Pitts invented the 1st computational model of a neuron which was a simple model but in early days of artificial intelligence that was a big deal. A few years later a physiologist named Rosenblatt build a model called Perceptron which is another word for a single layer feed forward neural network. Then later in 2006, G. E. Hinton et al. presented an autoencoder network which could reduce the dimensionality of any data by effectively initialising the weights [10]. This provided more efficient results than PCA. In case of neural networks, there exist various types of system architectures with distinctive features and characteristics. In real classification, there are two main networks i.e. recurrent networks and feed forward networks. There are some other popular neural networks which we are going to discuss in upcoming chapters such as back propagation neural network and convolutional neural network.

1.4.1.1 Feed Forward Neural Networks

In these types of networks, unidirectional links are present with absence any cycle. A feed-forward system registers an element of the info esteems which relies upon the weight characteristics; it does not include any interior state accepts for weights. These types of systems can actualize versatile adaptations of basic reflex operators or they can work as segments of more complicated specialists. The perceptron is also a feed forward neural

network which does not include any number of hidden layers. The learning becomes extremely easy utilizing these types of systems. Systems containing at least one hidden layer are termed as multi-layer systems. Perceptron is kind of layered feed forward system. Presently, feed forward systems with single layers are otherwise called perceptron. The figure 1.8 below shows a simple feed forward neural network containing only one hidden layer having three neurons.

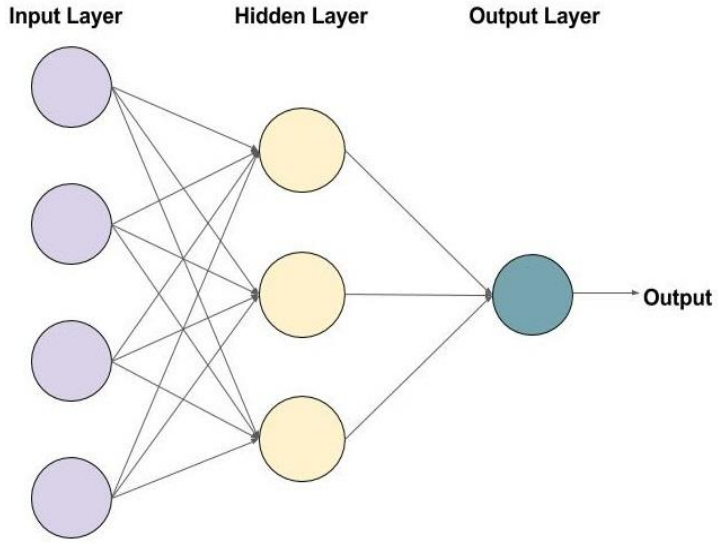


Fig. 1.8 A Feed Forward Neural Network [19]

1.4.1.2 Recurrent Neural Networks

In these types systems, connections can make discretionary topologies. Human mind could also be taken as an example of recurrent systems for instance. In recurrent neural networks, the activation is given back to the cells that generated it. These systems contain the interior state put away in activation levels of particular cells. In recurrent neural networks, the calculation is more complex than the feed forward neural systems. These networks can end up temperamental and learning is more troublesome. Hopfield systems are the best examples of this type of system. The figure 1.8 below shows a simple recurrent neural network.

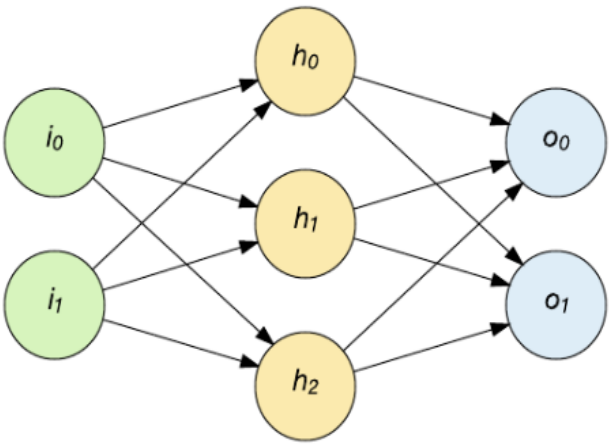


Fig. 1.9 A Recurrent Neural Network [24]

1.4.1.3 Back Propagation Neural Networks

Learning is applicable to a multi-layered feed forward systems like a 1- layered feed forward system also known as perceptron. The well-known technique for this purpose in case of a multi-layered feed forward system is known as back propagation algorithm. This approach partitions the contribution for all the weights. Like perceptron learning approach, back propagation endeavour to reduce the loss between each objective yield and the yield really processed by the system.

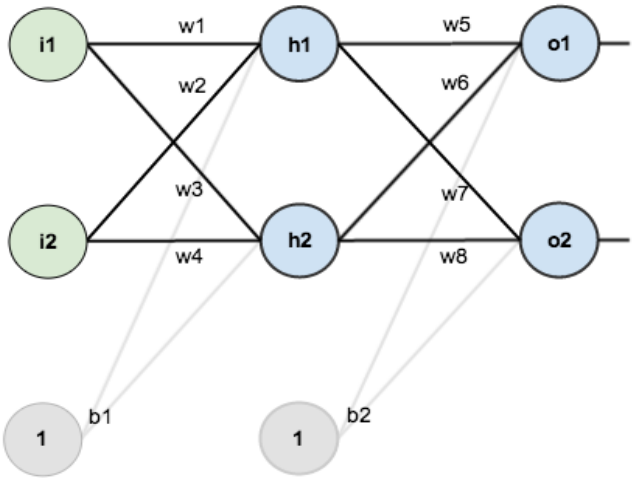


Fig. 1.10 A Back Propagation Neural Network [28]

First of all, the inputs are displayed to the system, and if the system registers a yield vector that match the objective, then no need to do anything. On the off chance that error is present, and then a difference of the yield and target is computed according to which the system adjusts the weights and hence decreases the error rate. The figure 1.10 above shows a simple back propagation neural network showing the weight mechanism for error reduction.

1.4.1.4 Convolutional Neural Networks (CNNs)

A convolutional neural network may contain a huge number of hidden layers and every layer has a capability to extract some features in a picture. Working of a CNN depends on three key point's i.e. local receptive fields; shared weights and biases; and activation and pooling [30]. The local receptive field is translated over an image to generate the feature maps from input layer to hidden layer neurons and for this purpose convolution layer is utilised. In case of CNNs the weights and bias values are same for all hidden layers that mean the network detects the same features in different regions of image. 3rd most important concept for this network is activation and pooling. The activation step applies the transformation to the output of each neuron by using an activation function. The output of the activation step is further transformed by applying a pooling step which reduces the dimensionality of feature maps by condensing the output of small regions of neurons into a single output.

1.5 APPLICATIONS OF BIOMETRIC RECOGNITION

Biometric recognition is a quickly developing innovation is generally utilized as a part of legal sciences, for example, crime detection, ATM's and jail security and it can possibly be utilized as a part of an expansive scope of regular citizen application purposes. These frameworks are being utilized widely in different areas. Some of those applications are discussed below.

1.5.1 Commercial Sector

Business uses of biometrics are to a great extent utilized by financial and banking administration divisions. Other most important areas in commercial sector where biometric systems are utilized are as follow:

- Account Access

- ATM’S Security
- Expanded Service Kiosks
- Online banking
- Telephonic transaction
- Personal Computer/Network Security
- E-Commerce

1.5.2 Government Sector

There exist various uses of biometric systems in the government driven projects and services. Most important of them where biometric identification plays an important role are for example the National ID cards and AADHAR cards. Some of the main application areas of biometric systems in government sector are as follow:

- National Identification Cards
- Voter ID and Elections
- Driver's licenses
- AADHAR Cards
- Benefits Distribution (social service)
- Employee Authentication

1.5.3 Health Care

Utilization of biometrics mat as a part of medicinal services for enrolment of the patients when they are admitted is the foremost application of biometric systems in this field. Patients can also make their transactions utilizing their biometric characteristics. It surely avoids misrepresentation instances of hospitalization and frauds of medications. Some other important applications in medical field are as follow:

- Patient identification
- Access to personal information of patients
- PC/Network access of hospital employees

1.5.4 Other Areas of Applications

There are some other important areas where biometric authentication frameworks play a vital role out of which detection of crimes and frauds is the most important. These are as follow:

- Forensics
- Military Programs
- Employee Authentication
- Fraud Detection

1.6 THESIS OUTLINE

This dissertation is divided into six chapters and rest of the elements of the dissertation are organized in the following manner:

CHAPTER 1

This chapter begins with the introduction about biometric recognition and approaches for multi-biometric fusion. After that a brief introduction about the deep neural network is given followed by the various applications of biometric recognition.

CHAPTER 2

Chapter 2 presents the literature review of some existing multi-biometric recognition methods which includes deep learning based face-recognition techniques, deep learning based gait-recognition techniques and multi-biometric fusion techniques. In the end motivation and research objectives are defined for the dissertation.

CHAPTER 3

This chapter provides an overview of the proposed methodology which includes detailed introduction about the deep convolutional neural network and all its layers such as convolutional, ReLU, max-pooling, fully connected and softmax layers.

CHAPTER 4

This chapter presents the work proposed in the thesis in details including the description of the database, proposed 13-layer DCNN architecture, training methodology, experiments conducted and their corresponding results on different database.

CHAPTER 5

This chapter presents some common noise attacks that can occur in the digital images such as salt and pepper noise, Gaussian noise and speckle noise. Experimental results for the proposed DCNN model are presented against all the three noise attacks mentioned above.

CHAPTER 6

This chapter summarizes the whole work conducted in this dissertation and in the end the possible future aspects are defined to further extend the thesis work.

CHAPTER 2
LITERATURE REVIEW

2.1 INTRODUCTION

With the rapid advancement in the field of technology and equipment, the area of individual recognition has gained interest of a lot of researchers in the past few years. The expanding requirements for video surveillance systems in business, law and military applications make the person identification architectures as one of the primary current application areas in the field of computer vision. But with the technology, the crimes are also increasing at the same pace hence; we require more powerful and robust security systems to get rid of this. Deep learning and neural networks come up with the solution for this and hence here we have fused the face and gait biometric traits in our work using a deep convolution NN. Though a lot of work has already been conducted in the field of person identification but a very less work has been done in this field by combining biometric fusion with deep learning. To the best of my knowledge, this is the first approach where deep learning is utilized in the biometric fusion of these two traits i.e. face and gait. This chapter presents the literature review of some prior work conducted in the field of person identification using multimodal biometric systems which includes deep learning based face-recognition techniques, deep learning based gait-recognition and some multi-biometric fusion techniques with or without utilizing deep learning approach.

2.2 DEEP LEARNING BASED FACE-RECOGNITION

Steve Lawrence *et al.* [1] in 1997 exhibited a hybrid neural-system that joins the picture sampling technique with a convolution neural system and a SOM neural system. SOM neural network first quantize the sample of pictures in a geological space in which the adjacent input values come closer in the yield space. In this way, it provides reduction in feature dimensions and infringement to small changes in the pictures. Also, the CNN accommodates little infringement to some properties like image interpretation, scaling, and rotation. The CNN here extricates progressively bigger feature sets in various levelled layers.

Hurieh Khalajzadeh *et al.* [3] A progressive structure based CNN was proposed to give the capacity for vigorous data handling. The weight sharing capacity of CNNs was taken into consideration as a level of progression to decreases quantity of free parameters and to enhance the speculation capacity. They utilized a little CNN to extract the features which passes over the entire info picture. Their structure provided the advantage of lesser parameters less training time with improved precision outcomes.

Rizoan Toufiq *et al.* [6] in the same year presented an approach for recognizing faces utilizing PCA with Back-propagation NN method. In their work, they combined the face and edge detection methods to extract face features which were later reduced using PCA and then fed to the neural network for classification. The classification of reduced features was done by utilizing a classifier based on backpropagation neural network.

Xueyi Ye *et al.* [9] they proposed a deep learning neural system employing the multi-layered nonlinear mapping and the semantic element extraction to detect the faces from images. They utilized the status likelihood of the neurons for displaying the status of the human cerebrum neuron. In addition, the quantity of the neurons with hidden layers diminished layer by Iayer to remove the excess data of the information and quicken the recognition process consolidating it with the skin shading recognition.

Chao Xiong *et al.* [11] a restrictive CNN architecture which was called c-CNN was presented in 2015 to deal with the issues related to multi-modality facial identification. Not quite the same as traditional convolutional neural networks which generally utilize the fixed kernels for convolution, their proposed architecture utilized dynamic kernels. Subsequently, their structure was not dependent on any earlier learning of biometrics different from the most prior techniques. They tried to solve the issues related to occlusion face identification and multi-view facial identification by utilizing the concept of decision tree.

Indra Budiman *et al.* [12] discussed about a few methods to conquer noise attacks in face identification assignment utilizing Gabor and NMF factorization. The commotions talked about in this examination comprise of Gaussian noise, speckle noise and salt-and-pepper noise. They utilized the KNN as a classifier for classification noisy images. They used the median filter to handle the effects of Salt and Pepper noise and used mean filter to handle the effects of Gaussian Noise and Speckle Noise both.

Yuanyuan Ding *et al.* [13] they presented a precisely planned noise resistant system utilizing deep learning approach to recognize facial images affected with commotion. The introduction of a multiple-input block in the last fully connected layer was done for extracting the features from the information picture. The outcomes such that receiver-operating characteristic curves on challenging noise databases demonstrated that the proposed system was noticeably better than some of the existing algorithms in this field.

Klemen Grm *et al.* [14] examined the impacts of various factors of noise such as salt and pepper noise and Gaussian noise etc. on the identification execution of four late CNN models. They explored the impact of factors identified with picture quality and model attributes, and checked their effects on the face identification execution of various deep CNN models. They tested the CNN models against the impacts of compression artefacts, contrast changes, blur, brightness variations and missing pixel values.

Umme Aiman *et al.* [15] they presented an adjusted neural system utilizing deep learning approach for the recognition of facial images from very small databases. The architecture was made out of convolutional neural networks, fully connected layers and the ReLU activation layers. The size of train database was enlarged with artificially created samples by the application of Poisson and Gaussian noises to all the images of the database. They exhibited that the expanded train dataset really enhances the speculation intensity of convolutional neural networks.

Hurieh Khalajzadeh *et al.* [16] later in 2014 introduced a combined framework, where CNN architecture and a LRC classifier were consolidated. A convolution neural network went through a training to distinguish and perceive the facial pictures, and then a Logistic regression classifier was utilized for classification of feature sets. This made the framework to deal with facial data containing illumination and pose varieties. Logistic regression classifier was utilized for classification of facial pictures.

Ze Lu *et al.* [17] they demonstrated that calibrating pre-prepared convolutional neural networks are unable to give attractive facial identification accuracy if teat and train databases have substantial contrasts. To overcome the issue, they decided to enhance the identification precision of convolutional neural networks by utilizing non-CNN characteristics. The

reciprocal data included in non-CNN features extraordinarily enhanced the recognition rate of convolutional neural networks and pre-prepared CNNs.

Musab Co Kun *et al.* [18] they presented an altered CNN model which included the application of a standardization task to two layers of the model to provide the acceleration to the network. This standardization task was named as batch normalization. Facial features were extracted using convolutional neural network architecture and for the classification of the face pictures, Softmax classifier was utilized.

Xi Yin et al. [20] they investigated a multi-tasking learning approach for identification of faces. Initially, they designed a multi-assignment CNN architecture in which the person ID was the primary purpose and the estimation of poses, light, and articulation were the second preferences. After that, they built up a weighting for the assignment of weights to each of the side task. They also designed a posture coordinated multi-tasking convolutional neural network by gathering distinctive poses to learn posture-based features.

2.3 DEEP LEARNING BASED GAIT-RECOGNITION

Jang Hee Yoo *et al.* [21] in 2008, it was the first time when they utilised the neural networks for the purpose of person identification. But here they used the gait biometrics only. In their work, they extracted a 2-D stick figure from the gait sequences, which was fed to the back-propagation neural network model for classification. In the end, they utilized the back-propagation neural network technique for classification.

Emdad Hossain *et al.* [22] they presented a novel multi-biometric technique utilizing deep learning approach to learn the gait features for individual recognition. They combined the principle component analysis method of feature reduction with some new classifier such as SVM, MLP and neural networks and it provided a big improvement in the precision of the proposed multi-biometric recognition system.

Wanli Ouyang *et al.* [23] they assembled a multi-modal deep learning model keeping in mind the end goal to extricate non-linear features from various parts of data sources. With the deep learning architecture, they extracted patterns from individual's body for posture

detection. The process for assessing body areas and the process of individual identification were picked up mutually utilizing a bound together deep architecture.

Kohei Shiraga *et al.* [25] they presented a technique for identification of people from their gait characteristics utilizing CNN architecture. The gait energy images were given to the proposed convolutional neural network architecture as the input data. They named their model as GEINet which was made out of two consecutive triplets of convolution, pooling, and standardization layers followed by 2 fully connected layers.

Binu M Nair *et al.* [26] they presented neural system based on a deep learning approach to examine stride of people from bad quality videos for constant discovery of strange movement/dangers in reconnaissance uses. The training of the proposed model was done by joining an opposite kinematic Groebner-dependent architecture for estimation of the body joint points from the stance. These edge directions of the upper and lower furthest points of the body were utilized for distinguishing danger designs. This system portrays the connection between the low-level picture/movement highlights and the kinematics related with the development of a person. The assessed joint point directions are then delegated as the dangers and non-dangers utilizing a K-NN search classifier.

Deepjoy Das *et al.* [27] they introduced a new technique for gait identification utilizing deep learning approach which was composed of auto-encoders (deep stacked) followed by a softmax regression layer as a classifier for the classification of individuals walking properties. The human stride outlines were regenerated by removing noise from them and after that the feature extraction process as carried out. The estimation of missing walk areas were evaluated utilizing straight interpolation method and after that the scaling and arrangement was done in the end before feeding it to the network.

Daksh Thapar *et al.* [29] they presented a 3-Dimensional deep CNN architecture for individual ID utilizing the human gait as the main source of input information under numerous perspectives. This technique was organised into two parts where it contained an arrangement for classification which at the start recognize the view angle. For each of the view angle, they designed a separate architecture which was programmed to recognize the individuals from a specific view direction.

Ngoc Dung *et al.* [33] they built up an anonymization technique for preventing the fake identification of individuals from their gait. The technique alters the characteristics of gait such that the individual can't be recognized while keeping up the instinctive nature of the walk. The modified the architecture by including another step, known as "noise walk". The CNN model makes this change by taking these two walks as information and yielding an anonymized walk. The technique was assessed utilizing the achievement rate and MOS. The achievement rate is the rate of fizzled stride ID; furthermore, the MOS is calculation of the expectation of the anonymized stride.

Chao Li *et al.* [34] they presented a new gait identification technique based on video sensors, which was named as 'deep gait', by utilizing deep CNN architecture and acquainted Joint Bayesian architecture. It was created by utilizing a pre-trained "profound" system VGG-D with no any calibration. The proposed model performed better than other models such as GEI, Gait Flow Image and Frequency-Domain Feature and so forth.

Omid Dehzangi *et al.* [36] a new method was presented for person identification utilizing the stride information of individuals and the time-frequency development of human step cycles keeping in mind the end goal to catch joint 2-D spectral and worldly examples of step cycles. At that point, they composed a DCNN architecture figuring out how to separate selective highlights from the 2-Dimensional extended stride cycles and mutually upgraded the I-Dimensional display. They gathered crude movement information from five inertial sensors put at the right knee, right hand wrist, and right lower leg; bring down back and chest of every individual synchronously with a specific end goal to examine the effect of sensor area on the gait distinguishing efficiency. At that point they exhibited two strategies for right on time and late multi-sensor combination to enhance the walk ID speculation execution.

Wu Liu *et al.* [39] they designed a proficient spatial-temporal gait identification utilizing a deep learning approach for human recognizable proof in which initially they composed a Siamese neural system dependent upon GEIs as input information for extracting the gait features from these gait energy images. Moreover, they employed the 3-D CNN architectures for learning the human walking properties i.e. C-3-D as the temporal step highlights. At long last, the GEI and C3D walking highlights were installed into a null space utilizing the NFST algorithm. Here in this space, these spatial-temporal highlights were adequately joined with

separate metrics, figuring out how to drive the comparability metrics to be little for sets of stride from a similar individual and huge for sets from various people.

Zhuowen Lv *et al.* [40] this paper presented the idea of GEIs that is the gait energy images which initially originated in 2004. They suggested that a series of walking picture outlines could be changed over in a solitary grey picture containing the characteristics of all the frames of gait images of the sequence. They demonstrated that a GEI totally keeps up all the features of the whole sequence of gait such as dynamic characteristics and shape etc. They discussed the challenges and issues related to gait energy images which could be faced in the research works.

2.4 MULTI-BIOMETRIC FUSION METHODS

Ping S. Huang *et al.* [41] here in this article, they tended to solve the issue of perceiving people progressively, information combination and then to track them by neuro-fuzzy estimation technique. This method consolidated eigen-space change with standard space change. This technique could be utilized to diminish information dimensions and to streamline the subject distinctness of various subjects at the same time.

Xiaoli Zhou *et al.* [43] in the year 2006, proposed a way to deal with intertwine side face and gait on include level where they utilized PCA and MDA techniques to get the side face and gait image characteristics separately. The Enhanced Side Face Image approach was utilized for extracting the face information from the images and Gait Energy Image approach was utilized for extracting the gait information from the images.

Xiaoli Zhou *et al.* [44] then again later in 2007, they exhibited a video based acknowledgment technique to perceive non-collaborating people at a separation from the video database, who open side perspectives to the camera. Data from 2 modalities i.e. side face and walk, was used for feature fusion. Here also, the Enhanced Side Face Image approach was utilized for extracting the face information from the images and Gait Energy Image approach was utilized for extracting the gait information from the images.

Zongyi Liu *et al.* [46] they investigated the likelihood of utilizing both walking and face information to improve the individual identification efficiency at a separation for outdoor

conditions and demonstrated that efficiency was essentially upgraded by blend of these two traits. For extracting the gait data, they utilized the HMM model i.e. hidden Markov model and to extract the facial information, a matching technique dependent upon elastic bunch graph was utilized. Then the averaging of these detected gait images was done utilizing eigen-gait approach on which the closeness between two stride groupings was dependent. The presented methodology altogether beat the HumanID Gait Challenge approach.

Xin Geng *et al.* [47] later in 2008, they proposed the same thing as mentioned in the above literature but this time they presented an efficient algorithm to dynamically adjust the fusion rules for the real time application purpose. They discussed about the adaptive combination of gait and facial information. Here, they discussed about the two parameters that could create a huge impact on adaptive combination of gait and facial information i.e. distance of individual from camera and the view point.

Emdad Hossain *et al.* [49] they presented a multi-biometric combination technique utilizing the PCA and LDA methods for individual ID out of low quality reconnaissance database of videos where information was extricated from stride and facial modalities. They demonstrated that the consolidated information via PCA and LDA when combined either by utilizing holistic approach or by hierarchical approach, could provide a great human ID efficiency could catch the multi-biometric characteristics even from bad quality video data.

Martin Hofmann *et al.* [51] they proposed a strategy which was depended on consolidating progressive gait identification approach with an adjusted low resolution facial identification approach. For that purpose, they tried a bit different algorithm known as alpha-Matte processing which utilized a mechanized segmentation strategy. It permitted the better development of highlights from the input pictures utilized for gait identification. A similar approach was additionally utilized for identifying the low-quality facial pictures from similar datasets.

Aparna Behara *et al.* [52] they presented 2 unique strategies for building up a solid individual identification architecture which was produced with the assistance of Emgu CV and Open CV. Utilization of two distinctive procedures i.e., Holistic depended approach and Model dependent approach was done for extraction of two distinct feature types. The model dependent approach was utilized for extricating some specific static and dynamic

characteristics from the gait and facial modalities. The Holistic depended approach was utilized for extricating some factual characteristics identified with the entire picture of the individual. At last these acquired characteristics were melded utilizing three approaches i.e. SVM, ANN, and Bayes approach.

B. A. Lathika *et al.* [54] to outperform the constraints of single biometric frameworks, they proposed to design a multi-biometric architecture comprising of a mixture of gait, ear and facial biometric information. Here, they utilized a wavelet transform to extract the features which depicted the proportion amongst black and white zones. For identification purpose, they utilized the ANN architecture to accomplish great acknowledgment rate with the availability of huge varieties in face images. The z-score strategy was employed for normalization of images for better combination outcomes.

Xianglei Xing *et al.* [56] later in the year 2015, they presented a coupled projection approach to combine the face and gait modalities in which the feature sets from both traits are projected into a common subspace to reduce the distance between them for fusion and hence the gait and facial features could be combined easily in this unified sub-space. They utilized the concept of GEIs for generating gait characteristics. KNN search classifier was utilized in the end for classification of results.

A. Derbel *et al.* [57] they presented an approach to manage the issues like little sized or bad quality pictures with differing postures, enlightenment, articulations, presence of glasses or caps, availability of beard and moustache, variations in light and illumination etc. by combining face acknowledgment and gait examination.

Ting Yu Fan *et al.* [58] later in the year 2017, they proposed architecture based on the deep neural networks for multi-biometric fusion of face and the ear modalities for person identification from a distance. First of all, face and ear features are extracted from the images using fast R-CNN and then training is done using CNN architecture and, in the end, Bayesian decision fusion technique is used for fusion of results.

Jianzheng Liu *et al.* [59] in 2016, they proposed a seven-layer deep neural network for person identification in which face image and motion history image were fused to feed to the network. This is the work which is a bit related to what we have proposed but instead of

MHIs we have fused GEIs with the face images. In their work, the first 6 layers of network act as feature dimension reducer or autoencoder and last layer is SoftMax classification layer. Instead of 7 layers we have proposed a 13-layers deep neural network.

Onur Can Kurban *et al.* [60] presented a deep neural network based approach in the year 2017 for person identification which utilized the score level fusion technique for the fusion of the facial images and gesture energy images. The VGG model was utilized for extracting the facial features from the face images and gesture energy image features were extracted by utilizing the energy imaging technique.

2.5 GAPS IN STUDY

Face and gait biometric fusion based identification techniques provide some promising outcomes, even though efforts are required to achieve higher identification results in the security areas. The single biometric architectures utilize the characteristics from a single modality and come against some issues like big error rates, variations in the image database, noised images, spoofing, and non-universality etc. The interdependent data combination plays an important role to overcome the limitations of single biometric recognition systems. The foremost issue in the multi-biometric fusion architectures is to select the exact approach for integration or fusion of data coming from different modalities. The reasons behind choosing the multi-biometric approach instead of single biometric systems are such as execution time, greater resistance, high demand of computational requirements, memory and last but the most important i.e. improved accuracy. This dissertation work includes the fusion of face and gait biometrics.

The reason for using these two traits specifically for fusion is that the characteristics which affect the process of face recognition such as occlusion and blur etc. doesn't affect the process of gait recognition. In the same way, face recognition process is unaltered by factors altering gait recognition such as clothing, subject to camera angle and carrying accessories etc. Hence both the traits are robust to each other which help to improve the accuracy. The task of identifying people automatically has always been a challenging one especially when we have to come across the issues such as large datasets and the robustness against the factors affecting recognition such as pose variation, subject to camera angle, illumination, poor quality data and occlusion etc. All these challenges motivated us to utilize the deep learning approach for the fusion of face and gait biometrics. Hence we decided to design an

architecture to identify people by fusing their gait and face biometric traits using a 13-layer deep convolution neural network.

2.6 OBJECTIVES

The main objectives of this research work are as follow:

1) To study the existing techniques in the field of face recognition, gait recognition and multi-modal fusion utilizing deep learning.
2) Preparation and fusion of the master face database images with the corresponding master gait database images.
3) Training and testing of the proposed 13-layer DCNN algorithm with the prepared databases for person identification.
4) Testing the proposed model against some common noise attacks such as salt and pepper noise, Gaussian noise and speckle noise.
5) Comparison of the outcomes of the proposed 13-layer DCNN model with the existing techniques in this field.

2.7 SUMMARY

The basic fundamentals and origin of the face and gait biometrics with the various techniques is discussed in this chapter. First of all, literature review is done for face recognition techniques utilizing deep learning. Then, we have discussed the existing gait recognition techniques which utilize the deep learning approach. Then, some existing techniques for multi-biometric fusion for person identification are discussed. On the basis of the literature review, we found some gaps in study which become the basis for our research objectives.

CHAPTER 3
OVERVIEW OF DEEP CONVOLUTIONAL NEURAL NETWORKS

3.1 INTRODUCTION

In the following segment, a concise depiction of the proposed algorithm has been provided, where two discriminative systems are joined together: a convolutional neural network and a Softmax regression layer. A convolutional neural network may contain a huge number of hidden layers and every layer has a capability to extract some features in a picture. But every hidden layer adds some complexity to the extracted features in a picture. Hence, we have utilised only nine hidden layers in our convolutional neural network. First layer is an input layer after that the hidden layers are the three triplets of convolution layer, ReLU layer and pooling layer. After that, there is a fully connected layer followed by a SoftMax regression layer. In the end, a classification layer is used to classify the recognition results. Let us have a brief knowledge about the history of neural network before moving forward in the chapter. In 1943, two early computer scientists named Mc Culloch and Pitts invented the 1st computational model of a neuron which was a simple model but in early days of artificial intelligence that was a big deal [31]. A few years later a physiologist named Rosenblatt build a model called Perceptron which is another word for a single layer feed forward neural network.

At that point later, G. E. Hinton et al. exhibited an auto encoder system, which could diminish the dimensionality of any information by viably initializing the weights [10]. This gave more productive outcomes than the PCA approach. In our work, instead of using an auto-encoder in particular for feature extraction and dimensionality reduction, we have designed our own deep convolution neural network (DCNN) which has a total of 13 layers including 9 hidden layers. Working of a CNN depends on three key point's i.e. local receptive fields; shared weights and biases; and activation and pooling. The local receptive field is translated over an image to generate the feature maps from input layer to hidden layer neurons and for this purpose convolution layer is utilised. In case of CNNs the weights and bias values are same for all hidden layers that mean the network detects the same features in different regions of image. 3rd most important concept for our network is activation and pooling. The activation step applies the transformation to the output of each neuron by using an activation function. We have used ReLU (rectified linear unit) as the activation function in

our network. Hence after convolution layer, reluLayer is used in the network. It maps the output of the neuron to highest positive value and if the output is negative then it maps it to zero. The output of the activation step is further transformed by applying a pooling step which reduces the dimensionality of feature maps by condensing the output of small regions of neurons into a single output. We have used max-pooling layer in our model for this purpose. A representation of the all the layers utilized in our proposed deep convolutional neural network architecture is presented in the Figure 3.1 below. A detailed introduction about all the layers utilized in our DCNN network is explained below in the remaining part of this chapter.

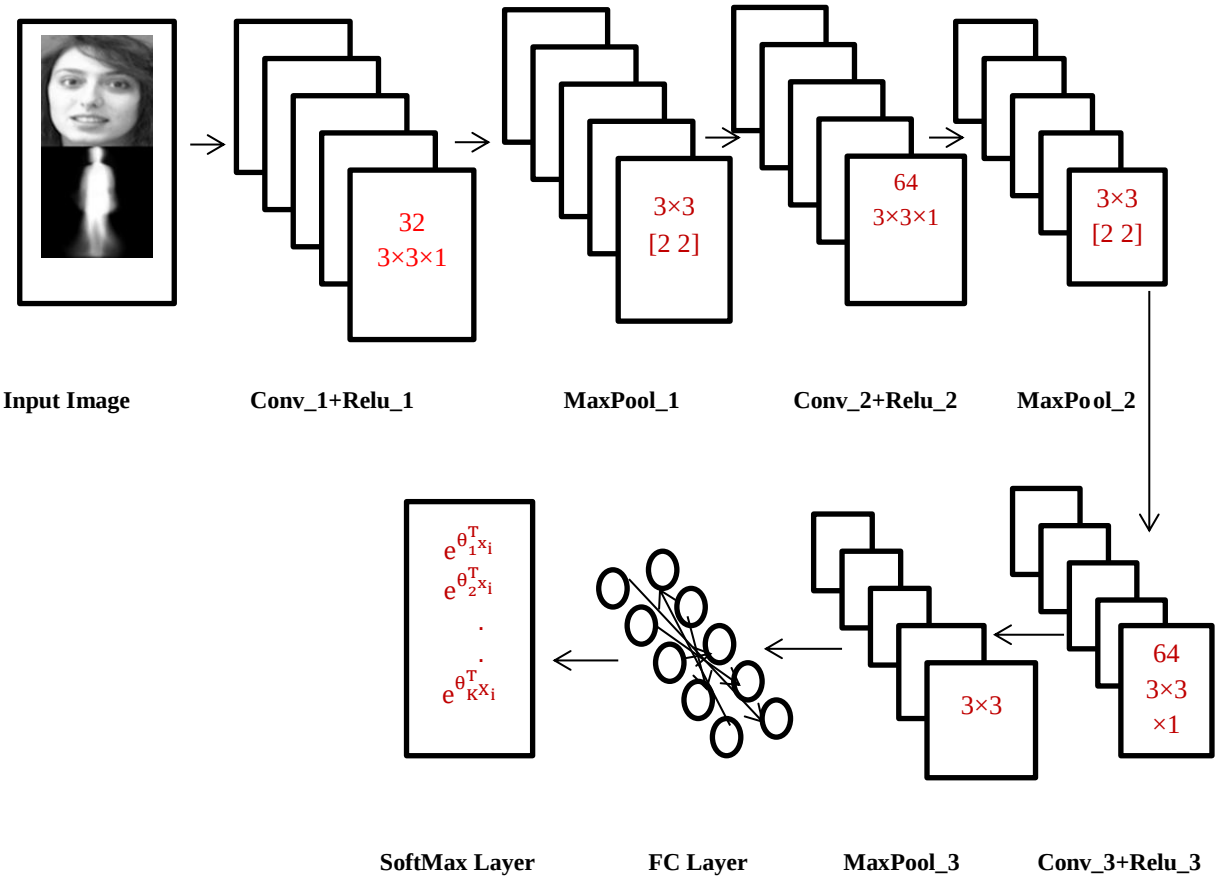


Fig. 3.1 Layers of Proposed DCNN model

3.2 CONVOLUTIONAL LAYERS

Here in the convolutional layer, the weights comprise of an arrangement of some kind of kernels which are arbitrarily created and have some learning ability. They utilize the back-propagation calculation method for this purpose. They are comprised of a couple of

neighbourhood associations which are completely interfaced with the past layer. Consequence of every bit convolved over full picture is known as the feature maps. The quantity of the feature maps is equivalent to the quantity of connected kernels in the corresponding layer. An example of the convolutional operation is shown in figure 3.2 which uses a filter of 3×3 region and outputs a feature map of 3×3 dimension. Feature maps for different filters combine to form a convolution layer. In our network, we have utilized three convolutional layers as shown in figure 3.1. The first convolutional layer produces a total of 32 3×3×1 convolutions having a stride of [1 1] and padding [0 0 0 0]. Second convolutional layer produces a total of 64 3×3×32 convolutions having a stride of [1 1] and padding [0 0 0 0]. Similarly, third convolutional layer produces a total of 64 3×3×32 convolutions having a stride of [1 1] and padding [0 0 0 0] same as second convolutional layer.

I =

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

(a) Input Image

k =

1	0	1
0	1	0
1	0	1

(b) Convolutional kernel

m =

4	3	4
2	4	3
2	3	4

(c) Feature map

Fig. 3.2 Convolution Layer Operation

The figure 3.2 above shows an example of how a convolutional operation is carried out by the convolutional layers to contribute to the feature maps. Here 'I' in figure 3.2 (a) is the 5×5 input image and 'k' in figure 3.2 (b) represents the 3×3-convolutional kernel which convolves over the input image to output a feature map shown in figure 3.2 (c). The equation 3.1 below represents the activation map of the convolution layer where $x^{i(p)}$ and $y^{j(p)}$ denote the ith input and the jth output activation maps, separately [16]. Also $a^{j(p)}$ here denotes the bias of the jth output map. The variable $N^{ij(p)}$ denotes the convolutional kernel residing in the jth output map and the ith input map, and * represents convolution.

$$y^{j(p)} = \max\left(0, a^{j(p)} + \sum_{i} N^{ij(p)} * x^{i(p)}\right) \tag{3.1}$$

3.3 RELU (RECTIFIED LINEAR UNIT)

Another most imperative idea for our system is the activation function. The activation step applies the change to the yield of every neuron by utilizing an activation function to them. ReLU also known as the rectified linear unit has been utilized as the activation function in our proposed 13- layer deep convolutional neural system. Henceforth after convolution layer, reluLayer is utilized as a part of the system. It maps the yield of the neuron to most elevated positive esteem and if the yield is negative then it maps it to zero. It is an elemental based task means it is employed per pixel and furthermore it presents non-linearity in the convolutional neural system [25]. Figure 3.3 shows the plot representing the operation of ReLU activation function. In our proposed network, we have utilized three Relu layers each after the convolution layer as shown in the proposed architecture block diagram in figure 3.1.

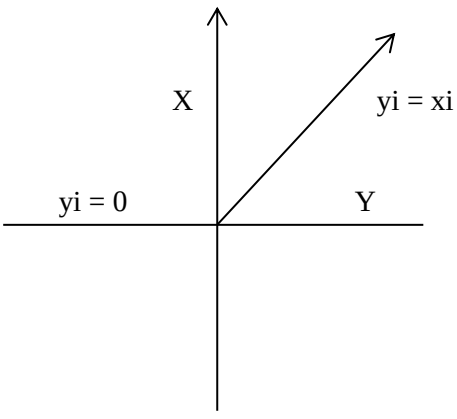


Fig. 3.3 ReLU Activation Operation

3.4 MAX-POOLING LAYERS

The fundamental capacity of this layer is to diminish the size of the convolution layers' yield, and also it delivers a restricted type of the translation in-variance. Once a particular component has been recognized by the convolution layer, just its rough area in respect to different highlights is kept. Every part of the convolution layer's yield is isolated into non-covering areas, and for every sub-region, the greatest term is considered in output. Mainly, the greatest value out of every cluster of neurons is utilized i.e. only the most important features of input image are chosen [36]. A regularly utilized shape in a max-pooling layer is with areas of size (2×2) having a stride of [2 2]. Depth dimensions of convolutional layer's output are kept unaltered. In our proposed deep neural network, we have utilized three Max-pooling layers each after a Relu activation layer as shown in the block diagram of proposed DCNN in figure 3.1. For all the three Max-pooling layers, we have chosen the following configuration: 3×3 max pooling with stride [2 2] and padding [0 0 0 0]. Equation 2 below represents the formula for the max pooling layer where $y_{j,k}^{i}$ denotes the neuron present at the ith output activation map.

$$y_{j,k}^{i} = \max\left(x_{js+p,kr+q}^{i}\right) \quad \text{for } 0 \le p, q < r \tag{3.2}$$

The figure 3.4 below presents an example showing the operation of the Max-pooling layer which utilizes a filter of pool-size 2×2 region with a stride of 2. Here figure 3.4 (a) represents an input feature map of size 8x8 which is needed to be max-pooled and figure 3.4 (b) represents the output of the max-pooling layer termed as the pooled feature map of size 4x4.

8	7	4	3	4	0	9	6
4	5	6	9	6	1	1	4
9	6	4	5	5	3	1	2
1	4	6	3	1	2	0	5
1	2	4	0	1	4	2	3
0	5	6	1	3	2	7	8
2	3	5	3	1	2	4	0
7	8	1	2	0	5	6	1

(a) Input Feature Map

8	9	6	9
9	6	5	5
5	6	4	8
8	5	5	6

(b) Pooled Feature Map

Fig. 3.4 Max-Pooling Layer Operation

Hence max-pooling layer reduced the size of the feature map from 8×8 to 4×4 keeping only the most important pixels in the feature set and eliminating the rest of all other pixels. Hence this is a better way of reducing the dimensionality of features. Here the filter of pool-size 2×2 pass over the whole image with a stride of 2 and selects the maximum value from each of the 2×2 region to contribute to the output pooled feature map.

3.5 FULLY CONNECTED LAYERS

At last, after a few max pooling and convolution layers, the higher potential work in the neural system is done by means of fully connected layers. In a fully connected layer, neurons have associations with all actuations in the past layer, as found in general neural systems. Their initiations can thus be figured with a multiplication of matrix took after the bias offset. The objective of the total fully connected structure is to tune the weight characteristics to make a stochastic probability portrayal of each class in light of the actuation maps created by the link of non-linearity, pooling and the convolution layers. Individually, the fully connected layers work indistinguishably to multi-layer perceptron layers but here, the input image layer is the only exception. The size of fully connected layer is same as the total number of subjects or classes in the datasets.

$$y^{(l)}(j) = f^{(l)}\left(\sum_{i=1}^{N^{(l-1)}} y^{(l-1)}(i).w^{(l)}(i,j) + b^{(l)}(j)(3)\right) \quad (3.3)$$

The equation 3.3 above represents the output $y^{(l)}(j)$ of the neuron j in a fully connected layer l [32]. Here, $\mathrm{N}^{(\mathrm{l}-1)}$ denotes the total number of neurons in previous layer, $w^{(l)}(i,j)$ denotes

weight for the connection from neuron j in the previous layer to the neuron j in the present layer. Also $b^{(l)}(j)$ represents the bias of neuron j in present layer.

3.6 SOFTMAX REGRESSION CLASSIFIER

Most of the times, the softmax actuation step is utilized by the fully associated layers for classifying the input images. The softmax layer for the most part is known as the Softmax regression classifier. It is a summed up type of paired strategic regression classifier expected to deal with multi-class classification assignments. The output from fully associated layer is taken by the Softmax layer and then it figures the likelihood framework for every one of the subject's class and after that the yield is fed to the classification layer to coordinate every one of the query picture to their classes as per this likelihood grid.

$$S_\theta(x_i) = \begin{bmatrix} a(y_i = 1|x_i;\theta) \\ a(y_i = 2|x_i;\theta) \\ . \\ . \\ a(y_i = K|x_i;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x_i}} \begin{bmatrix} e^{\theta_1^T x_i} \\ e^{\theta_2^T x_i} \\ . \\ . \\ e^{\theta_K^T x_i} \end{bmatrix} \tag{3.4}$$

Assume a total number of X classes and n number of labelled train images are in the database then, the Softmax classifier delivers a K-dimension vector, components of which aggregate to 1, where every component in the yield vector represents the evaluated probability matrix of each labelled image in the classes. The K-dimensional vector produced is represented in Equation 3.4 above where $\theta 1$, $\theta 2\ldots$ and θK represent the parameters which are randomly produced by utilizing back-propagation algorithm [35]. Here, $S_\theta(x_i)$ represents the hypothesis for estimating probability vector for each label.

$$C(\theta) = -\frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{K} 1\{y_i = j\} log \frac{e_{jX}^{\tau_i}}{\sum_{l=1}^{K} e_l^{T x_i}}\right] + \frac{\lambda}{2}\sum_{i=1}^{K}\sum_{j=0}^{m}\theta_{ij}^2 \tag{3.5}$$

The SoftMax classifier also uses a cost function for estimating the probability which is also called as the loss function. Equation 3.5 above represents the formulation for cost function in which the 2nd term of the equation represents the weight decay which helps to reduce the magnitude of weights to get away from overfitting.

3.7 SUMMARY

In this chapter, we have presented an overview about the deep convolutional neural networks. First of all a brief introduction and history of the convolutional neural networks is discussed. Then each layer utilized in our proposed algorithm i.e. convolutional layer, ReLU layer, max-pooling layer, fully connected layer and softmax regression layer are discussed in detail one by one with their mathematical equations. A block diagram representing the layers utilized in our proposed DCNN model is also shown in this chapter.

CHAPTER 4
PROPOSED WORK

4.1 PROPOSED ALGORITHM

Below in figure 4.1, the proposed multi-biometric deep learning model for person identification is presented. As contrasted to other systems, rather than utilizing a solitary trait that is either only face biometric or only gait alone, we have melded the pictures of faces with that of GEIs to feed to the presented 13-layer neural system. To start with, we change over the gait frames for each subject into a solitary GEI by taking the average of the entire walk cycle. This Gait Energy Image generation is done with the help of the condition specified in the section 3.3 i.e. equation (4.1). At this point, for the face information, detection of face is carried out initially from the face sequences to accomplish the ROI. And thus, faces are aligned by using a face alignment algorithm. Afterward, the aligned face pictures are changed over into grey level pictures so that they can be melded with the grey level Gait Energy Images. From this point, both the face and the gait pictures are resized into 100×100 size images and after that these image features are changed over into vectors of same size. Now, normalization of both the image vectors is done to minimize the size of feature vectors. Presently after this step, fusion of both image vectors i.e. face and gait is done. Now, the fusion output is fed as input to the presented deep learning model for classification.

As shown in figure 4.1 , our proposed deep convolution neural network is made up of 13 layers out of which first one is the input layer, then there are nine hidden layers, after that the network contains a fully connected layer which is later connected to a softmax regression classifier or a softmax layer in short. In the end the last layer is the classification layer for classification of the recognised images. We will explain configuration and parameters of the entire layers of deep convolution neural network one by one as follows:

4.1.1 Image Input Layer

First layer of the network is Image Input Layer. This is where we specify the size of the input image which is in case 200×100×1 for our database. It is related to the height, width, and the size of the channel i.e. whether the images are grey scale images or coloured images. As our images consist of the grey scale images hence the channel size is specified as 1. But for the

coloured images, the size would be 200×100×3 i.e. the channel size would be taken as 3. Transformation operations such as flipping or cropping of the images could also be specified at the image input layer in case if we want to avoid any chance of overfitting. But we don't need to do this because we have already done it before during the time of training by randomly training the image data and also the in the pre-processing phase. The equation 4.1 represents the image input layer matlab command used in our network.

$$\mathrm{I} = \mathrm{imageInputLayer}([200\ 100\ 1]) \tag{4.1}$$

4.1.2 Convolution Layer 1

Just after the image input layer, here comes the first convolution layer of the network. This is where the hidden layers start. This is the first hidden layer out of nine hidden layers in the network. This layer contains a total of two arguments out of which the first one is the filter size which specify the height and width of the convolution filter. The 2nd one indicates the total number of convolutional filters contributing in the feature maps. It represents the total number of neurons connecting to the output. The first convolutional layer produces a total of 32 3×3×1 convolutions having stride of [1 1] and padding as [0 0 0 0]. We can clearly see from equation 4.3, which represent the first convolutional layer of the network, that is 3 the first augment represent the size of filter as 3×3. And the 2nd augment i.e. 32 represent the number of convolutional filters.

$$\mathrm{Conv1} = \mathrm{convolution2dLayer}(3,32) \tag{4.2}$$

4.1.3 ReLU Layers:

Just after the first convolutional layer, we have to use a non-linear activation function to present the non-linearity in the convolutional neural system and we have used rectified linear unit function that is ReLU as an activation function here. We have already discussed in the third section that ReLU activation function maps the yield of the neuron to most elevated positive esteem and if the yield is negative then it maps it to zero. It is an elemental based task means it is employed per pixel and furthermore it presents non-linearity in the convolutional neural system. In our network we have used a total of three ReLU layers, each of which comes after each convolution layer in the network. The matlab command for ReLU

layer is pretty simple. We just need to write the command ‘reluLayer’ as shown in equation 4.3 which represents the first ReLU layer of our proposed DCNN. The detailed explanation of the ReLU activation function is explained in the third section.

$$\text{relu\_1} = \text{reluLayer} \tag{4.3}$$

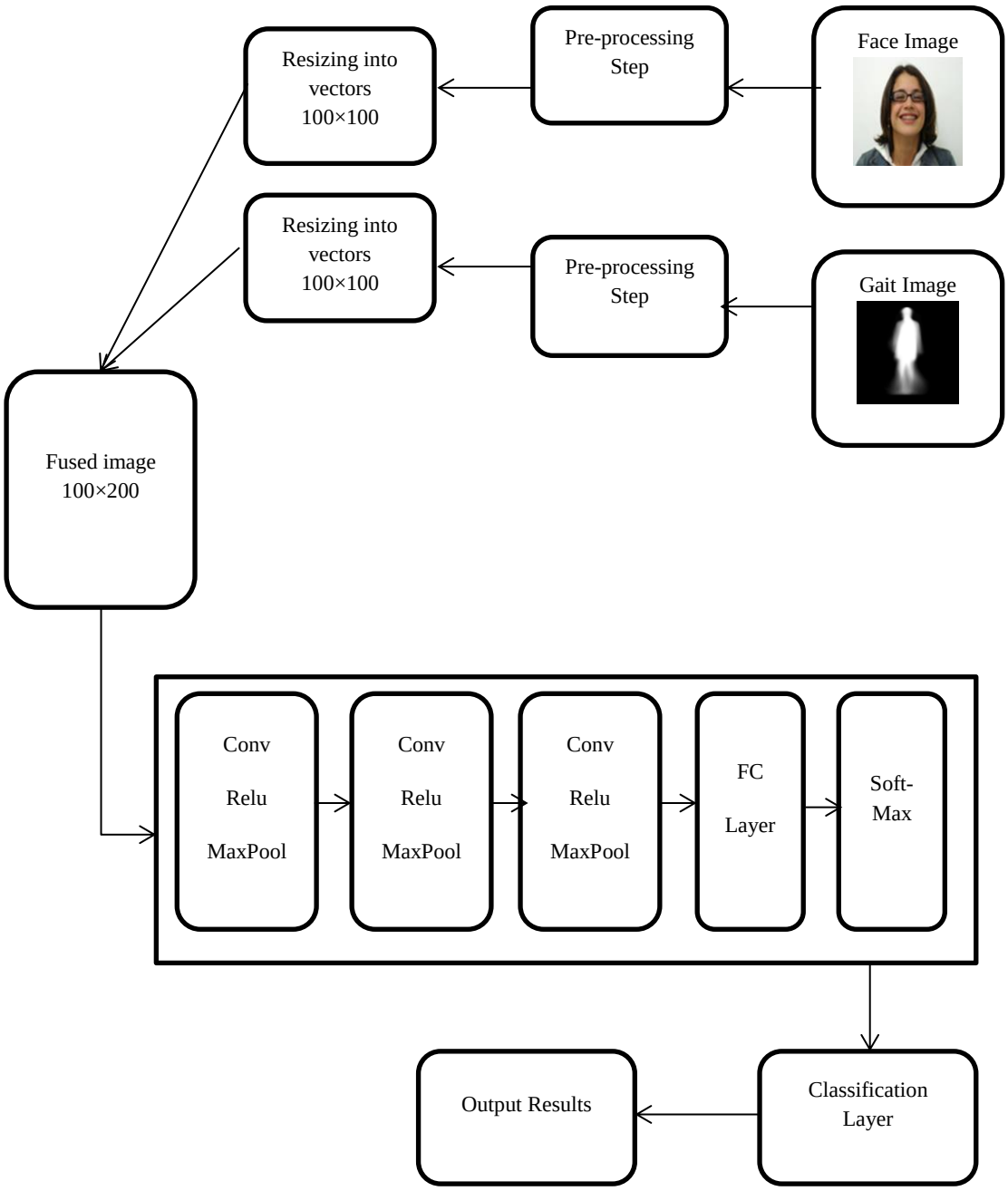


Fig. 4.1 Overview of the Proposed 13-layer Multi-Biometric Deep Convolutional Neural Network Model

4.1.4 Max-Pooling Layers

After the first convolutional layer and the ReLU activation layer we have to perform some kind of down-sampling operation on the feature maps. Basically this is done to reduce the number of parameters of the feature maps. Also this is another way of preventing overfitting. And the best method of doing this is to use a max-pooling layer after the ReLU layer. The functioning of max-pooling layer has already been discussed in section three with examples. The equation 4.4 below represents the matlab command for the first max-pooling layer used in our proposed DCNN network. There are also two arguments in the max-pooling layer too out of which the first argument represents the poolSize and the second argument is the stride which is an optional argument. Stride represents the step size taken by the function during training of the image that is during scanning the image. We have used a total of three max-pooling layers in our proposed deep convolutional neural network, each of which is used after each ReLU activation layer in the network. For all the three Max-pooling layers, we have chosen the following configuration: 3×3 max pooling with stride [2 2] and padding [0 0 0 0] as shown in the equation below.

$$\text{maxpool} = \text{maxPooling2dLayer}(3,'\text{Stride}',2) \tag{4.4}$$

4.1.5 Convolution Layer 2

After the first ReLU activation layer, the DCNN contains the second convolutional layer which is the fourth hidden layer of the network as well. Similar to the first convolutional layer, this layer also includes the first argument as 3 i.e. it also uses the filter size as 3×3×1. But the second argument is taken as 64 instead of 32 (as specified in first convolutional layer) i.e. it includes the total number of 64 filters to contribute to the feature maps. Hence, the second convolutional layer produces a total number of 64 3×3×1 convolutions with stride [1 1] and padding [0 0 0 0]. The equation 4.5 below represents the matlab command for second convolutional layer used in our proposed deep convolutional neural network. After the second convolutional layer, the network includes the second ReLU layer and the second max-pooling layer with the same configuration as used in the first max-pooling layer described in the above paragraph that is 3×3 max pooling with stride [2 2] and padding [0 0 0 0]. These three layers (second convolutional layer, second ReLU layer and the second max-pooling

layer) mentioned in this paragraph forms the fourth, fifth and sixed hidden layers of the deep convolutional neural network respectively.

$$\text{Conv2} = \text{convolution2dLayer}(3,64) \tag{4.5}$$

4.1.6 Convolution Layer 3

The DCNN includes the third convolutional layer after the second ReLU activation layer in the architecture which forms the seventh hidden layer of the DCNN model. This layer uses the same configuration of arguments as used in the second convolutional layer. That means this layer also uses the filter size as 3×3×1 and includes the total number of 64 filters of this size to contribute to the feature maps. Hence, the third convolutional layer produces a total number of 64 3×3×1 convolutions with stride [1 1] and padding [0 0 0 0]. The equation 4.6 below represents the matlab command for third convolutional layer used in our proposed deep convolutional neural network. After the third convolutional layer, the network includes the third ReLU layer and the third max-pooling layer with the same configuration as used in the first and the second max-pooling layer described in the above paragraphs that is 3×3 max pooling with stride [2 2] and padding [0 0 0 0]. The three layers (third convolutional layer, third ReLU layer and the third max-pooling layer) mentioned in this paragraph forms the seventh, eighth and ninth hidden layers of the deep convolutional neural network respectively.

$$\text{Conv3} = \text{convolution2dLayer}(3,64) \tag{4.6}$$

4.1.7 Fully Connected Layer

Every deep neural network includes one or more fully connected layers after the hidden layers i.e. convolutional, ReLU and pooling layers. Our deep convolutional neural network also includes a fully connected layer after the ninth hidden layer (third max-pooling layer) in the network. This layer forms the eleventh layer of our deep convolutional neural network. In a fully connected layer, neurons have associations with all actuations in the past layer, as found in general neural systems. This is the reason why it has the name as fully connected layer. The size of the output in the fully connected layer is same as the total number of subjects in the database. Hence in our experiments, the fully connected layer has the output

size equal to 40 for the first experiment and it is 124 for the second experiment. This is because for the first experiment, we have total number of 40 classes and for the second experiment; we have a total of 124 classes. The equation 4.7 below represents the fully connected layer in matlab format used in our deep convolutional neural network model. The detailed discussion about the working of fully connected layer is provided in the third chapter with its working and mathematical equations.

$$\mathrm{fc} = \mathrm{fullyConnectedLayer}(\mathrm{number\ of\ classes}) \tag{4.7}$$

4.1.8 Softmax Layer

Generally the fully connected layer utilizes the softmax activation function for the purpose of classification of images which is placed after the last fully connected layer. But our network has only one fully connected layer, hence it comes after that fully connected layer and forms the 12th layer of our proposed deep convolutional neural network. The softmax layer for the most part is known as the Softmax regression classifier. The output from fully associated layer is taken by the Softmax layer and then it figures the likelihood framework for every one of the subject's class and after that the yield is fed to the classification layer to coordinate every one of the query picture to their classes as per this likelihood grid. The detailed explanation about the fully connected layer and its working with its mathematical part are discussed in the third chapter.

4.1.9 Classification Layer

The thirteenth and the final layer of our proposed deep convolutional neural network is the classification layer which is generally placed after the Softmax Layer. The classification layer in general holds the name of the product utilized to train the system for classifying multiple classes, loss function, class labels and output size. The classification layer utilises the output of probabilities provided by the Softmax regression classifier as the input and then classifies the images according to this set of probabilities. There is a simple command to represent the classification layer which is shown in the equation 4.8 below.

$$\mathrm{c} = \mathrm{classificationLayer}() \tag{4.8}$$

4.2 CONCEPT OF GEIS

In the year 2004, the idea of GEIs that is gait energy images, was initially presented by Zhuowen et al. [40]. They suggested that a series of walking picture outlines could be changed over in a solitary grey picture containing the characteristics of all the frames of gait images of the sequence. This is done by averaging the entire cycle or succession by utilizing the accompanying condition:

$$\mathrm{GEI} = \frac{1}{\mathrm{K}} \sum \mathrm{B}\,(\mathrm{X}, \mathrm{Y}, \mathrm{K}) \qquad \text{for } \mathrm{K} = 1 \text{ to } \mathrm{k} \tag{4.9}$$

The equation 4.9 above represents the mathematical formulation for formation of a gait energy image from a sequence of gait images where, k in the equation demonstrates the aggregate of walking pictures in the given series of gait frames. The fundamental reason for generating this single grey level gait energy image is to lessen the processing time and to minimize the framework cost because a gait energy image convey the data of all the gait frames in a solitary picture which likewise decreases the measure of database. A GEI totally keeps up all the features of the whole sequence of gait such as dynamic characteristics and shape etc. Thus, the idea of GEIs is being used in this work presented. The gait energy image could be effortlessly influenced by a few factors, for example, wearing and carrying conditions, and so forth consequently we have utilized the CASIA B dataset for gait images to prove the worthiness of proposed DCNN. Since CASIA B dataset incorporate both the previously spoken varieties i.e. cloths and carrying bags.

(a) Gait Sequence

(b) Corresponding GEI

Fig. 4.2 Conversion of gait sequence into GEI

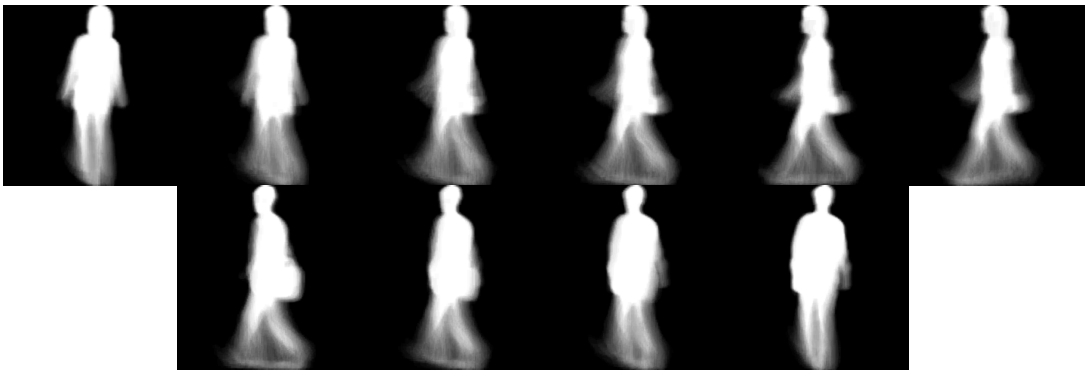
Figure 4.2 above demonstrates the sample pictures of the gait cycle and its converted GEI below it where Fig. 4.2(a) displays the gait images in the entire cycle and the Fig. 4.2(b) displays corresponding gait energy image of the gait cycle.

4.3 DATABASE DESCRIPTION

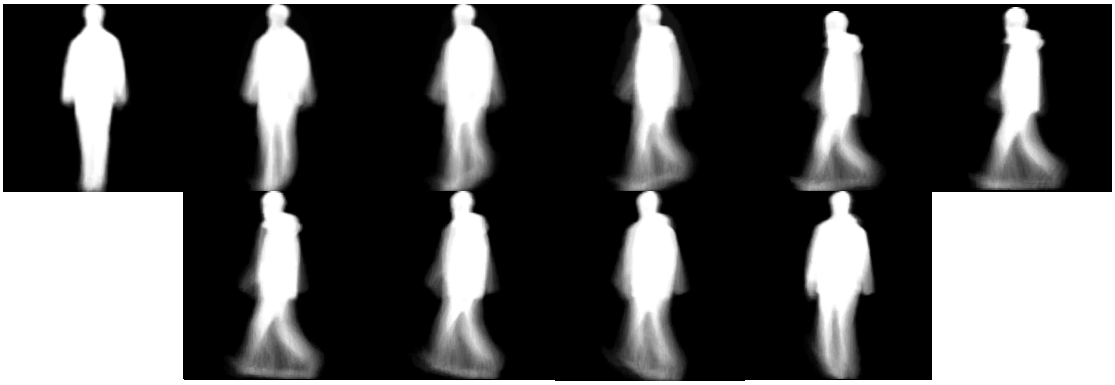
In our proposed work a total of three datasets are used to confirm the robustness of our proposed DCNN framework. The freely accessible CASIA Gait B Dataset has been utilized to represent the gait information. To represent the face biometric trait, two databases have been used to fuse with the CASIA Gait B Dataset. The first one is the ORL Face Database and the second one is FEI Face Database. The CASIA Gait B Database includes an aggregate of 124 subjects where each subject contains the walking sequences taken from ten different angles. Every class has two more varieties in the database in addition to the angle variation; these are the clothing and carrying material variation. Each of the class has first 2 sequences of individuals holding bags with them. The next 2 sequences are taken with individuals wearing coat. The other 6 gait groupings left are taken when individuals are walking normally. For every individual, each walking sequence is taken from 10 distinct angles ranging between 0 to 180 degrees. The second dataset is ORL Face Database which includes a total of 40 subjects in it, with each subject containing ten face images of same individual. The ORL dataset has been used in light of the fact that it is the most fundamental dataset for face recognition purpose and images are taken with different face expression and lighting variation. The third database used here is FEI Face Database which contains face images of 200 subjects with each subject having 14 face images of the person. Every face picture is changed over into 100×100 pixel size before fusing with gait images.

4.3.1 CASIA Gait B Database

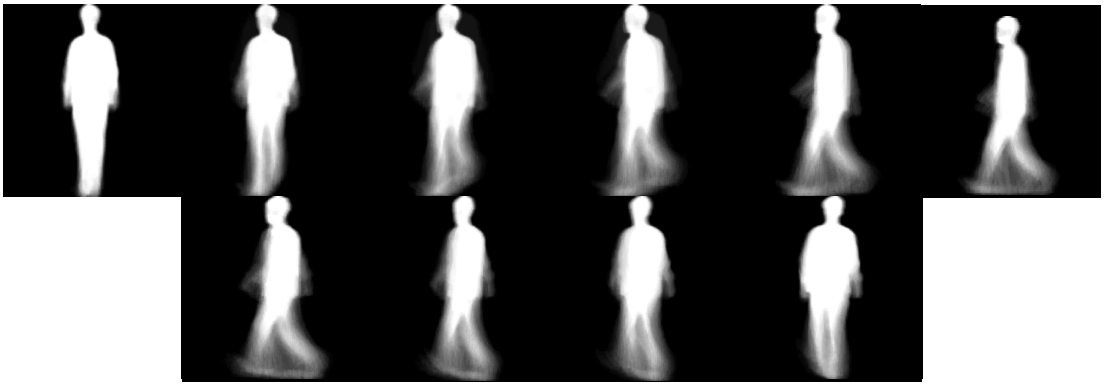
The CASIA Gait B Database is a substantial multi-view data set for gait recognition that was prepared in the year 2005 [37]. It included a total of 124 classes in which the gait information was captured from 11 viewing angles. There are basically three varieties in the database in particular which include: variation in viewing angle, variation in cloth types and the third one is the variation in carrying materials such as bags. Other than the video documents, regardless they also provide human outlines taken from video database. Some samples from the CASIA Gait B Database are provided in the figure 4.3 below.



(a) Sample Gait sequence with carrying bags



(b) Sample Gait sequence with wearing coat



(c) Sample Gait sequence with normal walking

Fig. 4.3 Samples from the CASIA Gait B Database

Figure 4.3(a) above displays a sample sequence from CASIA Gait B Database with the carrying bag in the hand of subject. Figure 4.3(b) above displays a sample sequence from CASIA Gait B Database with the subject wearing coat i.e. showing clothing variation. In the

end, figure 4.3(c) below displays the sample sequence from CASIA Gait B Database with the subject in normal walking condition.

4.3.2 ORL Face Database

The ORL Face Database includes an arrangement of pictures of faces captured between the years 1992 to 1994 at the Cambridge University Engineering Department [38]. The data-set was utilized as a part of the setting of a face acknowledgment venture completed in a joint effort with the Speech, Vision and Robotics Group. There are 10 unique pictures of every one of 40 different subjects. For a few individuals, the pictures were caught at various circumstances, differing in the light intensity, face appearances such as open or shut eyes, with smile or no smile, with spectacles or no spectacles.

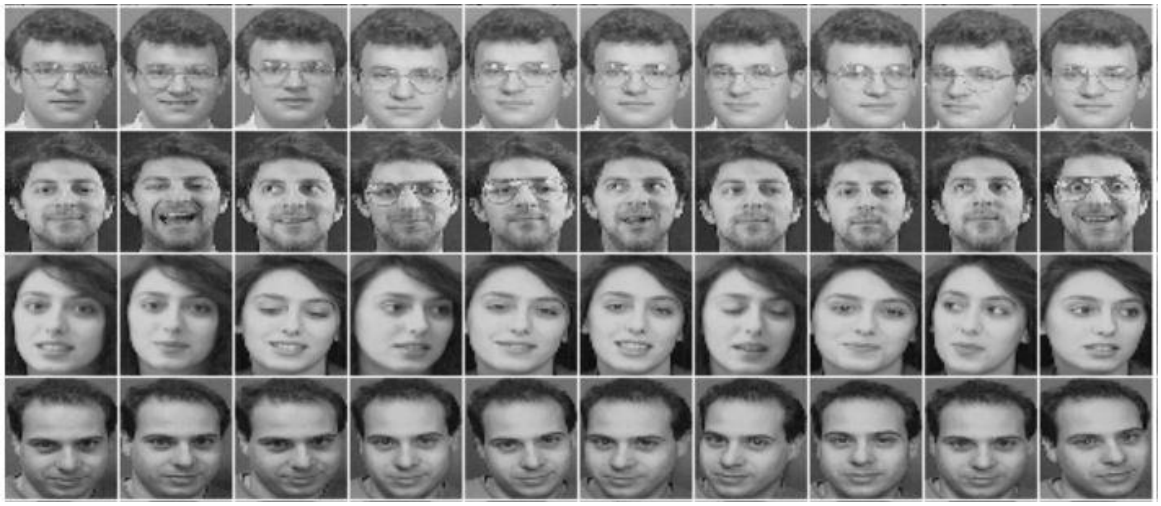


Fig. 4.4 Samples from the ORL Face Database

Every one of the pictures was taken in a dim uniform background with the individuals in a frontal and upright position. The images have PGM format with size of 92×112 pixels and 256 gray levels for every pixel. The figure 4.4 above shows some samples from the ORL Face Database which include images of 4 subjects from the database.

4.3.3 FEI Face Database

This is a Brazilian face data-set which includes an arrangement of pictures of faces captured between the years 2005 to 2006 at the AI Lab of FEI located at São Paulo, Brazil [42]. It

contains fourteen pictures for every one of 200 people, an aggregate of 2800 pictures. All pictures are coloured and captured in a white and uniform background with the individuals in a frontal and upright position with profile turn of around 180 degree maximum. Scale may fluctuate around 10% and each picture is of size 640×480 pixels. All the individuals in the database are understudies and faculties at FEI, in the vicinity of 19 to 40 years of age having different look, haircut, and embellishes. The count of female and male individuals is precisely equal and equivalent to 100. Likewise, they have also provided sub-set of the database made out of just frontal face pictures already adjusted to a typical layout with the goal that the pixel-wise highlights removed from the pictures relate generally to a similar area over all individuals. Figure 4.5 below shows some samples from the FEI Face Database.

Fig. 4.5 Samples from the FEI Face Database showing image variations

4.4 PRE-PROCESSING OF THE DATABASE

The pre-processing phase involves some of the important tasks which include face detection, resizing of images, alignment of face and gait images, normalisation of feature vectors and conversion of coloured images into grey scale images before fusion etc. The initial phase in the pre-handling stage is to discover the face region in the face pictures that means we first need to detect the faces from the images by using a suitable face detection algorithm to accomplish the ROI. There are a considerable number of techniques already available for face detection. Hence consequently, it isn't the point of focus of our work that how we are distinguishing faces. After the step of finding the faces, these detected faces are needed to be aligned in a proper manner which is our next pre-processing step. Side by side, we have also extracted the gait energy images from the gait sequences. Now the same work is also done with these GEIs to bring each of them into focus. After this, if the face images are coloured

images then they are changed over into grey level images so that they can be combined with the grey level gait energy images. From here, the next step in the pre-processing phase is resizing of the images hence, both types of images i.e. face and gaits are changed over into 100×100 pixel measure. It is carried out to decrease the computational time and cost of the framework.

4.4.1 Face Detection

The initial phase in the pre-handling stage is to discover the face region in the face pictures that means we first need to detect the faces from the images by using a suitable face detection algorithm to accomplish the ROI. There are a considerable number of techniques already available for face detection. Hence, face detection algorithm is not the focus of our proposed work that how we have detected the face. However, we have utilized the algorithm proposed by Viola and Jones for the detection of face regions from the images [45]. This helps us to get the region of interest for the images to perform recognition.

4.4.2 Face Alignment

After the step of finding the faces, these detected faces are needed to be aligned in a proper manner which is our next pre-processing step. Execution of face acknowledgment frameworks depends on the exactness of face arrangement and almost every face-alignment method depends upon the eye localization approaches. Consequently, we have utilized the eye localization algorithm for the alignment of face images to improve the recognition accuracy.

4.4.3 Conversion into Gait Energy Images

Side by side we also need to process the gait database by converting the gait sequence into the Gait Energy images. A gait energy image is generated by converting a series of walking picture outlines in a solitary grey picture containing the characteristics of all the frames of gait images of the sequence. The fundamental reason for generating this single grey level gait energy image is to lessen the processing time and to minimize the framework cost because a gait energy image convey the data of all the gait frames in a solitary picture which likewise

decreases the measure of database. The detailed mathematical explanation of gait energy images is given in section 4.2.

4.4.4 Resizing and Normalization

From here, the next step in pre-processing phase is resizing of the images hence, both the images i.e. face and gaits are changed over into 100×100 pixel measure. But before this, we need to convert the coloured images into the gray scale images if they are not already. This is done so that both the biometric features can be fused. After that normalization of these features is done in range of 0 to 1 to diminish the calculative part of the framework. The normalization tasks are needed to perform to expel variety impacts in the pictures. The landmarks focused on the face are used when pre-processing tasks are actualized.

4.5 FUSION OF FACE AND GAIT BIOMETRICS

As we have already discussed in the previous chapter that we chose these two types of biometrics, for fusion i.e. face and gait biometrics specifically because the elements altering the process of face recognition don't affect much the gait recognition process [56]. In the same way the characteristics altering the process of gait recognition don't affect much the face recognition process. Hence both the identification processes become robust to each other in this way which directly contributes to increasing the recognition accuracy of the model. Presently in our subsequent stage we will be fusing the face characteristics with the gait characteristics. For this reason, the characteristics for both the face and walking pictures are first changed over into vectors. After that normalization of these features is done in range of 0 to 1 to diminish the calculative part of the framework. Presently for each subject, first face picture and the relating walking frame is chosen for combination.

For combination, we have linked both component vectors by utilizing the vertical connection. Same above procedure is repeated for all the images for every class. Hence, finally we get a fusion vector of 200×100 pixels for each face and gait images which carry the data from both the biometric traits. From here, these melded vectors are given as input to the proposed multi-biometric deep learning model to train and test the model. Figure 4.6 demonstrates sample images of fused data where the CASIA Gait B Database is fused with ORL Face Database and figure 4.7 shows samples of fusion data where CASIA Gait B Database is fused with FEI Face Database.

Fig. 4.6 Sample fusion images of CASIA Gait B Database and ORL Face Database

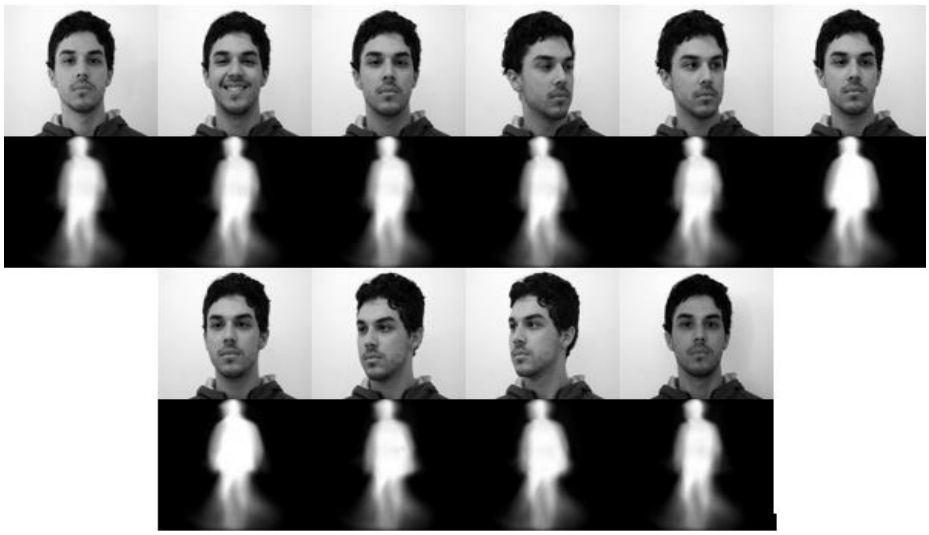
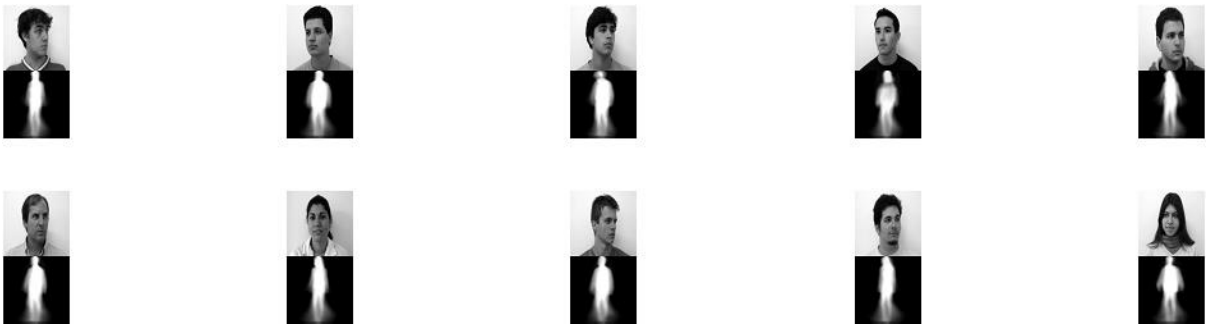


Fig. 4.7 Sample fusion images of CASIA Gait B Database and FEI Face Database

4.6 TRAINING METHODOLOGY

Our proposed 13-layer multi-biometric deep learning model is trained on the recent product of matlab i.e. matlab18a. Because the procedure of training needs a great deal of calculations and space henceforth, we have used NVIDIA GEFORCE GPU for this reason which have 8 GB RAM and Intel i7 processor. The training execution environment is right off the bat set to multi-GPU rather than single GPU because to train big datasets, it needs a ton of calculations. Our proposed DCNN display incorporates an input layer which takes the 200×100 pixel sized images as input. It also includes a few hidden layers or more precisely nine hidden layers

namely three sets of convolution, relu and max-pooling layers. After that there is a fully connected layer, a SoftMax layer and in the end a classification layer as we have just talked about in previous sections. Figure 4.8 below randomly demonstrates some of the training and testing samples fed to the proposed 13-layer deep convolutional neural network. training and test samples are non-overlapped.



(a) Random Samples for Training

(b) Random Samples for Testing

Fig. 4.8 Samples for training and testing

To begin with the first layer in our proposed DCNN model is an information layer or we say an input layer; where we characterize the measure of the sizing of pictures used for the training. For our situation, it is 200×100×1 for each picture. As our pictures are grey level pictures henceforth we have determined one at the last. Following layers are so called hidden layers which work as the feature extractor and also these layers reduce the dimensions of these features. Hidden layers are nine in number as discussed in previous section which includes the 3 sets of Convolutional, ReLu and Maxpooling layers. Where, at the convolutional layer the size of the filter is indicated together with the number of filters to be utilized for getting the feature maps. For the initial convolutional layer, a total of 32 quantities of filters having the filter size of [3 3] are taken. For the second and third

convolutional layer, total number of filters are taken to be as 64 of same size i.e. [3 3]. Next is the Relu actuation layer. This layer work as a non-linear activation function which maps the yield of the neuron to most elevated positive esteem and if the yield is negative then it maps it to zero as mentions in the 3rd section. Then, there comes the maximum pooling layer which is just used to take out overfitting. It also reduces the dimensions of the image features in addition to this by eliminating the less important features. Then, there comes a completely associated layer which is used to simply interface every one of the neurons of the last concealed layer with its neurons and after that to the Softmax layer. The output from fully associated layer is taken by the Softmax regression layer and then it figures the likelihood framework for every one of the subject's class and after that the yield is fed to the classification layer to coordinate every one of the query picture to their classes as per this likelihood grid. We have utilized the supervised learning technique has been utilized here in our work for the training purpose because we need to deal with the labelled databases of gait and face traits.

The training alternatives for all the experiments are set along these lines as mentioned below: training execution environment is taken as multi-gpu; the starting learn-rate is set to 0.0001; maximum epochs are set to 1000; L2Regularization is set to 0.0005; and a steady learning rate schedule is taken. Figure 4.9 and 4.10 shows the information about training progress like training time and training alternatives etc. for the first and second experiments (which will be explained in next section) respectively.

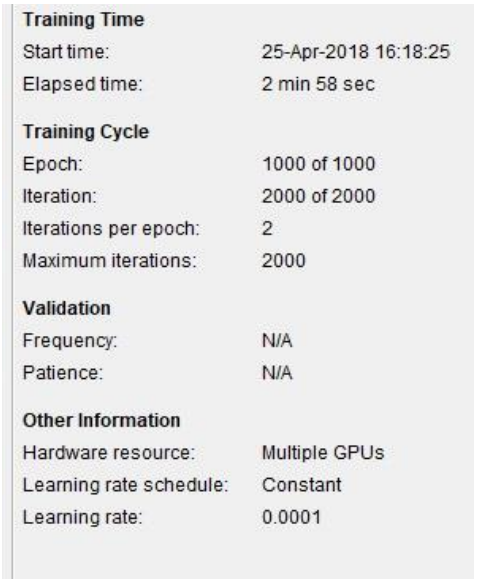


Fig. 4.9 Training information for ORL Face and CASIA Gait B fusion Database

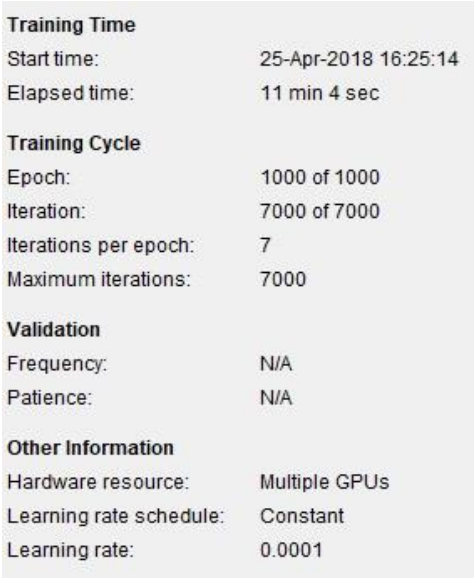
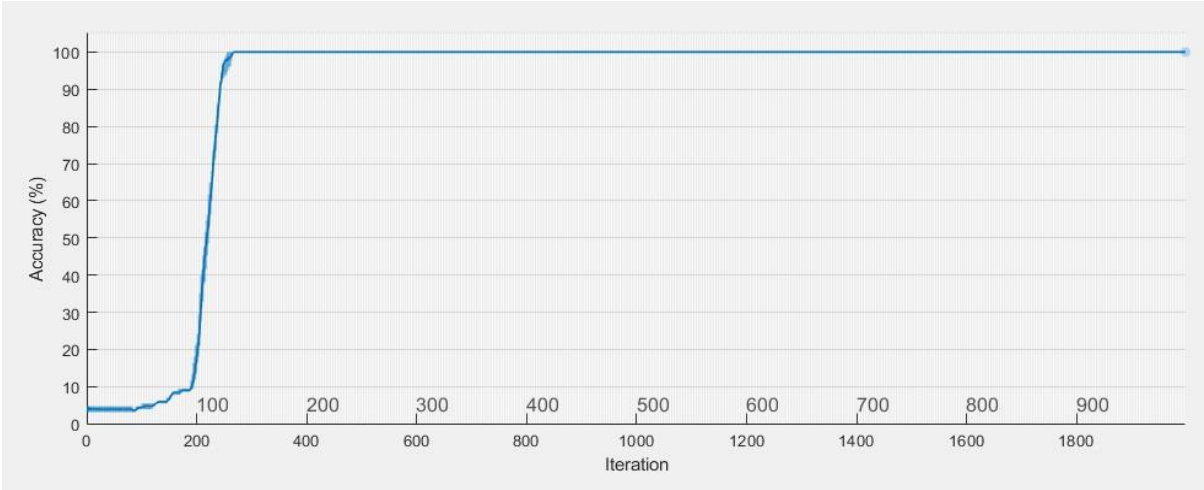


Fig. 4.10 Training information for FEI Face and CASIA Gait B fusion Database

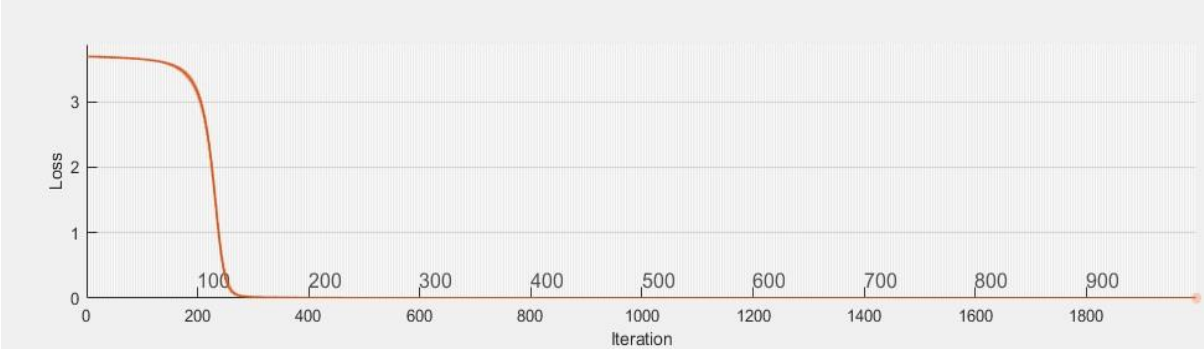
4.7 EXPERIMENTS AND RESULTS

Initially, a total of two experiments were carried out by utilizing the three databases mentioned above in the database description section i.e. CASIA Gait B Database, ORL Face Database and FEI Face Database. For the first experiment, we fused the face images of ORL Face Database with the corresponding gait images of CASIA Gait B Database where the initial 40 classes from both the datasets were selected for fusion. The evaluation protocol for training and testing our proposed DCNN model is selected as 4:1 (training: testing). For second experimental setup, we fused the face images of FEI Face Database with the corresponding gait images of CASIA Gait B Database. Here, we chose 124 subjects of FEI Face Database randomly to fuse with those of CASIA Gait B Database. Here also the evaluation protocol for training and testing our proposed DCNN model is selected as 4:1 (training: testing).

For both of the above setups, the training alternatives are set along these lines as follow: training execution environment is taken as multi-gpu; the starting learn-rate is set to 0.0001; maximum epochs are set to 1000; L2Regularization is set to 0.0005; and a steady learning rate schedule is taken. For both of the experiments, 100 % recognition accuracy is achieved when the model is tested upon the training data itself.



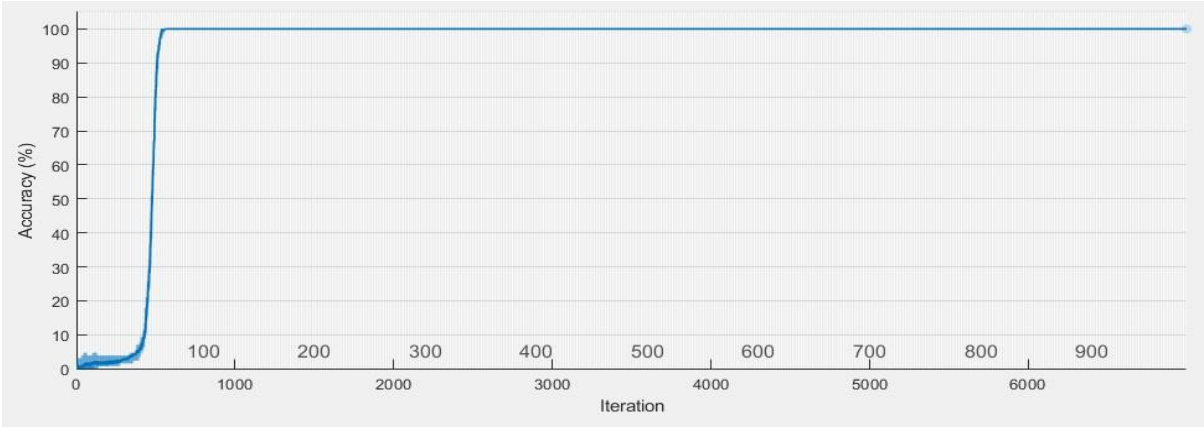
(a) Min-Batch Accuracy v/s Iterations Curve



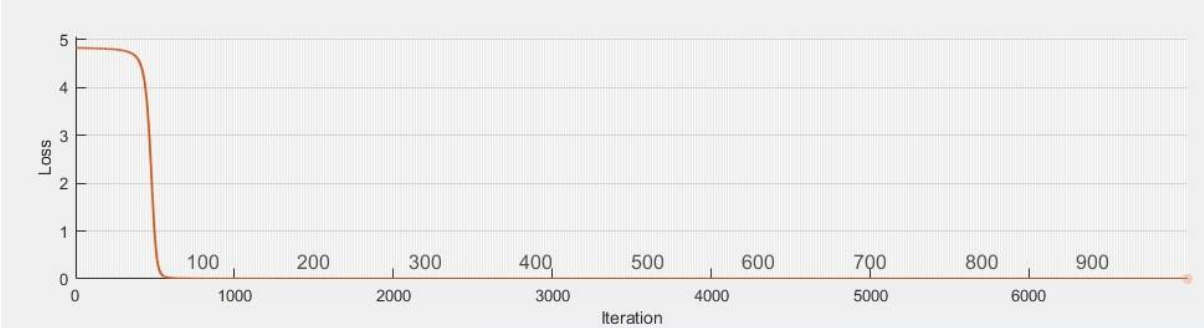
(b) Min-Batch Loss v/s Iterations Curve

Fig. 4.11 Training Progress Curves for (ORL Face Dataset + CASIA Gait Dataset B) fusion data: (a) Mini-batch Accuracy (b) Mini-batch Loss

During the training progress for the first experiment i.e. when the face images of ORL Face Database are fused with corresponding gait images of CASIA Gait B Database, a Min-Batch Accuracy of 100 % was reached till the 150th epoch and the Min-Batch Loss was decreased to 0.0003 till the 1000th epoch i.e. till the last one. This we can clearly see from training progress curves in figure 4.11 (a) and 4.11 (b). Similarly, for the second experiment i.e. when the face images of FEI Face Database are fused with corresponding gait images of CASIA Gait B Database, a Min-Batch Accuracy of 100 % was reached till the 79th epoch and the Min-Batch Loss was decreased to 0.0001 in the 1000th epoch. The Min-Batch Accuracy and Min-Batch Loss curves for 2nd experiment are shown in figure 4.12 (a) and figure 4.12 (b) respectively.



(a) Min-Batch Accuracy v/s Iterations Curve



(b) Min-Batch Loss v/s Iterations Curve

Fig. 4.12 Training Progress Curves for (FEI Face Database + CASIA Gait Dataset B) fusion data: (a) Mini-batch Accuracy (b) Mini-batch Loss

Training progress with respect to increasing number of epochs and iterations on ORL Face Dataset and CASIA Gait Dataset B fusion data is presented in table 4.1 below with their corresponding time elapse, minimum-batch accuracy and minimum-batch loss. The base learning rate for training the database is taken as 1.0000e-04 as shown in the table. Here, we can clearly see that minimum batch accuracy of 100% has been achieved in the 300th iteration or we can say in the 150th epoch. Also the minimum-batch loss has been reduced to 0.0003 for the first experiment.

Similarly, for 2nd experiment, the training progress with respect to increasing number of epochs and iterations on FEI Face Database and CASIA Gait Dataset B fusion data is presented in table 4.2 below with their corresponding time elapse, minimum-batch accuracy and minimum-batch loss. Here also the base learning rate for training the database is taken as 1.0000e-04. We can see from the table 4.2 that minimum batch accuracy of 100%

has been achieved in the 550th iteration that is in the 79th epoch. For the second experiment minimum-batch loss has been reduced to 0.0001 as we can see in the table 4.2 below.

Table 4.1 Training progress per iteration on test database for first experiment

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Mini-batch Loss	Base Learning Rate
1	1	00:00:02	5.47%	3.6883	1.0000e-04
50	100	00:00:11	3.91%	3.6593	1.0000e-04
100	200	00:00:20	10.94%	3.3248	1.0000e-04
150	300	00:00:29	100.00%	0.0242	1.0000e-04
200	400	00:00:38	100.00%	0.0062	1.0000e-04
250	500	00:00:47	100.00%	0.0034	1.0000e-04
300	600	00:00:57	100.00%	0.0023	1.0000e-04
350	700	00:01:06	100.00%	0.0017	1.0000e-04
400	800	00:01:15	100.00%	0.0014	1.0000e-04
450	900	00:01:24	100.00%	0.0011	1.0000e-04
500	1000	00:01:33	100.00%	0.0009	1.0000e-04
550	1100	00:01:43	100.00%	0.0008	1.0000e-04
600	1200	00:01:52	100.00%	0.0007	1.0000e-04
650	1300	00:02:01	100.00%	0.0006	1.0000e-04
700	1400	00:02:10	100.00%	0.0006	1.0000e-04
750	1500	00:02:19	100.00%	0.0005	1.0000e-04
800	1600	00:02:28	100.00%	0.0005	1.0000e-04
850	1700	00:02:37	100.00%	0.0004	1.0000e-04
900	1800	00:02:46	100.00%	0.0004	1.0000e-04
950	1900	00:02:55	100.00%	0.0004	1.0000e-04
1000	2000	00:03:04	100.00%	0.0003	1.0000e-04

Table 4.2 Training progress per iteration on test database for second experiment

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Mini-batch Loss	Base Learning Rate
1	1	00:00:00	0.00%	4.8237	1.0000e-04
50	350	00:00:33	3.91%	4.6882	1.0000e-04
79	550	00:00:52	100.00%	0.0530	1.0000e-04
100	700	00:01:06	100.00%	0.0075	1.0000e-04
150	1050	00:01:39	100.00%	0.0022	1.0000e-04
200	1400	00:02:12	100.00%	0.0012	1.0000e-04
250	1750	00:02:45	100.00%	0.0008	1.0000e-04
300	2100	00:03:18	100.00%	0.0006	1.0000e-04
350	2450	00:03:52	100.00%	0.0005	1.0000e-04
400	2800	00:04:25	100.00%	0.0004	1.0000e-04
450	3150	00:04:58	100.00%	0.0003	1.0000e-04
500	3500	00:05:31	100.00%	0.0003	1.0000e-04
550	3850	00:06:04	100.00%	0.0002	1.0000e-04
600	4200	00:06:38	100.00%	0.0002	1.0000e-04
650	4550	00:07:11	100.00%	0.0002	1.0000e-04
700	4900	00:07:44	100.00%	0.0002	1.0000e-04
750	5250	00:08:17	100.00%	0.0002	1.0000e-04
800	5600	00:08:50	100.00%	0.0002	1.0000e-04
850	5950	00:09:24	100.00%	0.0002	1.0000e-04
900	6300	00:09:57	100.00%	0.0001	1.0000e-04
950	6650	00:10:30	100.00%	0.0001	1.0000e-04
1000	7000	00:11:03	100.00%	0.0001	1.0000e-04

An identification percentage of 98.75 % is achieved when our model is tested upon test data prepared for first experiment i.e. when the face images of ORL Face Database are fused with

corresponding gait images of CASIA Gait B Database and an identification percentage of 97.50 % is achieved when it is tested upon test data prepared for the second experiment i.e. when the face images of FEI Face Database are fused with corresponding gait images of CASIA Gait B Database. Table 4.3 below shows the identification rate of the proposed DCNN model for both the experiments. Our model consumed only 2 min 58 sec in the 1st experiment and around 11 minutes in the 2nd experiment for training the entire database as shown in the table 4.4. It demonstrates the efficiency and fastness of our architecture. We can clearly see from the outcomes we got, that our model is superior to the other single biometric systems. This is because; the fusion features carry greater info than a single modality.

Table 4.3 Recognition accuracy of proposed DCNN model on test datasets

Database	**Accuracy**	**Mini-batch Loss**
ORL Face Dataset + CASIA Gait Dataset B	98.75 %	0.0003
FEI Face Database + CASIA Gait Dataset B	97.50 %	0.0001

Table 4.4 Average training time of proposed DCNN model

Database	**Number of subjects**	**Samples per subject**	**Average training time**
ORL Face Dataset + CASIA Gait Dataset B	40	10	2 min 58 sec
FEI Face Database + CASIA Gait Dataset B	124	10	11 min 4 sec

To further evaluate our proposed 13-Layer deep convolutional neural network algorithm, we have compared our work with several other existing state-of-art techniques. When contrasted with [60], where they could accomplish the error rate of around 0.0128, we have essentially lessened this error rate to only 0.0001 in our second experimental setup as shown in the table 4.3. Also, we compared our work with the work displayed in [59] where a maximum accuracy of around 95.5% could be accomplished, we have achieved a huge improvement to

this too as we have achieved a maximum accuracy of 98.75 % for the first experiment and a maximum accuracy of 97.50 % for the second experiment. Comparison of our proposed approach in contrast with rest of the existing approaches in this area has been presented in the table 4.5 below. We can clearly see from this table that our method has provided better results in contrast with other existing approaches in this field.

Table 4.5 Comparison of proposed DCNN model with some prior techniques

Approach	**Classification Method**	**Maximum Accuracy**
Jianzheng Liu [59]	Softmax Regression	95.50 %
A. Derbel [57]	PCA+Euclidean Distance	97.40 %
Xin Geng et al. [47]	Hausdorff Distance	86.67 %
Xiaoli Zhou et al. [44]	Nearest Neighbor	91.30 %
Xianglei Xing et al. [56]	KNN Search	98.25 %
Proposed Method	SoftMax Classifier	**98.75 %**

4.8 SUMMARY

In this chapter, we have presented the proposed work and the results obtained for the same. Initially, we have described the proposed algorithm layer by layer in detail. A detailed description of the databases is provided after that. After that the proposed algorithm is discussed step by step such as pre-processing of the databases, fusion of the face and gait biometrics and the training methodology utilised. In the end, proposed algorithm is tested upon the two fusion databases and the outcomes are compared with the existing techniques which outperformed almost all of the existing techniques in this field as shown in the table above.

CHAPTER 5
NOISE ATTACKS

5.1 INTRODUCTION

The same two experiments mentioned in the previous chapter, were carried out again but this time noise attacks were added in all the test images. We added three types of noise attacks to the test images of all the three datasets which are Salt and Pepper noise, Gaussian Noise and Speckle Noise. For reducing the noise effects from images, some image de-noising techniques were also applied to the test images which include mean filter and median filter. Here also, the evaluation protocol for training and testing the proposed neural network is selected as 4:1 (training: testing) for both the experiments. Training options are chosen as same for both of the experiment as specified above in the experiments and results section. To test our model for Salt and pepper noise, we added Salt and pepper noise with a noise density of 0.02 to all the test images of all three databases. Besides Salt and pepper noise, we tested our model after adding Gaussian noise to the test images with a zero mean and variance equal to 0.003. We also added the Speckle noise to the test images with a variance equal to 0.01.

5.2 SALT AND PEPPER NOISE

Salt-and-pepper noise is a type of noise now and again observed on pictures. It is otherwise called impulse noise. That type of noise could be generated by sharp and sudden unsettling influences in the picture. It presents itself as scantily happening white and dark pixels. A successful noise lessening strategy for this sort of noise is a median filter. We can also say that for the case of pixels, salt and pepper noise implies which are with high frequencies. Hence, for salt noise the estimations of this noise is high i.e. from 255 to 200, and for the pepper noise the estimations of this noise compose is low i.e. from 5 to 0 [48]. We think about salt-and-pepper noise, for which a specific measure of the pixels in the picture are either dark or white subsequently the name of this noise is salt and pepper. Salt-and-pepper noise can be utilized to display artefacts in the CCD or in the transmission of the picture. For the probability r (for $0 \leq r \leq 1$) that a pixel is debased, the salt-and-pepper noise in a picture can be introduced by setting a small amount of r/2 haphazardly chosen pixels to dark, and another portion of r/2 haphazardly chose pixels to white. Below command in equation 1 is

used to add salt and pepper noise to the images in matlab where D is the density of the salt and pepper noise, I is the input image in which noise has to be introduced and I1 is the output noisy image. In our experiments, we have added Salt and pepper noise with a noise density of 0.02 to all the test images of all three databases.

$$I1 = \text{imnoise}(I, 'salt \& pepper', D) \tag{5.1}$$

In our experiments, we tried some filters to handle and remove the effects of salt and pepper noise which are mean filter, Gaussian filter, and median filter. We passed each of the test images through above three filters before feeding to our proposed model for fusion and then tested the accuracy results. Out of all the three filters, the median filter provided the best classification accuracy percentage. Hence, we chosen to use the median filter to de-noise the images with salt and pepper noise for both the experiments conducted in our work.

5.2.1 Median Filter

Let us have a brief introduction about median filter. Median filtering is a non-linear technique utilized to expel noises from the pictures. This is generally utilized as it is extremely viable at expelling noises while protecting edges. It is especially powerful at expelling 'salt and pepper' type commotion. This filter processes by traveling through the picture pixel by pixel, supplanting each an incentive with the median benefit of neighbouring pixels. We call this pattern of neighbours as the "window", which slides, pixel by pixel over the whole picture 2 pixels, over the whole picture. The median is figured by initially arranging all the pixel terms from the window into numerical request, and after that supplanting the pixel being considered with the centre pixel term. The matlab command for median filter for filtering out the noise from the image is shown by the below equation i.e. equation 5.2. Here I1 is the input noisy image affected by the salt and pepper noise and J is the output image after removing the noise using median filtering.

$$J = \text{medfilt2}(I1) \tag{5.2}$$

Averaging of the neighbourhood pixels can stifle disengaged out-of-go noise, however the symptom is that it likewise obscures sudden changes, for example, line highlights, clear edges, and other picture subtle elements all comparing to high spatial frequencies. The

median filtering technique is a successful one that can, to some degree, recognize out-of-extend separated noise from true blue picture highlights, for example, lines and edges. In particular, the median filter changes the value of a pixel by the median of the neighbourhood pixels, rather than the average of all pixels in an area. A proper sorting algorithm is required for finding out the median of the set of neighbourhood pixel values [50]. The equation 5.3 given below represents the mathematical formula for median filter where w indicates the neighbourhood for the user pixel which is located at the location [m, n] in the picture.

$$y[a, b] = \text{median}\{x[i, j], (i, j)\} \in w \tag{5.3}$$

To calculate the output of the median filter, first of all we consider a pixel in the picture and its neighbourhood pixel values. Now to find out the median of these neighbourhood pixel values, these values are sorted in an increasing order of numbers and then the middle value is chosen from these values which is the median value and then the noisy pixel is replaced by this value. Figure 5.1 below shows an example to explain the working of median filter. Here 'I' is the image with a noisy pixel shown in figure 5.1 (a) where the bold and circled pixels i.e.150 represents the noisy pixel in the image.

I =

123	124	125	127	135
122	**124**	**126**	**127**	234
118	**120**	**150**	**125**	0
119	**115**	**119**	**123**	0
111	117	110	120	130

(a) Noisy Image

W =

124	126	127	120	150	125	115	119	123

(b) Window with noisy pixel

Wp =

115	119	120	123	**124**	125	126	127	150

(c) Processed window

Ip =

123	124	125	127	135
122	**124**	**126**	**127**	234
118	**120**	**124**	**125**	0
119	**115**	**119**	**123**	0
111	117	110	120	130

(d) Processed image

Fig. 5.1 Example of median filtering

W in the figure 5.1 (b) represents the window with the noisy pixel. Now pixels of this window are sorted in the increasing order of numbers and then the middle value of that sequence i.e. 124 is the median filter. The noisy pixel 150 is then replaced by this number in the image. The window Wp in figure 5.1 (c) represents the processed window and the image Ip is the output of the median filter referred to as the processed image in figure 5.1 (d).

5.2.2 Results for Salt and Pepper Noise

The evaluation protocol for training and testing the proposed DCNN is selected as 4:1 (training: testing) i.e. 20 percent of images are selected for testing and rest of images are used to train the network. Training options are chosen as same to that are specified above in the fourth chapter. Figure 5.2 below presents the images showing effects of Salt and Pepper noise and median filter on the test images where figure 5.2 (a) shows the sample test images, figure 5.2 (b) shows the test images after adding Salt and Pepper noise to them and finally figure 5.2 (c) shows the test images after de-noising them using median filter.

Figure 5.3 below presents the training progress curves for first experiment i.e. when ORL Face Dataset is fused with the CASIA Gait Dataset B in presence of salt and pepper noise to feed to the proposed DCNN for classification. Here figure 5.3 (a) shows the minimum-batch accuracy v/s iteration curve and figure 5.3 (b) shows the minimum-batch loss v/s iteration curve for 1st experiment in presence of Salt and pepper noise. We can analyse from the training progress curves that the minimum-batch accuracy has reached to 100% in the 300^{th} iteration.

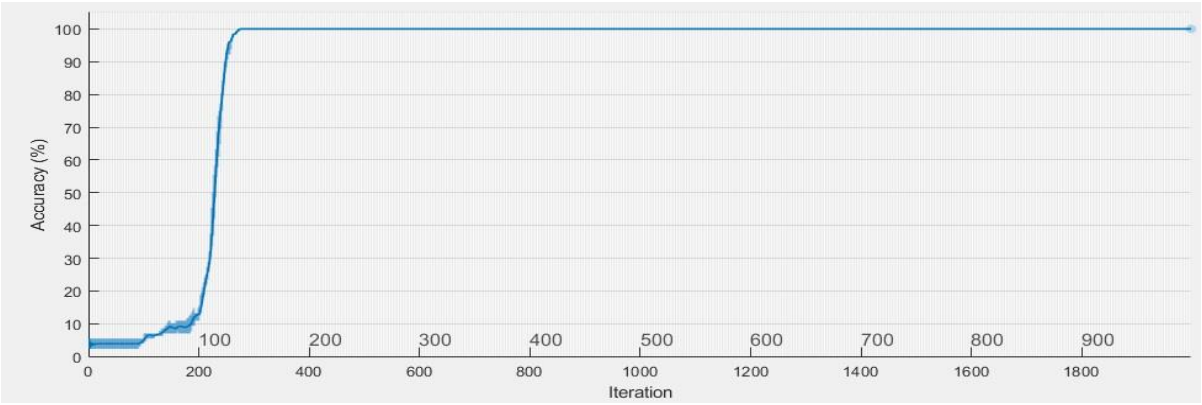
(a) Sample test images

(b) Images after adding Salt and Pepper noise

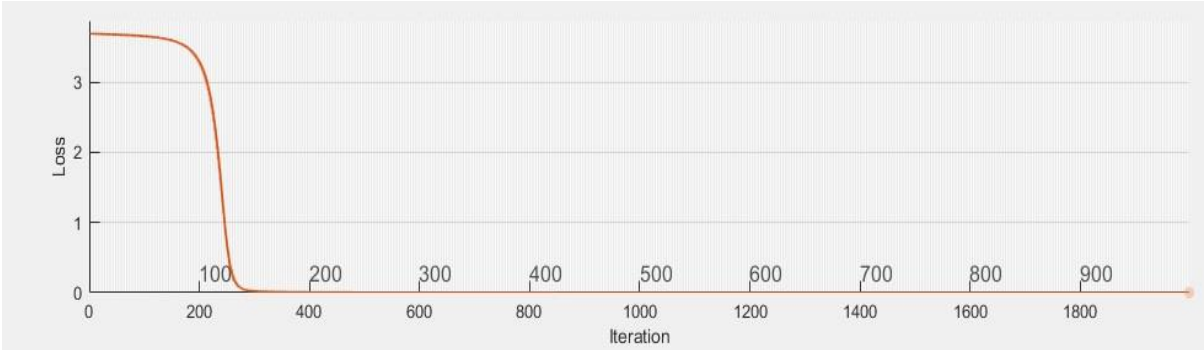
(c) De-noising with median filter

Fig. 5.2 Effects of salt and pepper noise and median filter on test images

Similarly, Figure 5.4 below presents the training progress curves for second experiment i.e. when FEI Face Dataset is fused with the CASIA Gait Dataset B in presence of salt and pepper noise to feed to the proposed DCNN for classification. Here figure 5.4 (a) shows the minimum-batch accuracy v/s iteration curve and figure 5.4 (b) shows the minimum-batch loss v/s iteration curve for 2nd experiment in presence of Salt and pepper noise. We can analyse from the curves that the minimum-batch accuracy has reached to 100% in the 550^{th} iteration.

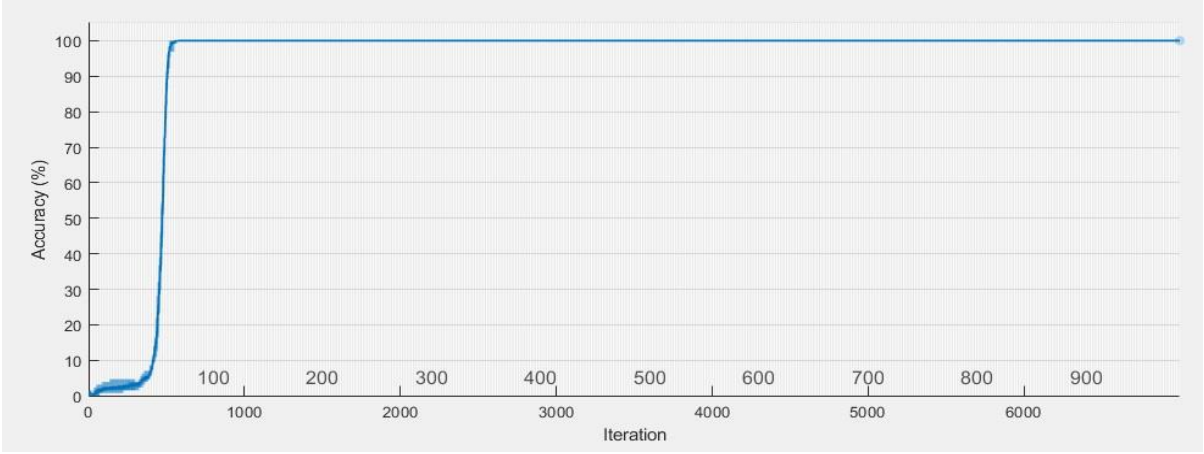


(a) Min-Batch Accuracy v/s Iterations

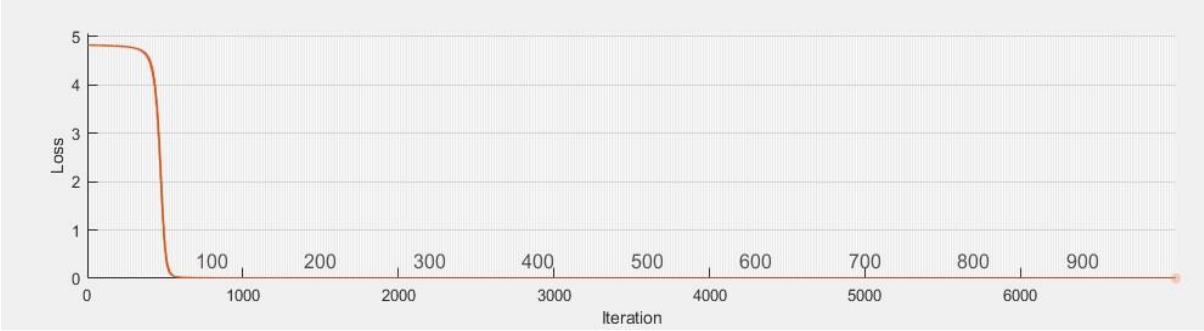


(b) Min-Batch Loss v/s Iterations

Fig. 5.3 Training Progress Curves for ORL Face Dataset + CASIA Gait Dataset B in presence of Salt and pepper noise



(a) Min-Batch Accuracy v/s Iterations



(b) Min-Batch Loss v/s Iterations

Fig. 5.4 Training Progress Curves for FEI Face Dataset + CASIA Gait Dataset B in presence of Salt and pepper noise

Table 5.1 Accuracy on test databases in presence of Salt and Pepper noise

Database	Accuracy	Mini-batch Loss	Average training time
ORL Face Dataset + CASIA Gait Dataset B	97.50 %	0.0003	3 min 1 sec
FEI Face Database + CASIA Gait Dataset B	97.18 %	0.0001	11 min 7 sec

Table 5.1 above shows the identification accuracy, minimum-batch loss and average training time for the first and second experiment when our DCNN model is tested upon test databases in presence of Salt and Pepper noise. For the first experiment, we are able to achieve the identification accuracy of 97.50 % by consuming an average training time of 3 minutes and 1 second whereas minimum-batch loss has been reduced to 0.0003. Similarly for the second experiment, we are able to achieve the identification accuracy of 97.18 % by consuming an average training time of 11 minutes and 7 seconds whereas minimum-batch loss has been reduced to 0.0001.

5.3 GAUSSIAN NOISE

In simple words, Gaussian noise is measurable artefact which has a PDF equivalent to the ordinary distribution and this noise has also the name Gaussian distribution. As such, the terms that this type of noise can go up against are Gaussian-disseminated. The white Gaussian noise is the exceptional case for Gaussian noise where the terms at some time pair are uncorrelated and distributed indistinguishably [53]. In the digital pictures, the main source of generation of this type of noise is due to the process of acquisition of images. For example the sensor commotion is created because of the bad lighting or potentially increased temperature and also due to the transmission etc. The equation 5.4 below shows the matlab command to generate the Gaussian Noise in the digital images where ‘a’ indicates the mean and vg indicates the variance for the Gaussian Noise. To test the accuracy of our proposed DCNN model we have added the Gaussian noise to the test images with a zero mean and variance equal to 0.003 for both the experiments.

$$\mathrm{Ig} = \mathrm{imnoise}\left(\mathrm{I}, '\,\mathrm{gaussian}', \mathrm{a}, \mathrm{v_g}\right) \tag{5.4}$$

In our experiments, we tried some filters to handle and remove the effects of Gaussian noise which are mean filter, Gaussian filter, and median filter. Out of all the three filters, the mean filter provided the best classification accuracy percentage for Gaussian noise. Hence, we chosen to use the mean filter to de-noise the images with Gaussian noise for both the experiments conducted in our work.

5.3.1 Mean Filter

Mean filter approach is a straightforward, natural and simple to actualize technique to smoothen the pictures, i.e. diminishing the measure of variation in intensity between two pixels. Usually mean filter is utilized to decrease commotion in pictures. Mean filter just changes every pixel term in a picture with the average or mean estimation of its neighbouring pixels including itself. It eliminates those pixels that are unrepresentative to their environment. Concept of mean filter can be normally understood as a convolution channel. Similar to different convolutions it is also dependent on a kernel representing the size and shape of the area to be tested while computing the mean.

$$I1 = \text{conv2}\left(I, \frac{\text{ones}(3)}{9}, 'same'\right) \tag{5.5}$$

In general, a 3×3 size filter kernel is utilized but we can also utilize bigger ones such as 5×5 kernels for more extreme smoothing. A little kernel can be connected more than once keeping in mind the end goal to create a comparable however not indistinguishable impact as a single pass with a bigger kernel. The equation 5.5 above represents the matlab command for mean filter utilized in our proposed architecture. Here 'I' is the input noisy image affected by the Gaussian noise and I1 is the output image after removing the noise using mean filtering approach.

1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

Fig. 5.5 3x3 kernel

Processing of the mean filter is as simple as the convolution process. Simply the computation of direct convolution of a picture with its kernel completes the mean filtering operation. Figure 5.4 above represents a kernel of size 3x3 for mean filter.

5.3.2 Results for Gaussian Noise

Here also the evaluation protocol for training and testing the proposed DCNN is selected as 4:1 (training: testing) i.e. 20 percent of images are selected for testing and rest of images are used to train the network. Training options are chosen as same to that are specified above in the fourth chapter. Figure 5.6 below presents the images showing effects of Gaussian noise and mean filter on the test images where figure 5.6 (a) shows the sample test images, figure 5.6 (b) shows the test images after adding Gaussian noise to them and finally figure 5.6 (c) shows the test images after de-noising them using mean filter.

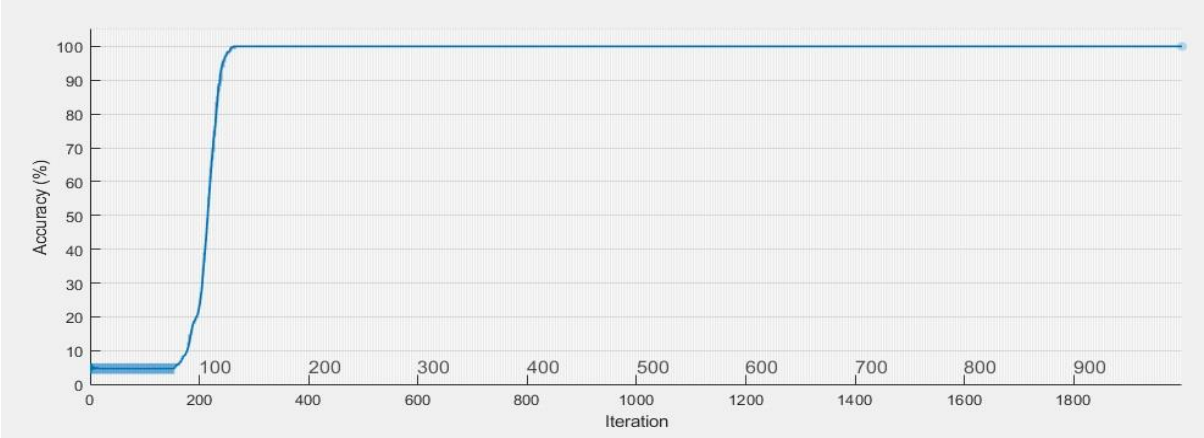
(a) Sample test images

(b) Images after adding Gaussian noise

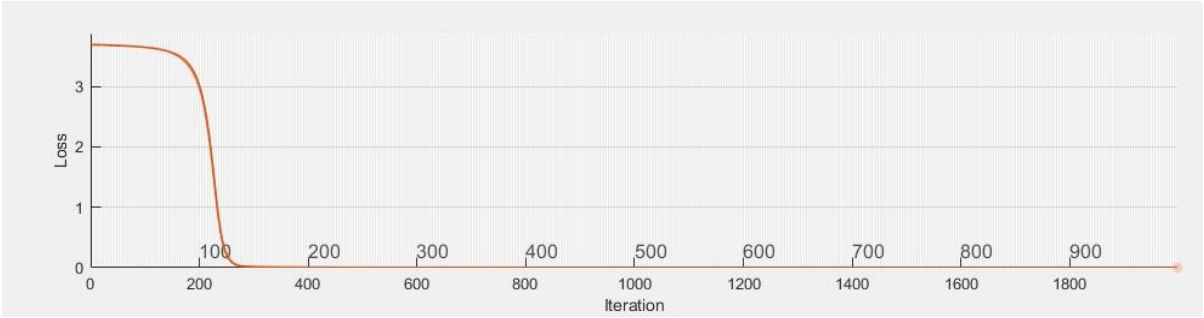
(c) De-noising with mean filter

Fig. 5.6 Effects of Gaussian noise and mean filter on test images

Figure 5.7 below presents the training progress curves for first experiment i.e. when ORL Face Dataset is fused with the CASIA Gait Dataset B in presence of Gaussian noise to feed to the proposed DCNN for classification. Here figure 5.7 (a) shows the minimum-batch accuracy v/s iteration curve and figure 5.7 (b) shows the minimum-batch loss v/s iteration curve for 1st experiment in presence of Gaussian noise. We can analyse from the curves that the minimum-batch accuracy has reached to 100% in the 300th iteration.



(a) Min-Batch Accuracy v/s Iterations

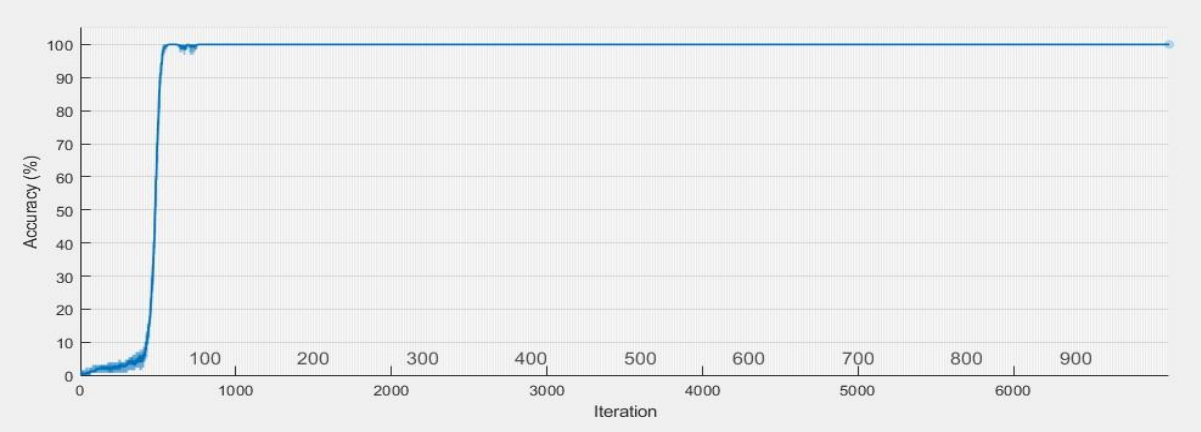


(b) Min-Batch Loss v/s Iterations

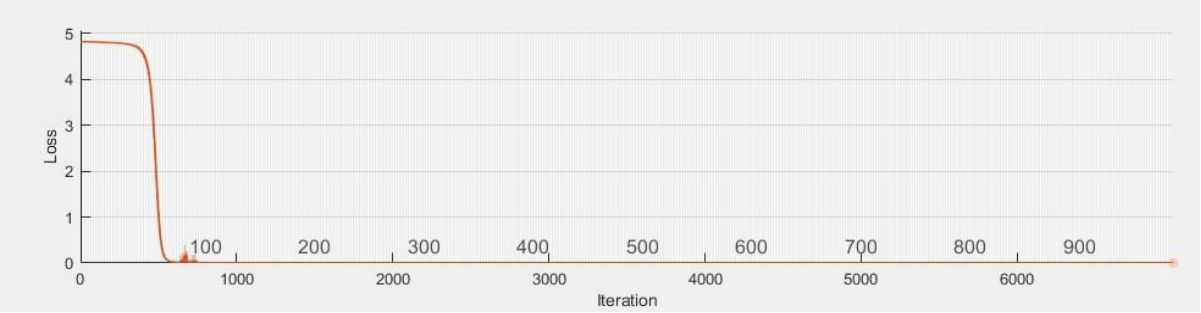
Fig. 5.7 Training Progress Curves for ORL Face Dataset + CASIA Gait Dataset B in presence of Gaussian noise

Similarly, Figure 5.8 below presents the training progress curves for second experiment i.e. when FEI Face Dataset is fused with the CASIA Gait Dataset B in presence of Gaussian noise to feed to the proposed DCNN for classification. Here figure 5.8 (a) shows the minimum-batch accuracy v/s iteration curve and figure 5.8 (b) shows the minimum-batch loss

v/s iteration curve for 2nd experiment in presence of Gaussian noise. We can analyse from the curves that the minimum-batch accuracy has reached to 100% in the 600th iteration.



(a) Min-Batch Accuracy v/s Iterations



(b) Min-Batch Loss v/s Iterations

Fig. 5.8 Training Progress Curves for FEI Face Dataset + CASIA Gait Dataset B in presence of Gaussian noise

Table 5.2 Accuracy on test databases in presence of Gaussian noise

Database	Accuracy	Mini-batch Loss	Average training time
ORL Face Dataset + CASIA Gait Dataset B	93.75 %	0.0004	3 min 0 sec
FEI Face Database + CASIA Gait Dataset B	95.97 %	0.0001	11 min 8 sec

Table 5.2 above shows the identification accuracy, minimum-batch loss and average training time for the first and second experiment when our DCNN model is tested upon test databases in presence of Gaussian noise. For the first experiment, we are able to achieve the identification accuracy of 93.75 % by consuming an average training time of 3 minutes whereas minimum-batch loss has been reduced to 0.0004. Similarly for the second experiment, we are able to achieve the identification accuracy of 95.97 % by consuming an average training time of 11 minutes and 8 seconds whereas minimum-batch loss has been reduced to 0.0001.

5.4 SPECKLE NOISE

This is a kind of information drop-out noise that is also known by the name intensity spikes. This kind of noise is generated due to the mistakes occurred during the transmission of information. The affected pixels due to speckle noise are set to the most extreme pixel values giving a snow like resemblance in the picture. This type of noise is a granular noise that naturally resides in the images and downgrades the nature of the computerized pictures. It also debases the nature of the SAR, dynamic radar and medical ultrasound pictures. This type of commotion is generally measured by the level of pixels which are affected or undermined. The equation 5.6 below shows the matlab command to generate the Speckle Noise in the digital images where vg indicates the variance for the Speckle noise. Here, 'I' is the input image without any noise and I1 is the output image after adding the Speckle noise to the image.

$$I1 = \text{imnoise}(I,' \text{speckle}', Vs) \tag{5.6}$$

In our experiments, we tried some filters to. Out of all the three filters, the mean filter provided the best classification accuracy percentage for Speckle noise. Hence, we chosen to use the mean filter to de-noise the images affected with Speckle noise for both the experiments conducted in our work. We have already discussed about the mean filters in detail in the above section handle and remove the effects of Speckle noise which are mean filter, Gaussian filter, and median filter.

5.4.1 Results for Speckle Noise

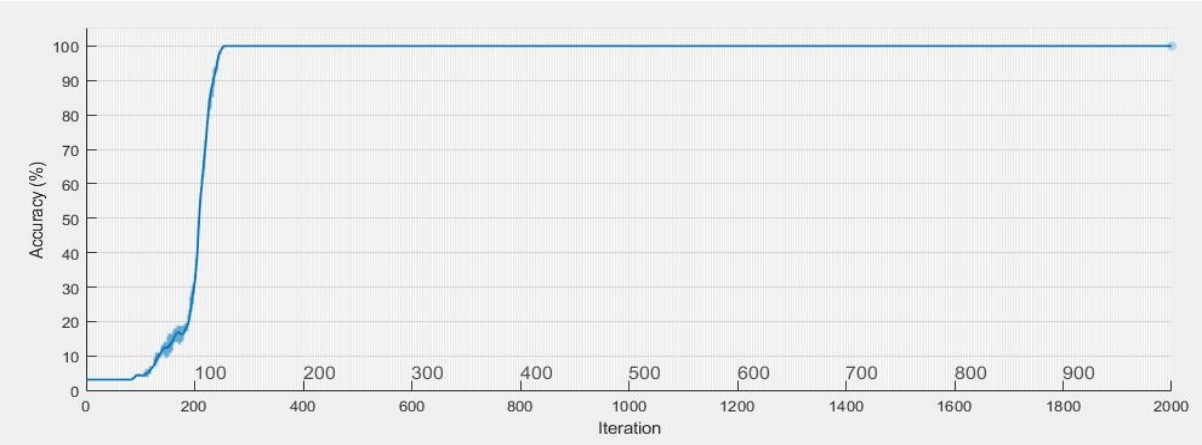
Here also, we have utilized the same evaluation protocol for training and testing the proposed DCNN that is 4:1 (training: testing) that means 20 percent of images are selected for testing and 80 % of the images are utilized for training the deep neural network. Training options are also kept the same as specified above in the Salt and Pepper noise and Gaussian noise sections. Figure 5.9 below presents the images showing the effects of the Speckle noise and the mean filter on the test images where figure 5.9 (a) shows the sample test images, figure 5.6 (b) shows the test images after adding Speckle noise to them and finally figure 5.6 (c) shows the test images after de-noising these images using the mean filter. Figure 5.10 presents the training progress curves for the first experiment i.e. when ORL Face Dataset is fused with the CASIA Gait Dataset B in presence of the Speckle noise to feed to the proposed DCNN for classification.

(a) Sample test images

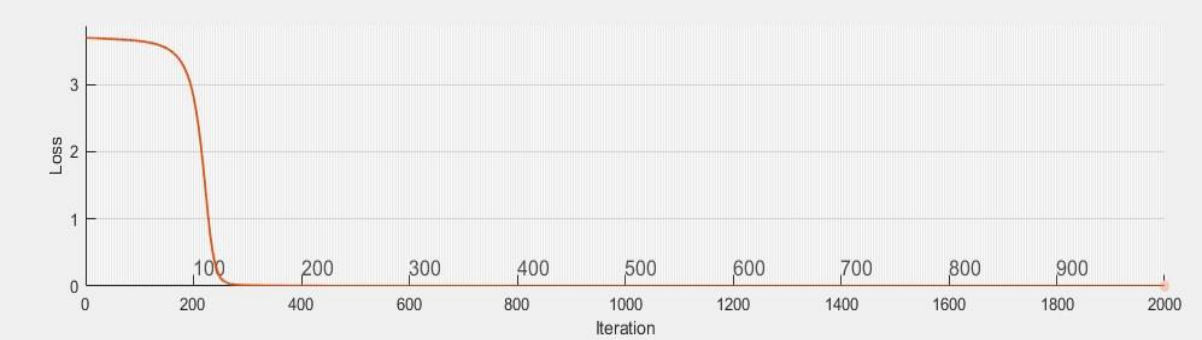
(b) Images after adding Speckle Noise

(c) De-noising with mean filter

Fig. 5.9 Effects of Speckle noise and mean filter on test images



(a) Min-Batch Accuracy v/s Iterations



(b) Min-Batch Loss v/s Iterations

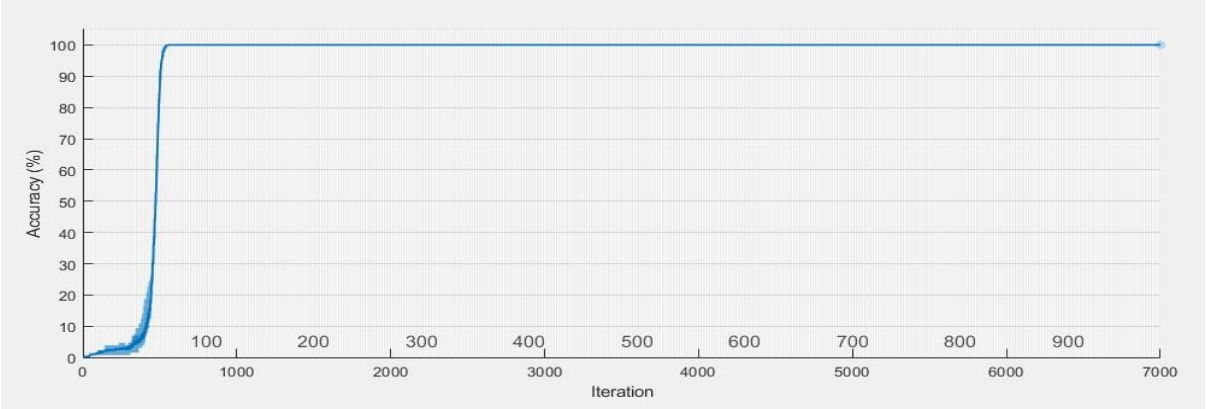
Fig. 5.10 Training Progress Curves for ORL Face Dataset + CASIA Gait Dataset B in presence of Speckle noise

Table 5.3 Accuracy on test databases in presence of Speckle noise

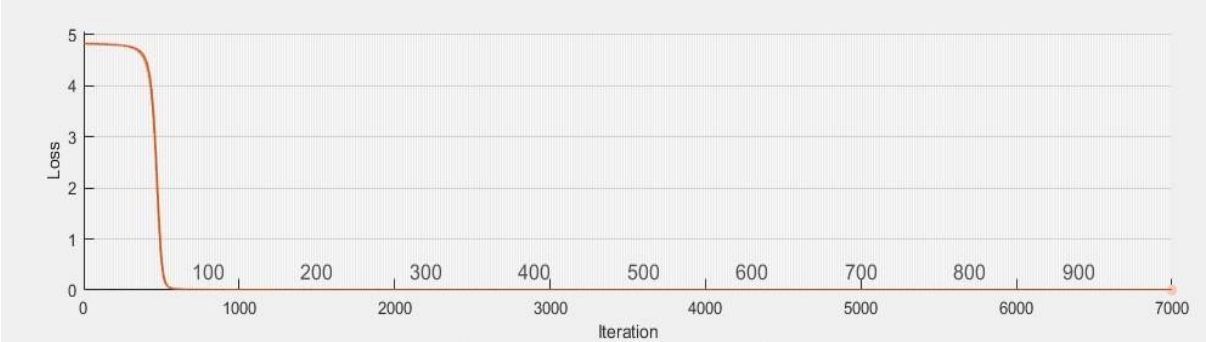
Database	**Accuracy**	**Mini-batch Loss**	**Average training time**
ORL Face Dataset + CASIA Gait Dataset B	95.00%	0.0003	2 min 59 sec
FEI Face Database + CASIA Gait Dataset B	95.56 %	0.0001	11 min 10 sec

Here figure 5.10 (a) shows the minimum-batch accuracy v/s iteration curve and figure 5.10 (b) shows the minimum-batch loss v/s iteration curve for 1st experiment in presence of

Speckle noise. We can analyse from the curves that the minimum-batch accuracy has reached to 100% in the 300th iteration. Similarly, Figure 5.11 below presents the training progress curves for second experiment i.e. when FEI Face Dataset is fused with the CASIA Gait Dataset B in presence of the Speckle noise to feed to the proposed 13-layer deep convolutional neural network for classification.



(a) Min-Batch Accuracy v/s Iterations



(b) Min-Batch Loss v/s Iterations

Fig. 5.11 Training Progress Curves for FEI Face Dataset + CASIA Gait Dataset B in presence of Speckle noise

Table 5.3 above shows the identification accuracy, minimum-batch loss and the average training time for the first and second experiment when our proposed 13-layer deep convolutional neural network is tested upon the test databases in presence of the Speckle noise. For the first experiment i.e. when ORL Face Dataset is fused with the CASIA Gait Dataset B in presence of the Speckle noise, we are able to achieve the identification accuracy of 95 % by consuming an average training time of 2 minutes and 59 seconds whereas the

minimum-batch loss has been reduced to 0.0003 for the same experiment. Similarly for the second experiment i.e. when FEI Face Dataset is fused with the CASIA Gait Dataset B in presence of the Speckle noise, we are able to achieve the identification accuracy of 95.56 % by consuming an average training time of 11 minutes and 10 seconds whereas the minimum-batch loss has been reduced to 0.0001.

5.5 SUMMARY

In this chapter, experimental results for the proposed DCNN model are presented against some common noise attacks such as salt and pepper noise, Gaussian noise and speckle noise. We have compared all our noise attacks outcomes with few existing techniques in this field utilizing the same three noises that we have utilized in our experiments and then same filters to de-noise the images i.e. median and mean filters. Our results outperformed all those existing techniques in presence of noise. We compared our noise results with [12], where they could achieve a recognition accuracy of only 90.35 %, 91.40 % and 93.20 % in presence of Salt and Pepper noise, Gaussian Noise and Speckle Noise respectively whereas we achieved a recognition accuracy of 97.5 %, 93.75 % and 95 % respectively for the same.

CHAPTER 6
CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION

This research work is concentrated on the issues related to multi-biometric fusion. But simple recognition algorithms cannot handle the large datasets hence deep learning comes out to be a great remedy to overcome this problem. In our work, we have introduced a novel technique of biometric fusion for individual recognition at a distance utilizing a 13-layer Deep Convolutional Neural Network model where the face and gait biometric traits are chosen for fusion. To extricate the gait characteristics, we have utilized the concept of Gait Energy Images where the entire gait cycle is converted in a single grey level GEI. For the face biometrics, face images are passed through face detection algorithm first of all and then alignment of face images is done to get the ROI. Some pre-processing steps are applied to both the biometrics before carrying out the fusion step such as resizing of the images and normalization in the range of 0 to 1 to reduce the computational complexity of the proposed architecture. The resulting face images are also converted into gray scale images if they are not already to fuse with the gray scale GEIs.

To fuse the face and gait features, both the images are 1st converted into feature vectors of same size and then these vectors are connected utilizing the vertical concatenation which is later fed to the proposed 13-layer DCNN model for classification and verification. The proposed DCNN model is composed of 13 layers including an image input layer, nine hidden layers, one fully connected layer, a softmax regression classifier and in the end a classification layer for classification of images. The hidden layers are composed of three triplets of convolution layer, ReLU layer and max-pooling layer connected in sequence. The proposed algorithm is tested upon three publically available databases: ORL Face Dataset, FEI Face Database and CASIA Gait Dataset B.

- When ORL Face Dataset is fused with CASIA Gait Dataset B, we achieved a recognition accuracy of 98.75 % and when FEI Face Database is fused with CASIA Gait Dataset B, we achieved a recognition accuracy of 97.50 %. We have done a significant improvement to the work presented in [59] where a maximum accuracy of only 95.5% could be achieved. As compared to [60], where they were able to achieve the error rate of 0.0128,

we have significantly reduced the error rate to 0.0001. We have outperformed other existing techniques in this field proposed by Xianglei Xing et al. [56], Xiaoli Zhou et al. [44], Xin Geng et al. [47] and A. Derbel [57].

- We have also tested our model after adding three well known noise attacks to both face and gait test images which are Salt and Pepper noise, Gaussian Noise and Speckle Noise. When ORL Face Dataset is fused with CASIA Gait Dataset B, a recognition accuracy of 97.5 %, 93.75 % and 95 % is achieved in presence of Salt and Pepper noise, Gaussian Noise and Speckle Noise respectively. Similarly when FEI Face Database is fused with CASIA Gait Dataset B, a recognition accuracy of 97.18 %, 95.97 % and 95.56 % is achieved in presence of Salt and Pepper noise, Gaussian Noise and Speckle Noise respectively.
- We utilized the median filter to de-noise the images affected with Salt and Pepper noise and mean filter was utilized to de-noise the images affected with Gaussian Noise and Speckle Noise both. Our results outperformed the work proposed in [12], where they could achieve a recognition accuracy of only 90.35 %, 91.40 % and 93.20 % in presence of Salt and Pepper noise, Gaussian Noise and Speckle Noise respectively.

6.2 FUTURE SCOPE

In future, we will try to combine our proposed deep convolutional neural network model with some of the existing popular feature fusion algorithms such as Canonical Correlation Analysis (CCA) [56] and Discriminant Correlation Analysis (DCA) [55] etc. for further improving the identification rates. We can further carry out our research work by testing our DCNN model on bigger datasets which contain the pictures which are already noisy and aren’t pre-processed.

REFERENCES

[1] Lawrence S *et al.* (1997). Face recognition: A convolutional neural-network approach, IEEE transactions on neural networks, 8(1), 98-113.

[2] Alsaadi IM (2015). Physiological biometric authentication systems, advantages, disadvantages and future development: a review, *International Journal of Scientific & Technology Research*, *4*(8), 285-289.

[3] Khalajzadeh H, Mansouri M, and Teshnehlab M (2013). Hierarchical structure based convolutional neural network for face recognition, *International Journal of Computational Intelligence and Applications*, 12(03), 1350018.

[4] Wikipedia. Biometrics. Available at http://en.wikipedia.org/wiki/Biometric.

[5] Zewail R et al. (2004). Soft and hard biometrics fusion for improved identity verification, *47th Midwest Symposium on Circuits and Systems, MWSCAS'04*, IEEE, Vol. 1, pp. I-225.

[6] Toufiq R and Islam MR (2014). Face recognition system using PCA-ANN technique with feature fusion method, *2014 IEEE International Conference in Electrical Engineering and Information & Communication Technology (ICEEICT)*, pp. 1-5.

[7] Jain AK, Ross A and Prabhakar S (2004). An introduction to biometric recognition, *IEEE Transactions on circuits and systems for video technology*, *14*(1), 4-20.

[8] Poh N and Kittler J (2010). Multimodal information fusion, In *Multimodal Signal Processing*, pp. 153-169.

[9] Ye X *et al.* (2015). Deep learning network for face detection, *2015 IEEE 16th International Conference in Communication Technology (ICCT)*, pp. 504-509.

[10] Hinton GE and Salakhutdinov RR (2006). Reducing the dimensionality of data with neural networks, *science*, *313*(5786), 504-507.

[11] Xiong C *et al.* (2015). Conditional convolutional neural network for modality-aware face recognition, *2015 IEEE International Conference in Computer Vision (ICCV)*, pp. 3667-3675.

[12] Budiman I *et al.* (2016). The effective noise removal techniques and illumination effect in face recognition using Gabor and Non-Negative Matrix Factorization, *IEEE International Conference in Informatics and Computing (ICIC)*, pp. 32-36.

[13] Ding Y *et al.* (2017). Noise-resistant network: a deep-learning method for face recognition under noise, *EURASIP Journal on Image and Video Processing*, *2017*(1), 43.

[14] Grm K. *et al.* (2017). Strengths and weaknesses of deep learning models for face recognition against image degradations, *IET Biometrics*, *7*(1), 81-89.

[15] Aiman U and Vishwakarma VP (2017). Face recognition using modified deep learning neural network, *2017 8th IEEE International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-5.

[16] Khalajzadeh H, Mansouri M, and Teshnehlab M (2014). Face recognition using convolutional neural network and simple logistic classifier, *Soft Computing in Industrial Applications*, Springer, Cham, pp. 197-207.

[17] Lu Z, Jiang X, and Kot A (2017). Enhance deep learning performance in face recognition, *2017 2nd IEEE International Conference on Image, Vision and Computing (ICIVC)*, pp. 244-248.

[18] Coşkun M *et al.* (2017). Face Recognition Based on Convolutional Neural Network, *2017 International IEEE Conference on Modern Electrical and Energy Systems (MEES).*

[19] Gupta V. Understanding feedforward neural networks. Available at https://www.learnopencv.com/understanding-feedforward-neural-networks/ (Accessed on 10th June 2018).

[20] Yin X and Liu X (2017). Multi-Task Convolutional Neural Network for Pose-Invariant Face Recognition, *IEEE Transactions on Image Processing*.

[21] Yoo JH *et al.* (2008). Automated human recognition by gait using neural network, *IEEE First Workshop on* Image *Processing Theory, Tools and Applications, IPTA 2008*, pp. 1-6.

[22] Hossain E and Chetty G (2013). Multimodal feature learning for gait biometric based human identity recognition, *International Conference on Neural Information Processing*, Springer, Berlin, Heidelberg, pp. 721-728.

[23] Ouyang W, Chu X, and Wang X (2014). Multi-source deep learning for human pose estimation, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2329-2336.

[24] Gupta DS. Fundamentals of deep learning – Introduction to recurrent neural networks. Available at https://www.analyticsvidhya.com/blog/2017/12/introduction-to-recurrent-neural-networks/ (Accessed on 11th June 2018).

[25] Shiraga K *et al.* (2016). GEINet: View-invariant gait recognition using a convolutional neural network, *2016 International IEEE Conference* on *Biometrics (ICB)*, pp. 1-8.

[26] Nair BM and Kendricks KD (2016). Deep network for analyzing gait patterns in low resolution video towards threat identification. *Electronic Imaging*, *2016*(11), 1-8.

[27] Das D and Chakrabarty A (2016). Human Gait Recognition using Deep Neural Networks, In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, ACM, p. 132.

[28] Mazur M. A step by step backpropagation example. Available at https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/ (Accessed on 11th June 2018).

[29] Thapar D *et al.* (2018). VGR-net: A view invariant gait recognition network, *2018 IEEE 4th International Conference* on *Identity, Security, and Behavior Analysis (ISBA)*, pp. 1-8.

[30] Krizhevsky A, Sutskever I and Hinton GE (2012). ImageNet classification with deep convolutional neural networks, In *Advances in neural information processing systems*, pp. 1097-1105.

[31] McCulloch WS and Pitts W (1943). A logical calculus of the ideas immanent in nervous activity, *The bulletin of mathematical biophysics*, *5*(4), 115-133.

[32] Al-Waisy AS *et al.* (2017). A multi-biometric iris recognition system based on a deep learning approach, *Pattern Analysis and Applications*, 1-20.

[33] Tieu NDT *et al.* (2017). An approach for gait anonymization using deep learning, *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pp. 1-6.

[34] Li C *et al.* (2017). Deepgait: a learning deep convolutional representation for view-invariant gait recognition using joint Bayesian, *Applied Sciences*, *7*(3), 210.

[35] Zeng R *et al.* (2015). Quaternion softmax classifier, *Electron Lett IET*, 50(25), 1929–1930.

[36] Dehzangi O, Taherisadr M and Changalvala R (2017). IMU-based gait recognition using convolutional neural networks and multi-sensor fusion, *Sensors*, *17*(12), 2735.

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

[38] Guo G, Li SZ and Chan K (2000). Face recognition by support vector machines, *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 196-201.

[39] Liu W *et al.* (2018). Learning Efficient Spatial-Temporal Gait Features with Deep Learning for Human Identification, *Neuroinformatics*, 1-15.

[40] Lv Z *et al.* (2015). Class energy image analysis for video sensor-based gait recognition: A review, *Sensors*, *15*(1), 932-964.

[41] Huang PS, Harris CJ and Nixon MS (1998). Visual surveillance and tracking of humans by face and gait recognition, *IFAC Proceedings Volumes*, *31*(29), 113-118.

[42] Thomaz CE and Giraldi GA (2010). A new ranking method for principal components analysis and its application to face image analysis, *Image and Vision Computing*, *28*(6), 902-913.

[43] Zhou X and Bhanu B (2006). Integrating face and gait for human recognition, *IEEE Conference on Computer Vision and Pattern Recognition, CVPRW'06*, pp. 55-55.

[44] Zhou X and Bhanu B (2006). Feature fusion of face and gait for human recognition at a distance in video, *18th International IEEE Conference on Pattern Recognition, ICPR 2006*, Vol. 4, pp. 529-532.

[45] Viola P and Jones MJ (2004). Robust real-time face detection, *International journal of computer vision*, *57*(2), 137-154.

[46] Liu Z and Sarkar S (2007). Outdoor recognition at a distance by fusing gait and face, *Image and Vision Computing*, *25*(6), 817-832.

[47] Geng X *et al.* (2008). Adaptive fusion of gait and face for human identification in video, *IEEE Workshop on Applications of Computer Vision, WACV 2008*, pp. 1-6.

[48] Jayaraj V and Ebenezer D (2010). A new switching-based median filtering scheme and algorithm for removal of high-density salt and pepper noise in images, *EURASIP journal on advances in signal processing*, *2010*(1), 690218.

[49] Hossain E and Chetty G (2011). Multimodal face-gait fusion for biometric person authentication, *9th International IEEE Conference on Embedded and Ubiquitous Computing (EUC), 2011 IFIP*, pp. 332-337.

[50] Lei P (2004). Adaptive median filtering, In *Seminar Report, Machine Vision,* Vol. 140.

[51] Hofmann M (2012). Combined face and gait recognition using alpha matte preprocessing, *5th IAPR International Conference on Biometrics (ICB),* IEEE, 2012, pp. 390-395.

[52] Behara A and Raghunadh MV (2013). Comparison of gait-face fused human recognition techniques, *International Journal of Electrical, Electronics and Data Communication,* Vol.1, pp. 2320-2084.

[53] Rasmussen CE (2004). Gaussian processes in machine learning, In *Advanced lectures on machine learning*, Springer, Berlin, Heidelberg, pp. 63-71.

[54] Lathika BA and Devaraj D (2014). Artificial neural network based multimodal biometrics recognition system, *2014 International IEEE Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT),* pp. 973-978.

[55] Haghighat M, Abdel-Mottaleb M and Alhalabi W (2016). Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition, *IEEE Transactions on Information Forensics and Security*, *11*(9), 1984-1996.

[56] Xing X, Wang K and Lv Z (2015). Fusion of gait and facial features using coupled projections for people identification at a distance, *IEEE Signal Processing Letters*, *22*(12), 2349-2353.

[57] Derbel A, Vivet D and Emile B (2015). Access control based on gait analysis and face recognition, *Electronics Letters*, *51*(10), 751-752.

[58] Fan TY, Mu ZC and Yang RY (2017). Multi-modality recognition of human face and ear based on deep learning, *2017 International IEEE Conference on Wavelet Analysis and Pattern Recognition (ICWAPR),* pp. 38-42.

[59] Liu J, Fang C and Wu C (2016). A fusion face recognition approach based on 7-layer deep learning neural network, *Journal of Electrical and Computer Engineering*, *2016*.

[60] Kurban OC, Yildirim T and Bilgiç A (2017). A multi-biometric recognition system based on deep features of face and gesture energy image, *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA),* pp. 361-364.

LIST OF PUBLICATIONS

1. **Sharma A *et al.*** (2018). Multimodal fusion of face and gait biometrics for person identification using deep convolution neural networks – *Communicated* (in Scopus).
2. **Sharma A *et al.*** (2018). A multi-biometric person identification system using face and gait fusion: A deep learning approach – *Communicated* (in SCI index).