Multi-Modal Biometric-based Human Identification Using Deep Convolutional Siamese Neural Network

Miss Hsu Mon Lei Aung

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Electrical Engineering
Department of Electrical Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2021

การระบุตัวบุคคลด้วยชีวมาตรหลายโหมดโดยใช้โครงข่ายประสาทสยามแบบคอนโวลูชันเชิงลึก

น.ส.ชู มอง เล อง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

Thesis Title	Multi-Modal Biometric-based Human Identification Using Deep Convolutional Siamese Neural Network
By	Miss Hsu Mon Lei Aung
Field of Study	Electrical Engineering
Thesis Advisor	Associate Professor CHARNCHAI PLUEMPITIWIRIYAWEJ, Ph.D.
Thesis Co Advisor	Professor Kazuhiko Hamamoto, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Doctor of Philosophy

..... Dean of the FACULTY OF ENGINEERING

(Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

DISSERTATION COMMITTEE

..... Chairman

(Assistant Professor Somkiat Wangsiripitak, Ph.D.)

..... Thesis Advisor

(Associate Professor CHARNCHAI PLUEMPITIWIRIYAWEJ, Ph.D.)

..... Thesis Co-Advisor

(Professor Kazuhiko Hamamoto, Ph.D.)

..... Examiner

(Associate Professor SUPAVADEE ARAMVITH, Ph.D.)

..... Examiner

(Associate Professor NISACHON TANGSANGIUMVISAI, Ph.D.)

..... Examiner

(Assistant Professor SUREE PUMRIN, Ph.D.)

ชู มอง เล อง :
การระบุตัวบุคคลด้วยชีวมาตรหลายโหมดโดยใช้โครงข่ายประสาทสยามแบบคอนโวลูชันเชิงลึก. (Multi-Modal Biometric-based Human Identification Using Deep Convolutional Siamese Neural Network) อ.ที่ปรึกษาหลัก : ชาญชัย ปลื้มปิติวิริยะเวช, อ.ที่ปรึกษาร่วม : คาซึฮิโค ฮามาโตโม

การรู้จำแบบไบโอเมตริกซ์เป็นงานที่สำคัญในระบบรักษาความปลอดภัย แม้ว่าไบโอเมตริกซ์ใบหน้าจะได้รับการยอมรับอย่างมากและเป็นประโยชน์มากสำหรับการรู้จำมนุษย์ แต่อาจถูกขโมยและเลียนแบบได้ง่าย นอกจากนี้ยังมีความท้าทายในการรับข้อมูลใบหน้าที่เชื่อถือได้จากกล้องความละเอียดต่ำ ในทางตรงกันข้าม เมื่อเร็ว ๆ นี้มีการใช้ไบโอเมตริกซ์ท่าเดินเพื่อการรู้จำ ซึ่งเชื่อถือได้มากกว่าจากภาพในกล้องความละเอียดต่ำ อย่างไรก็ตาม การรู้จำมนุษย์ยังคงเป็นปัญหาเนื่องจากขาดรายละเอียดทั้งตัวในระยะทางสั้น ๆ นอกจากนี้ ระบบไบโอเมตริกซ์แบบเดียวยังคงมีข้อจำกัดกับปัจจัยภายในของแต่ละคุณลักษณะ เมื่อเร็ว ๆ นี้ โครงข่ายประสาทเทียมคอนโวลูชันเชิงลึก (deep Convolutional Neural Network หรือ deepCNN)ได้ถูกนำมาใช้อย่างแพร่หลายในหลายสาขาเพื่อรู้จำอย่างไรก็ตาม ข้อมูลที่ใช้ฝึกฝนจำนวนมากจำเป็นต้องติดฉลากกำกับก่อน การได้มาซึ่งข้อมูลไบโอเมตริกซ์และการติดฉลากสำหรับการสร้างชุดข้อมูลขนาดใหญ่ยังคงเป็นปัญหาอยู่ ในวิทยานิพนธ์นี้ เราขอเสนอแนวทางการรวมไบโอเมตริกซ์สองชุดเข้าด้วยกันโดยใช้โครงข่ายประสาทเทียมเชิงลึกกับโครงข่ายประสาทเทียมสยาม(Siamese Neural Network)สำหรับการเรียนรู้การรู้จำมนุษย์ โมเดลเครือข่ายที่เสนอจะเรียนรู้คุณลักษณะทั้งในปริภูมิภาพและปริภูมิเวลา ที่เลือกคุณลักษณะจากท่าทางการเดินและคุณลักษณะเด่นจากภาพใบหน้า คุณลักษณะไบโอเมตริกซ์ทั้งสองที่สกัดได้ถูกรวมเข้ากันเป็นเป็นกระบวนการในระดับเซนเซอร์สำหรับการรู้จำหลายรูปแบบ การศึกษานี้ทำการทดลองกับชุดข้อมูลท่าทางการเดินของฐานข้อมูล CASIA-B ที่เปิดเผยต่อสาธารณะ ชุดข้อมูลใบหน้าของ Yale-B และชุดข้อมูลวิดีโอท่าทางการเดินของผู้ใช้ 25 ราย โมเดลที่เสนอมีความแม่นยำในการจำแนกประเภท 97.3 % ด้วยคะแนน 0.97 F1 และ อัตราความผิดพลาด(ERR) 0.004 โมเดล SNN ที่เสนอยังมี True Positive Rate (TPR) อยู่ที่ 90.4% สำหรับท่าทางการเดินและ 89.7% TPR สำหรับใบหน้าและ 98.4% TPR ใเมื่อใช้ทั้งสองรูปแบบผลการทดลองแสดงให้เห็นว่าระบบสามารถจำแนกบุคคลโดยคุณลักษณะที่เรียนรู้เกี่ยวกับภาพใบหน้าGait energy (GE) และLow-resolution (LR) ได้โดยตรงการประเมินประสิทธิภาพการรู้จำหลายรูปแบบที่เสนอมานั้นยังเข้ากันได้เมื่อเปรียบเทียบกับวิธีการรู้จำหลายรูปแบบอื่นๆ

สาขาวิชา	วิศวกรรมไฟฟ้า	ลายมือชื่อนิสิต .....
ปีการศึกษา	2564	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6171408921 : MAJOR ELECTRICAL ENGINEERING

KEYWORD: Deep Convolutional Neural Network, Multimodal Biometrics, Transfer Learning, Siamese Neural Network

Hsu Mon Lei Aung : Multi-Modal Biometric-based Human Identification Using Deep Convolutional Siamese Neural Network. Advisor: Assoc. Prof. CHARNCHAI PLUEMPITIWIRIYAWEJ, Ph.D. Co-advisor: Prof. Kazuhiko Hamamoto, Ph.D.

Biometric recognition is a critical task in security control systems. Although face biometric has long been granted the most accepted and practical biometric for human recognition, it can be easily stolen and imitated. It also has challenges getting reliable facial information from the low-resolution camera. In contrast, a gait physical biometric has been recently used for recognition. It can be more complicated to replicate and can also be taken from reliable information from the poor-quality camera. However, human body recognition has remained a problem since the lack of full-body detail within a short distance. Moreover, the unimodal biometric system still has constraints with the intrinsic factors of each trait. Recently, a deep Convolutional Neural Network (deepCNN) has been firmly applied to many fields in recognition research. Nevertheless, it needs a lot of labelled data for training. Biometrics data acquisition and labelling for creating large-scale datasets are still problematic. In this thesis, we propose a multimodal approach by combining two biometrics using a deep Convolutional Neural Network with a distance learning based Siamese Neural Network for human recognition. The proposed network model learns discriminative spatio-temporal features from gait and facial features. The extracted features from the two biometrics are fused into a common feature space at the feature level and sensor level methods for multimodal recognition. This study conducted experiments on the publicly available CASIA-B gait dataset, Yale-B faces dataset and a walking videos dataset of 25 users. The proposed model achieves a 97.3 % classification accuracy with an 0.97 F1 score and a 0.004 Equal Error Rate (EER). The proposed SNN model also achieves a 90.4% True Positive Rate (TPR) on gait and 89.7 % TPR on face modality, and 98.4% TPR on the multimodal system. The experimental results demonstrate that the system can classify people by learned features on Gait energy (GE) and Low-resolution (LR) face images directly. The proposed multimodal recognition performance evaluation is compatible in comparison to other multimodal recognition methods.

Field of Study:	Electrical Engineering	Student's Signature .....
Academic Year:	2021	Advisor's Signature .....
		Co-advisor's Signature .....

ACKNOWLEDGEMENTS

Throughout the writing of this doctoral thesis, I have received help and support indeed from the kind people around me.

First of all, I would like to express my deepest gratitude and appreciation to my advisor, Associate Professor Dr Charnchai Pluempitiwiriyawej, whose expertise was invaluable in formulating the research topic and methodologies. I also wish to thank my co-advisor, Professor Dr Kazuhiko Hamamoto, Tokai University. Their patience, motivation, guidance, encouragement, knowledgeable advice, timeless suggestions and polishing of the writing to improve the English expression have helped me to a very extent to accomplish my doctoral study.

Next, my sincere gratitude goes to all Professors of my thesis committee members for their valuable and supportive comments and the time they shared with me. I am also thankful to all Electrical Engineering (EE) Department professors for sharing and teaching practical background knowledge for me to understand the problematic theories and concepts very quickly.

I would like to acknowledge the AUN/SEED-Net (JICA) Scholarship, Doctoral Degree Sandwich Program and the top-up scholarship from the EE Department, Chulalongkorn University, for their financial support during my study period. My thanks are extended to all the staff of the EE Department, ISE and AUN/SEED-Net, for their support.

In addition, I thank all my friends who supported, understood and encouraged me throughout my study period.

Last but not least, I especially want to thank my parents and my sisters for their love, caring, sacrifice and support.

Hsu Mon Lei Aung

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Motivation and Research Problem

Person identification is critical in biometric authentication, security control, and video surveillance systems. Although face, iris, and fingerprints have been used as the primary physical biometrics to recognize people recently [1-4], it is challenging to use them reliably in a typical surveillance unconstrained environment at a considerable distance from the camera. Moreover, low resolution (LR) face recognition currently focus on using the high resolution (HR) face images by generating the corresponding LR image and mapping function synthetically [5]. Gait trait, the unique walking style of an individual, is a biometric that is less sensitive to distance and quality of capturing devices and cannot also be faked effortlessly. However, it also has limitations due to clothing, carrying bags, or environmental circumstances[6]. Due to the various encounter difficulties of unimodal biometric obtaining feature patterns, the recognition performance decreased.

On the other hand, a multi-modal system supports more reliable information extraction than unimodal [7-9] for recognition accuracy improvement. Different fusion-level methods are employed for the information fusion of various modalities. Since the feature set retains more valuable information about the input pattern than the score-level [10], integrating biometrics information at the early-stage fusion improved recognition performance.

Even though machine learning (ML) methods have been widely used for biometrics features extraction from raw data and apply classifiers for recognition, they have limitations on feature discrimination and selection automatically in various task domains. Deep learning (DL), a new subcategory of machine learning, has been developed using Artificial Neural Networks (ANNs) with multiple hidden layers for extracting features from low-level to abstract-level Layers by Layer. DL techniques with parallel and distributed data computing, adaptive feature learning, reliable fault tolerance, and hardy robustness characteristics[11], especially deepCNNs have recently

been utilized in biometrics recognition systems. The extensive training time, the massive amount of data and expensive and powerful GPUs for processing requirements are the significant problems in deep CNN models. However, Transfer learning (TL) could solve these problems by reusing the trained model for new tasks. It is also a machine learning method that can take learned features on one task and leverage them on a new task. TL performed well in transferring knowledge from the domains to target tasks with little data. It also spectacularly reduces cost and training time, resulting in improved performance on related problems.

Moreover, if the new class with a small amount of data is needed to set in the biometric system for recognition, the CNN's models must retrain the whole models. It is one of the most critical problems for CNNs based methods. Siamese Neural Network (SNN), also anointed a twin network, has an identical network structure of two subnetworks that share all parameters, weights, and biases. SNN [12] is based on similarity metric learning, and the network model's input requires only the image pairs. The distance metric learning-based differencing layer connects the two subnetworks. SNN is also called a one-shot classification model that can accurately make predictions with a single training sample of a new class.

According to the significant performance of deep CNN techniques in different recognition tasks, this study intends to investigate the deep transfer learning-based CNN approach and Siamese Neural Network (SNN) to recognize a person at a distance in video using two biometric traits, face, and gait, with a small amount of data.

1.2 Objectives

The primary purpose of this dissertation is to develop strategies as follows:

- To implement a deep CNN model with a small amount of training samples for biometric recognition
- To compare and analyze the proposed model and other state-of-art models
- To propose a transfer learning based effective feature extraction and classification model for multi-modal recognition

- To develop a transfer learning-based similarity metric learning SNN model for multi-modal recognition

1.3 Scope of Work

- Video sequences that include a person with normal walking conditions
- The entire body of a person with low-resolution that has at least one complete gait cycle

1.4 Research Contribution

The main contributions of this research are:

- The study proposed a small deep CNN network model that performs multi-modal recognition with a small amount of training sample, not more than 100 images for each subject.
- This research also conducted a transfer learning-based effective feature extraction, and classification model for feature-level fusion based multi-modal recognition.
- Transfer learning-based similarity metric learning Siamese Neural Network model is also developed for multi-modal verification and recognition.
- The proposed SNN method applied the combination of easy-triplets and Semi-hard triplets mining for triplets input selection.

1.5 Dissertation Organization

This dissertation is arranged into six chapters. Chapter 1 describes the objectives of this study, the scope of work, the research contributions, and the dissertation organization. Chapter 2 presents the literature review on deepCNNs, TL and SNN. In Chapter 3, some general backgrounds are provided. Chapter 4 consists of the proposed methodology and models. The experimental setup, training and testing data sequences are explained in Chapter 5. This chapter also describes the experimental result, analysis, and discussion of some significant findings. This dissertation ends with

Chapter 6, summarizing the research findings and outlining the research's contribution and discussing the research limitations and future research work.

CHAPTER 2

LITERATURE REVIEW

While the unimodal system has proven reliable in many studies, they have limitations such as noisy data sensing, absence of distinguished data representation, and non-universality properties of the modalities [13]. As a result of single modality-based recognition problems, multi-modal recognition methods that combine data of multiple modalities have been developed. Several studies have proposed multi-modal bio-metrics recognition systems with various combinations of different biometrics. This section discusses studies that applied different machine learning approaches in multi-modal biometrics systems.

Zhou et al. [14] presented a multi-modal feature-level fusion-based recognition at a long distance in a surveillance environment. The system applied Principal Components Analysis and Multiple Discriminant Analysis (MDA) methods to decrease the extracted features' dimensions and select the most critical face and gait traits. Also, in [15], a study used Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) methods for feature extraction and selection. The extracted face and gait features were fused with hierarchical and holistic fusion methods. Zhang et al. [16] proposed a canonical correlation analysis method for ethnicity identification based on a feature-level fusion of two modalities. The Local Binary Patterns (LBP) operator extracted frontal face features, and gait features were characterized by Spatio-temporal expression of lateral gait. In [17], the projection of heterogeneous features of gait and face modalities into a unified space is fused by the projecting features for human recognition. Also, in [18], complicated face features from the wavelet discrete transformation method and soft face characteristics associated with the skin colour of front view images and GE image features were fused with the score-level by the Choquet integral method Particle Swarm Optimization (PSO) technique. According to the learning approaches mentioned above for gait and face multi-modal biometric systems, they need to manually choose essential and specialized feature extraction

and selection in the given images. When the number of classes to classify increases, the feature extraction process becomes more complex and challenging.

In deep learning-based biometrics systems, the deep network automatically extracts the features from the input modalities. These deep learning abilities have solved the limitations of other machine learning algorithms. Therefore, deep learning-based biometrics recognition techniques have been developed recently.

Ding et al. [19] proposed CNNs based deep learning framework of multi-modal face information-based recognition. A set of CNN is implemented for extracting features of multiple face modalities. Stacked Auto-Encoder is operated for feature fusion. In [20], the extracted elements of multi-stream CNNs with face, iris, and fingerprint modalities were fused by multi-level features abstraction for identification. [21] also represented Deep Belief Network (DBN) based facial features extraction, recognition process, and CNN-based left and right irises feature extraction and classification processes. The resulting scores are combined by multi-modal biometrics identification for score and rank-level fusion methods. Also, in [22], face, fingerprint, and iris traits recognition systems using deep learning template matching techniques were designed. Contourlet transform and local derivative ternary methods were utilized for feature extraction. The weighted rank-level algorithm fused the extracted features. Kim W et al. [23] presented a deep CNN of finger-vein and finger shape modalities. The near-infrared camera sensor captured the finger images. The matching distance scores of features were fused by weight sum, product, and perceptron methods. Boucherit et al. [24] developed a merge CNN scheme based on identical CNNs with different input images for finger vein recognition. The merged CNN was built by the optimal CNN structure of fusion of two images with various qualities.

Although conventional CNNs based biometrics recognition has significant performance, it requires a massive amount of training data, computation resources. Moreover, if the new class with a small amount of data is needed to set in the biometric system for recognition, the CNN's models must retrain the whole models. It is one of the most critical problems for CNNs based methods. The deep CNN models achieved low performance when insufficient training data problems occurred.

TL is often used to handle inadequate biometrics data issues by transferring pre-trained knowledge from source to related target domains. In [25], iris and periocular modalities-based deep transfer learning methods were proposed for recognition. VGG model was used for feature extraction, and feature selection applied Binary Particle Swarm Algorithm. Fusion of two modalities used matching score and feature-level fusion methods. Lumini et al. [26] presented the various deep learned models with transfer learning for the automated plankton recognition system. Tao Z et al. [27] proposed bidirectional feature extraction and transfer learning-based deep CNN for finger-vein recognition. The original image features and the rotated image features were fused by feature-level concatenation. Therar et al. [28] used a CNN with a transfer learning approach for multi-modal right iris and left iris biometrics recognition. The feature extraction and classification were performed using the proposed deep learning and multi-class SVM algorithms. In [29], finger-vein and finger knuckle print features were extracted using three popular CNNs interact with transfer learning. The proposed fusion approaches combined these features for classification. Zhu et al. [30] developed a CNN-based deep transfer learning for the human identification framework. The framework learned and transferred optimal motion dynamics representation knowledge from the source domain tasks to the target domain.

Based on the automated discriminative feature extraction ability of deep CNNs methods, we proposed a deep transfer learning strategy based CNN model to perform multi-modal recognition with a limited amount of data that is not more than 100 samples per subject in this study. This research developed a pre-trained model as feature extractors and classification model for face, gait, and multi-modal recognition using the early-stage fusion method.

Schroff et al. [34] presented a FaceNet that used a deep CNN. It's learning based on Euclidean L2 distances embedded learning per image for the embedding space directly related to face similarity. In [35], deep neural networks automatically learn the robust and discriminative gait features. The distance metric learning-based Siamese network with a contrastive loss layer is operated to compute the similarity metrics for gait pairs. The paper [36] presented a model based on deepSNNs with a supervised loss function for facial expression recognition (FER). The transfer learning

from the verification process to an identification process is used as a joint identification-verification representation. Song et al. [37] proposed a pairwise differential SNN with a mask generator for finding similarities between occluded facial blocks and corrupted feature elements.

The learned generator established a binary mask dictionary and combined the feature discarding mask and original face features by multiplication for robust occlusion face recognition. Yu et al. [38] also proposed a deep SNN for kinship verification that included three parts: feature extraction from two input face images, extracted feature fusions based on element-wise operations, and decided to verify using a similarity score. DD et al. [39] presented a signature and EEG based multi-modal SNN (mSNN) for person verification. EEG signal features are extracted from an EEG encoder, and signature features are acquired from the image encoder. Then the two encoded features are combined to calculate similarity for verification.

Table 1 Survey of Multi-modal Biometrics Recognition Systems

Methods	Biometrics	Methods	Datasets
Zhou et al. [14]	face(side) + gait	MDA , PCA	Database of 46 people
Hossain et al. [15]	face(side) + gait	PCA,LDA, Holistic or Hierarchical Fusion	Publicly available video database of human actions
Zhang et al. [16]	face + gait	LBP, CCA	own database including 36 walking people
Xing et al.[17]	face +gait	Coupled Projection	ORL, CASIA
Ghalleb AE et al. [18]	face(hard,soft)+ gait	Score level fusion, Choquet integral	CASIA Gait A and B
Soleymani et al.[20]	face + iris + fingerprint	feature fusion with multi stream CNNs	BIOMDATA, BioCop, CASIA-Iris Thousand, Notre Dame-IRIS
Al-Waisy et al.[21]	face + two side irises	Deep Belief Network (DBN), CNN	FERET, CASIA V1.0 , MMU1, SDUMLA-HMT
Derbel et al.[31]	face +gait	PCA, NN, two descriptors	database of 27 people
Kurban et al. [32]	face+Sgesture	VGG , PCA	EURECOM, Kinect Face, BodyLogin Gesture
JH Koo et al.[33]	face + body	VGG Face-16, ResNet-50	DFB-DB1, ChokePoint

CHAPTER 3

BACKGROUND THEORIES

A Biometric system is a method that recognizes or identifies human beings by their biometrics. Two kinds of biometrics, physical and behavioural biometrics, are used for recognition.

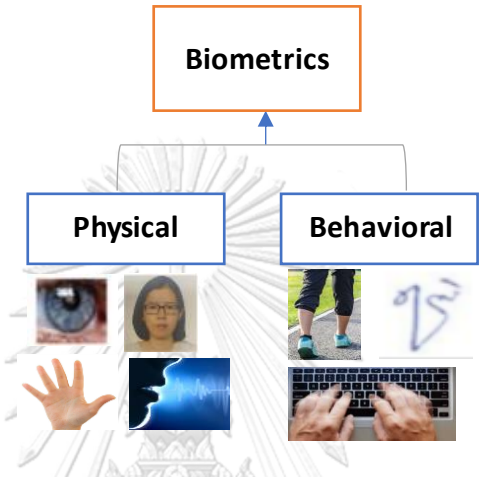


Figure 1 Two types of biometrics

3.1 Face Detection

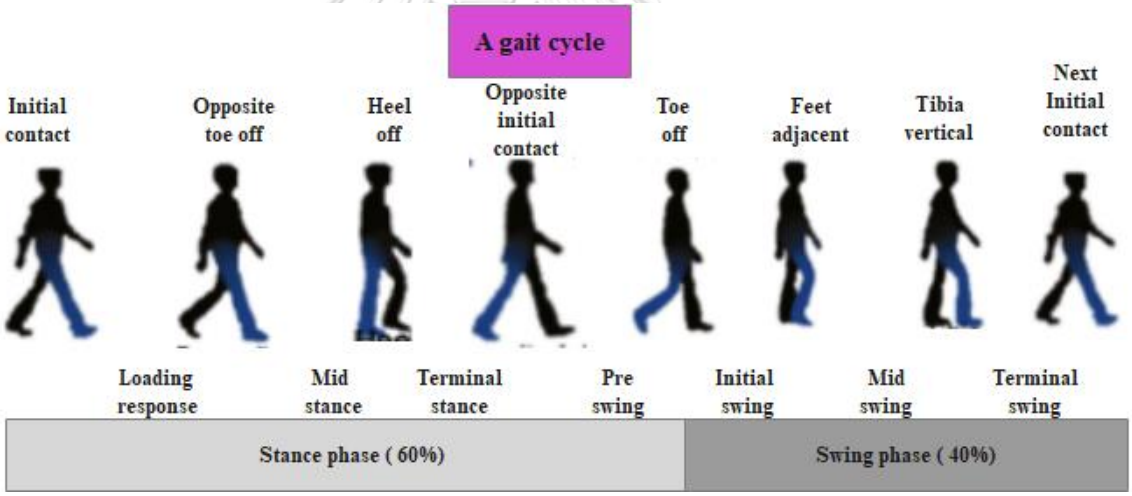
Face detection is the technique to detect the face of a person before face recognition. The method is concerned with discovering faces in video frames or images and recording the face region if found. This procedure is fundamental in surveillance and security control systems. There are many challenges in face detection techniques. The illumination condition, camera characteristics, and partial or large occlusion on the face can distress the face appearance, leading to the issue of face detection and recognition. In this research, the earliest first step of preprocessing for LR face images extraction is face detection in surveillance videos.

3.2 Gait and Gait Energy (GE) Image

Gait is a walking style of people, and everybody has their own unique walking patterns. The gait is the only biometric to cooperate long distances, providing crucial alternative information. It means gait features can take quite a considerable range

without interacting with the subject and using low-quality image or video capturing devices. The main reason for using gait biometrics for recognition at a distance is that gait is not sensitive to distance and video or image resolution. Although early gait recognition methods directly used silhouettes images extracted from the raw input videos sequences, various environmental changing conditions have affected every silhouette image. Therefore, different gait representation images are used for gait feature extraction.

GE is an image that describes a human walking action sequence in a single image. GE image describes the spatiotemporal information of gait, which means the movement of the human being is described in an image while maintaining temporal information. The GE image representation is obtained from the averaging operation on a complete gait cycle of the human silhouette images into one image.



src=https://www.researchgate.net/profile/Cheng-Zhang-82/publication/322958273/figure/fig1/AS:962701790543912@1606537401838/A-complete-cycle-of-the-human-gait-Besides-humen-identification-the-gait-analysis-can.gif

Figure 2 A complete gait cycle of the human

A complete gait cycle is defined that the step starting the heel strike of one foot and continues until the preparation of the same foot's heel strike for the next step.

3.3 Deep Convolutional Neural Network

ML is a core type of artificial intelligence where computer machines can learn from historical data through related algorithms to make predictions on new data. Many ML methods have been used in biometrics systems for recognition intentions. The ML methods extract features from raw biometrics data by various feature extraction techniques and apply classifiers to examine the extracted features for recognition. Although feature extraction techniques have an essential role in ML, these methods do not always perform well with various biometrics or other domains. Moreover, they also need to select and extract critical features manually [40]. Furthermore, standard features extraction of traditional learning methods also has a limited ability for features discrimination. The main advantage of machine learning is to improve generalization ability, which means examining a model that can react and adapt new data and make correct predictions after getting trained on a training set.

Deep learning, a new subcategory of machine learning, has been developed using Artificial Neural Networks (ANNs) with multiple hidden layers for extracting features to Layer by Layer. Deep learning models can also automatically learn features from raw data and get more valuable features expression that overcomes the limitations and difficulties of feature selection and extraction of other machine learning methods.

A typical Convolutional Neural Network (CNN) composed of convolution layers performs convolution of the input with its kernels or weights. It permits the image input to be provided into the system directly and then passed through various convolution layers, like the preprocessing phase in the traditional deep learning techniques. This process is conducted in the network, defining how the convolution layers extract the features to classify the input adequately. A CNN is also called the end-to-end process because it carries input and produces output instantly, learning about the convenient features.

3.4 Types of Layers of a Typical Convolutional Neural Network

Three main layers are involved in a CNN: convolution Layer, pooling layer, and fully connected Layer.

3.4.1 Convolution layer

The convolution layer, the main building block, performs convolution operation on the input image data using filters to produce feature maps in CNN. Convolution is the linear mathematical operation between matrixes that creates an output matrix. The filter is a feature detector that detects edges, shapes, and other features in the image layer by layer, and the size is much less than the image size. The convolutional process is performed by sliding the filters over the input locally. Each filter has the same weight and bias through the input during this process. The number of feature maps generated by the convolution layer is the same as the number of filters. The general calculation of an output feature map is shown in Equation (1).

$$output = B + \sum (w * input) \qquad (3.1)$$

where B and w are bias and weight of the filter, $input$ is an m x n size image containing one or more channels.

The fundamental elements of a convolution layer are stride, which refers to the number of pixels the filter to skip and padding that retains the feature map dimension. The output feature map size relies on the input map size, filter size, padding and stride values.

3.4.2 Pooling Layer

The pooling layer also called a subsampling layer, decreases feature maps size to reduce network complexity and computational cost and hence control overfitting. It merges the adjacent pixels of a particular area of the feature map into a single value. Average pooling and max pooling are widely used methods in a typical CNN.

3.4.3 Fully Connected Layer

The fully connected layers are the last few layers in the CNN. The input to the first layer is the output from the final convolution or pooling layer, flattened and then fed into it. Neurons in this layer have complete connections to all activations in the previous layer.

3.4.4 Non-linear Activation

Activation and dropout are not considered as actual layers in a CNN. However, the activation function is the critical factor for deep CNN performance. Every deep neural network has an activation function followed by the convolution process to learn and recognize complex information from complicated data inputs such as videos, images, speech, etc. It also performs arbitrary complex mappings between inputs and outputs. Activation functions are needed for neural networks for introducing non-linearity into these networks. The differentiable feature of activation, an important feature, can provide a backpropagation optimization strategy of deep CNN. Although sigmoid and tanh functions were the most-popular non-linearity activations, the Rectified Linear Unit (ReLU) has recently been operated as the standard for deep learning.

3.5 Transfer Learning

The extensive training time, the massive amount of data and expensive and powerful GPUs for processing requirements are the significant problems in deep CNN models. The deep CNN models also achieved low performance when insufficient training data problems occurred. However, Transfer learning (TL) could solve these problems by reusing the trained model for new tasks. It is also a machine learning method that can take learned features on one task and leverage them on a new task. TL is performed well for the task with a limited amount of data for training a whole model from scratch. It also spectacularly reduces cost and training time, resulting in improved performance on related problems.

3.6 Machine Learning Algorithms for Classification Task

Supervised learning is the one method of ML to make the model enable to predict future outcomes after training based on past data. It generates a function that maps inputs to desired outputs. Classification is the essential method of pattern recognition and ML, and it is a supervised learning approach. The Supervised machine learning classification algorithms which categorize the data from the preliminary information include the following: Support Vector Machine (SVM), Logistic Regression, Decision Tree, K- nearest Neighbors, Gradient Boosting and Random Forest.

3.7 Siamese Neural Network

A Siamese neural network (SNN) is a category of CNN network architecture and generally consists of two identical CNNs as subnetworks. SNN is based on the similarity metric learning method, and the same configured subnetworks share all parameters, weights and biases. Individuals can learn the hidden representation of an input vector with feedforward and backpropagation strategies for comparing their outputs by a distance function. The parallel subnetwork instances are joined at the top of the networks by a distance-measure loss function that calculates embedding distances between embeddings from the previous layers. The contrastive loss is the most used function in SNN. It is defined as follows:

$$Loss(x_1, x_2, Y) = (1 - Y)\frac{1}{2}(D_w)^2 + \frac{1}{2}Y\, max(0,\, m - D_w)^2 \qquad (3.2)$$

Where x_1, x_2 are two input samples, Y is a binary function for indication, m is margin and D_w in equation 3.3 is defined as the Euclidean distance between the embeddings of the subnetworks.

$$D_w = \sqrt{(f_w x_1 - f_w x_2)^2} \qquad (3.3)$$

where f_w is a particular embedding output of the subnetworks. The output of the Siamese neural network is the semantic similarity between two input vectors. Unlike traditional approaches that allocate binary similarity labels to pairs, SNN desires to

fetch the output embedding vectors closer to input pairs labelled as similar and dismiss the feature vectors if the input pairs are dissimilar.

One inspiring task is a classification under the restriction that we may only observe a single example of each possible class before predicting a test instance. This task is called one-shot learning or one-shot classification. Siamese network is a technique that can handle one-shot learning.

CHAPTER 4

METHODOLOGY

In this section, the general architecture of the proposed deep CNN model based multi-modal biometrics recognition system is explained in Section 4.1. The detailed descriptions of the proposed approach are followed in the subsequent Sections 4.2 - 4.5.

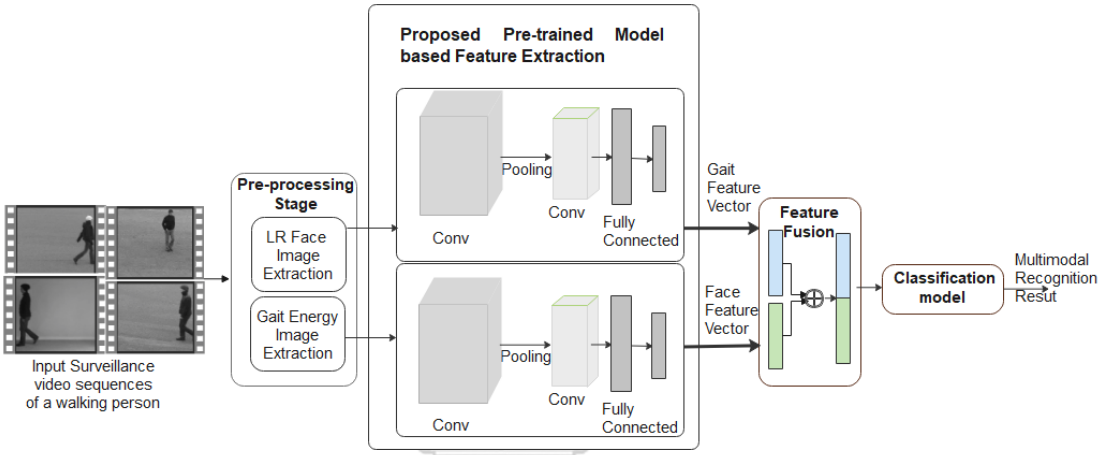


Figure 3 The general architecture of the proposed deepCNN model bases multi-modal biometrics system

4.1. Proposed System Overall Procedure

A typical multi-modal recognition framework mainly comprises biometrics data acquisition and pre-processing, feature extraction, extracted feature fusion, and classification parts. The proposed multi-modal system is shown in Figure 3 generally. In the early first step, the sequences of walking people frames were acquired from the surveillance videos, and then human silhouette images were extracted by background subtraction. A person is walking in various directions in every video sequence. The adequately aligned single sequence of the silhouette of gait is averaged into a GE image. At the same time, the proposed system detects the face region in the video frame and extracts it as an LR face input image. Next, face and gait features are acquired by the proposed deep CNN model and concatenated these features directly at the

feature level fusion. A unified feature vector is passed through the classification model for multi-modal recognition. The step-by-step flowchart of the proposed system is shown in Figure 4.

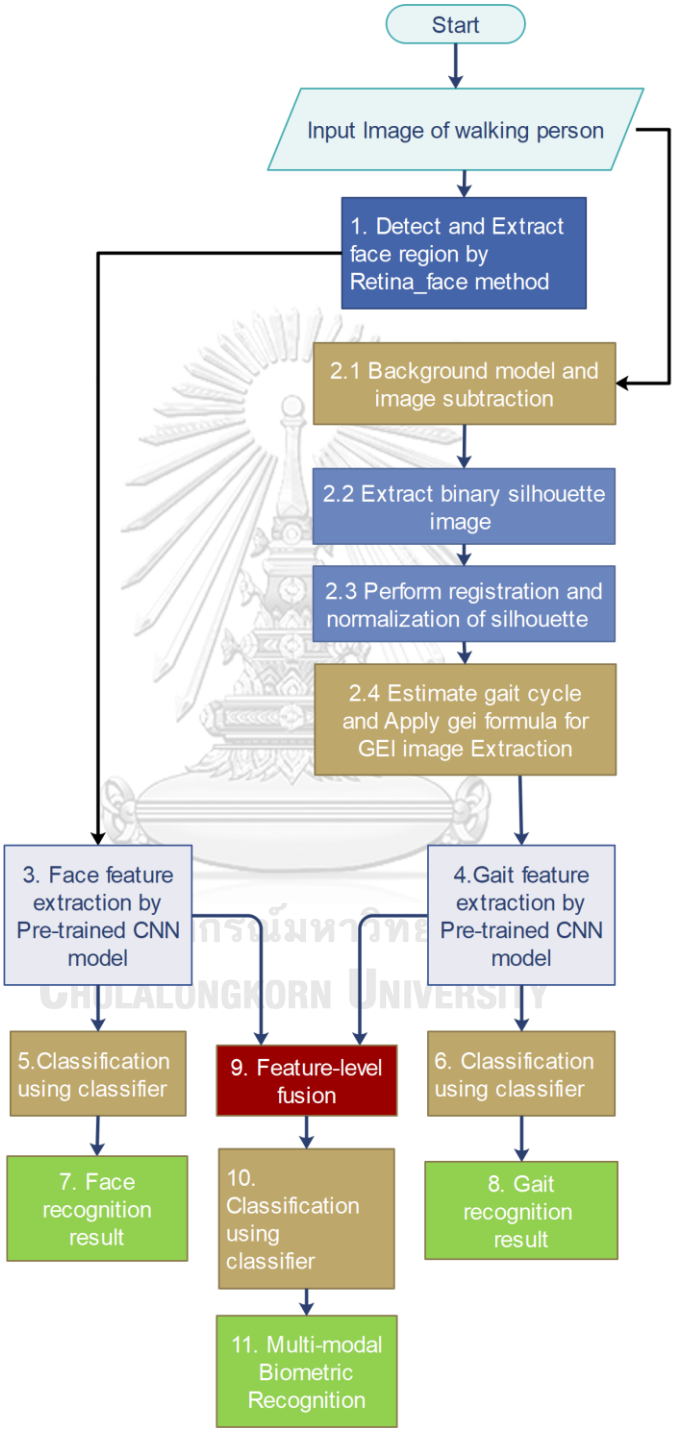


Figure 4 Flowchart of the proposed framework

4.2. Detection and Extraction of LR Face and Gait Energy Image as Pre-processing

The first preprocessing step is the face detection of a person for acquiring an LR face image from the input frames. A deep learning-based cutting-edge facial detector, RetinaFace [41], is used to detect the face. It is a robust stage pixel-wise facial localization method that puts face locations and scales on feature pyramids. The mobilenet0.25 with 1.7 M model size is used as the pre-trained model to detect face. It predicted face score, face region by bounding box, and five facial landmarks (right eye, left eye, nose tip, right mouth, and left mouth) on the face. The method performs 87percent face detection accuracy on the input videos approximately. Based on the face detection result, the system obtains the LR face region images from the frames.

The proposed system uses a GE image to represent a human walking gesture sequence in a single image for gait feature extraction and recognition. The GE image averaging action can retain the original gait sequences data and less sensitivity to a silhouette image noise[6]. In the GE image extraction, the first step is the foreground moving object detection. The simplified Self-organized Background Subtraction method (simplified SOBS) [42] is used for sensing foreground objects from the background model that uses a self-organizing neuronal map by mapping every color pixel into a weight vector. It uses the median filter with consecutive frames to construct the initial background model. The Euclidean distance equation in [43] calculates the minimum distance between the input and the current background model by the image's HSV hexagonal color space (h, s, v) to find the best match, as shown in Equation 4.1.

$$d\big(b, I(x,y)\big) = \sqrt{(v_b s_b cos(h_b) - (v_I s_I cos\,(h_I))^2 + (v_b s_b sin(h_b) - (v_I s_I sin\,(h_I))^2 + (v_b - v_I)^2} \quad (4.1)$$

Where $(h_b s_b, v_b)$ is the HSV color space of each neuron and (h_I, s_I, v_I) is for each pixel. The required distance value must not be larger than thresholds that were set automatically employing Otsu's method. The found best match was defined as a background pixel, and other pixels were defined as the foreground object component.

After getting the human walking binary silhouette image sequences, the proposed system performs the horizontally center-aligned and rectangle-based size-normalized silhouette images. The grey-level GE image is an efficient Spatio-temporal gait expression technique for human walking properties in a complete gait cycle for individual recognition by gait. Each silhouette image of a walking person describes the space normalized energy image, and GE image is the average cycle of the silhouette images into one image as the time normalized accumulated energy image.

A human gait cycle is formulated of two phases: the stance and the swing. The step starts the heel strike of one foot and continues until the preparation of the same foot heel strike for the next step is defined as a complete gait cycle. A GE image is defined from gait silhouette images sequences as follows:

$$G(x,y) = \frac{1}{N}\sum_{t=1}^{N} B_t(x,y) \qquad (4.2)$$

where N means the number of frames in a silhouette gait sequence of a cycle, t means the frame number in the image sequence, B is binary silhouette image and x and y are 2D image coordinate values. The proposed system considered an N value range from 20 to 25 frames per cycle. The Extraction result of LR face region image and GE image are shown in Figure 5.

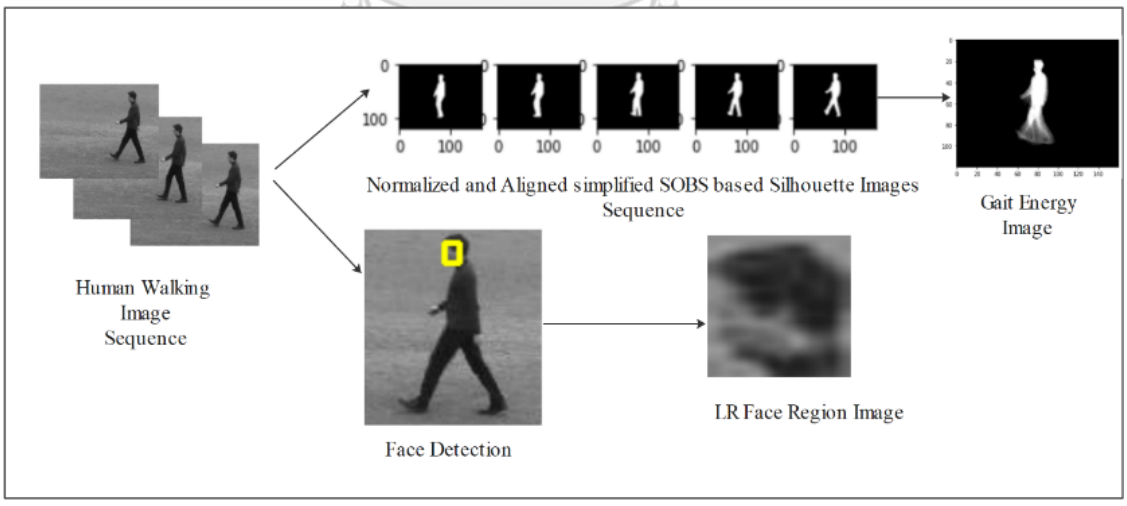


Figure 5 Example of GE image and LR face image extraction

4.3. Proposed Deep Convolutional Neural Network Architecture

This study uses LR face and GE image instead of HR face and raw gait sequence as the deep convolutional neural network input. The details of the proposed CNN model architecture is described in Table 1. The model composed of nine learned layers and uses a fixed input image size of 224x224 pixels with a channel. The first layer is the convolution layer, the main building block of the network, with 32 filters. The input volume is convolved with these filters of dimension 3x3 to produce an output feature map. The feature maps size for each layer is calculated as follow:

$$FeatureMap_{out} = \frac{W - F + 2P}{S} + 1 \qquad (4.3)$$

where W is the input volume size, F means filter size; padding borders size is P, and S is the stride size. The filter size 3x3, pooling size 2x2 and stride value 2 are used in our proposed model. The padding with value 1 is used in the first two pooling layers. The first layer is followed by dropout with 0.5 rates and Rectified Linear Unit (ReLU) transfer function as activation prevented slower learning speed and overfitting in the learning process. Trainable parameters of each layer are the number of well learnable parameters of the layer affected by the backpropagation process. The formula in Equation 4.4 is used to calculate the learnable parameters in each layer. These parameters situate only in the convolution layer and fully connected layers of the network. Every pooling layer of the proposed model performs the maximum subsampling operation to decrease the feature map dimensionality and retain the most critical information.

$$Trainable\ Parameters = \big((filter\_size * conv\_Depth) + 1\big) * \#filters \qquad (4.4)$$

The output from the previous convolution layers is flattened before being connected to the last FC layers. These layers contain the weight and bias together with the neurons for connecting these neurons between two different layers. The flattened vector then goes through an FC layer and a dropout layer of 0.5 that dropped

out 50 percent of the nodes from the network randomly for preventing the cause of overfitting issues. The last layer of the proposed model is also an FC layer and output layer of the proposed network that represents the output classification result of the classes of the related datasets.

Table 2 Description of Proposed base Network Model

Layer Number	Type of Layer	Number of Filters	Size of Features Map	Trainable Parameters
0	Image Input	-	224x224x1	
1	Conv2d-1	32	224x224x32	320
2	dropout	P= 0.5		
3	MaxPool2d_1	1	112x112x32	
4	Conv2d-2	32	110x110x32	9,248
5	MaxPool2d_2	1	55x55x32	
6	Conv2d-3	64	53x53x64	18,496
7	Conv2d-4	64	51x51x64	36,928
8	MaxPool2d-3	1	25x25x64	
9	Conv2d-5	128	25x25x128	73,856
10	Conv2d-6	128	23x23x128	147,584
11	MaxPool2d-4	1	11x11x128	
12	Conv2d-7	256	9x9x256	295,168
13	Conv2d-8	256	7x7x256	590,080
14	MaxPool2d-5	1	3x3x256	
15	Flatten	-	2304x1	
16	Fc1	-	1024x1	2,360,320
17	dropout	P= 0.5		
18	Output Layer	-	#Classes	

4.4. Proposed CNN Model-based Feature Extraction, A Classification Model Architecture and Multi-modal Features Fusion

The proposed model performs recognition on reasonably large datasets. We also consider using it for other tasks where it has a small amount of data. The deep learning model requires a large amount of data to get better performance than other techniques and needs expensive GPUs and takes a long training time that increases the computational cost. The transfer learning (TL) technique can overcome the limitations of deep learning methods by reusing a trained model on a specific task as part of the training process for a different task. In this step, the proposed deep CNN model was performed as a feature extractor by transferring knowledge from the base model to the classification model, as shown in Figure 6. There are two independent feature extractors one is for gait, and the other is for the face that does not share their weight values and had trained on CASIA B gait and Yale B face datasets separately. The proposed method freezes the pre-trained layers of a base model to preserve existing learning generic features in feature extraction. Then, the learned features are used as input for a new, smaller classification model. The model comprised two FC layers to learn additional information on a new dataset and perform a classification process for unimodal recognition.

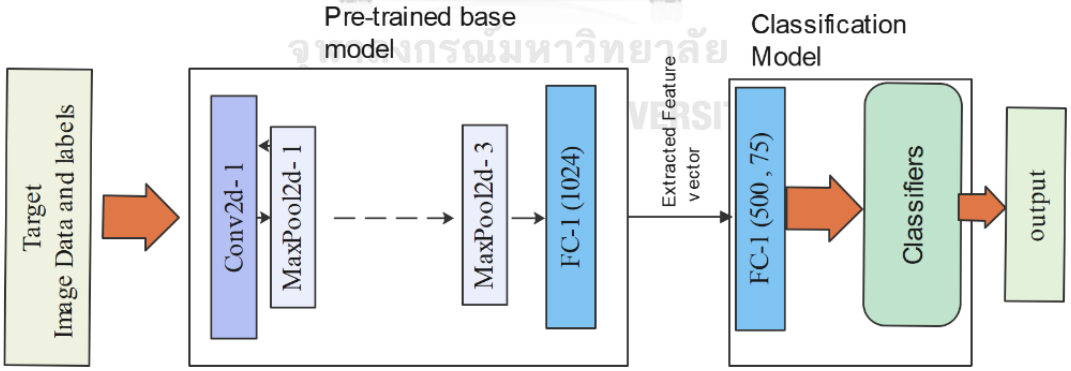


Figure 6 Feature extraction from the pre-trained model-based feature extractors and classification model

For the multi-modal biometrics recognition, the proposed system fused the extracted features from the two feature extractors. The method conducted a feature fusion method on all probable combinations of face features and gait features to

generate the maximum number of the concatenated feature vectors. The integrated feature vector is passed through the classifier layer for the multi-modal recognition process.

4.5. Siamese Neural Network Based Multi-modal Recognition

SNN or a twin network, has an identical network structure of two subnetworks that share all parameters, weights, and biases. SNN is based on similarity metric learning, and the network model's input requires only the image pairs. The distance metric learning-based differencing layer connects the two subnetworks. SNN is also called a one-shot classification model that can accurately make predictions with a single training sample of a new class.

This study intends to investigate the Siamese Neural Network (SNN) to identify a person from two biometric traits, face, and gait. First, a gait and face biometrics-based SNN is proposed to combine two sensor-level modalities for multi-modal recognition. After that, the proposed method applied Semi-hard triplets mining for an anchor, positive and negative images to the training inputs pairs. In the training stage, the three images (anchor, positive and negative) as a triplet image input separately enter the three parallel CNNs, passes them to extract 75- Dimensional embeddings, and then compute their loss function and optimize a ranking loss to fine-tune the model. For testing, a test image is sent into only one CNN and then perform the feedforward network computation to extract the features in the testing stage. Finally, the proposed model presents the similarity score as an output for person recognition.

4.5.1. Proposed Methodology

The proposed system takes the combination of two biometrics images as a single sensor-level image as early-level fusion, and the model takes in that image and generates a dimensional embedding for it. The input image pairs are passed through these subnetworks with the same parameters and weights. The output embeddings calculate the difference based on the distance learning between inputs. And then, the proposed SNN classifies the output using a similarity score for the inputs are the same person or not.

4.5.1.1. Pre-processing

Each person's computed GE images with 224✕224 size and same size face images are fused vertically and resized to the same size with a single channel using bicubic interpolation over square pixels. The sample of face, GE and merged images can be found in Figure 4.5. Both input images are pre-processed before passing them into the proposed subnetworks.

Figure 7 (a) Face Image (b) GE Image

4.5.1.2. Siamese Neural Network (SNN) Model

Figure 8 gives a detailed architecture of the proposed Multi-modal Siamese Neural Network (multiSNN) based on a person identification framework. The network consists of three subnetwork instances with identical architecture, The network architecture is same with the previous proposed model with two extra FC layers is used as the based SNN to extract the embeddings from inputs. Nonlinear activation function ReLU is chosen in the hidden layers of the proposed model for faster learning. The input size of 224✕224✕1 image pixel is used for these networks. These parallel networks share the same weights and joint in the top by a triplet loss function those measures embedding distances between three embeddings. The primary purpose of triplet ranking loss is to predict the relative distance between the model inputs as the dissimilarity between the anchor and positive images must be less than between the anchor and negative images pair. The input triplets images are selected by applying easy triplets and semi-hard triplets mining methods. The triplet loss is defined in equation 4.5.

$$Loss = D(A,P) + \alpha < D(A,N) \qquad (4.5)$$

where α is a margin that defines how distantly away the dissimilarities should be. D is the distance value, A means an anchor image, P is positive, and N is negative image. The margin value 0.2 is used for the proposed system.

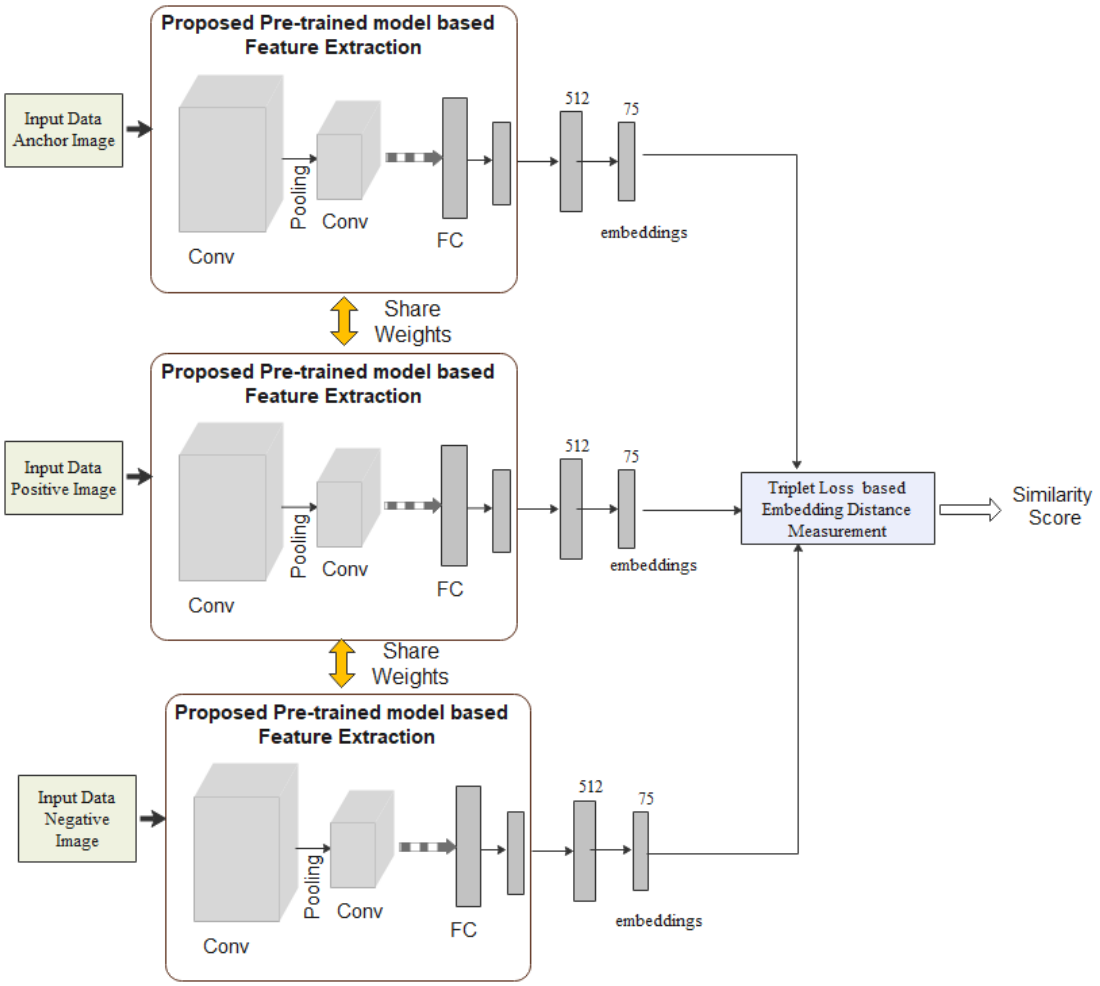


Figure 8 Details Description of the Proposed Multi-modal Siamese Neural Network (multiSNN):

In the training stage, the three images (anchor, positive and negative) as a triplet image input separately enter the three parallel CNNs, passes them to extract 75-Dimensional embeddings, and then compute their loss function and optimize a ranking loss to fine-tune the model. After the network training, the positive sample embeddings are closed to the anchor embedding values, and the negative samples values are far from the anchor.

To train the proposed SNN network, we combine two types of triplet mining for input triplets. The first is a simple triplets mining method by choosing an anchor image and randomly sampling positive and negative images. The easy triplets have a 0-loss learning because the distance between the anchor and positive plus margin is less than between anchor and negative samples. Semi-hard triplets mining is applied to the training inputs to avoid the no learning and long learning time. There is a positive loss in this mining, but the negative is not closer to the anchor than the positive distance. A batch of training data is fed and randomly generated triplets with all samples and calculated the loss. Batch size value 5 and the number of hard samples 50 are defined to train the network.

The test image is sent into only one CNN and then perform the feedforward network computation to extract the features in the testing stage. Finally, the proposed model presents the similarity score as an output for person recognition. The SNN model performance is evaluated by AUC and sensitivity metrics.

CHAPTER 5

EXPERIMENTAL SETUP AND RESULTS

This study was intended to design an effective multi-modal biometrics recognition system in surveillance environments where the object is far from the cameras.

5.1 Datasets

The proposed model is trained and tested on the publicly available CASIA-B gait dataset [44] for gait recognition and Yale-B faces dataset [45] for face recognition.

5.1.1 CASIA-B Gait Dataset

CASIA-B gait dataset contains human silhouettes and GE images of 124 people. GE images for each class are extracted from the silhouette sequences from the 11 angles that start from 0 to 180 degrees of the camera.

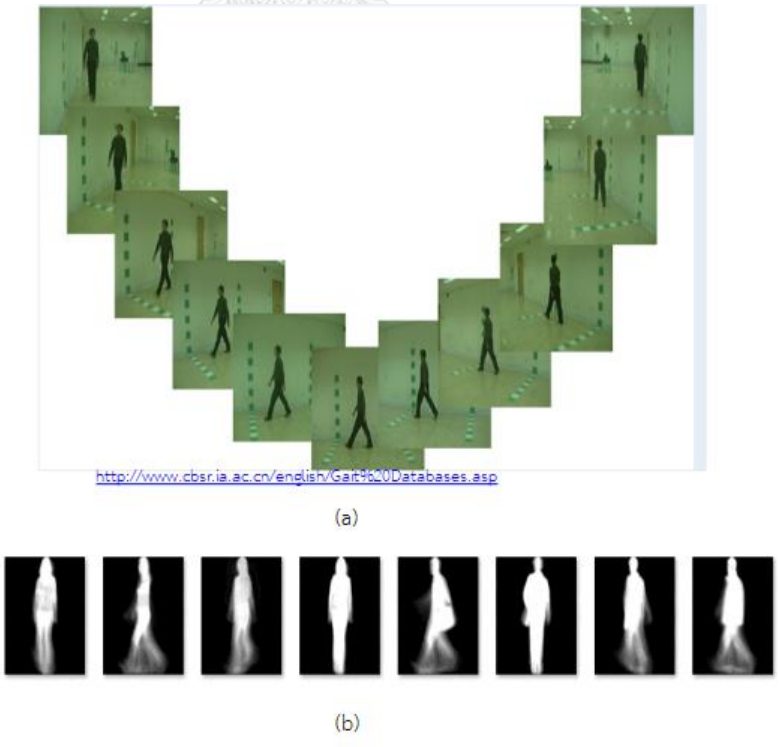


Figure 9 CASIA-B Dataset (a) raw video frames of 11 angles (b) GE images

The ten sequences of a person are divided into three walking situations: six for normal walking conditions, two for walking with a bag, and the last two for the subjects wearing their coats while walking.

5.1.2 Yale-B Face Dataset

In the face dataset, face images of 38 people individually were used as input for extracting face features. Each subject consists of 64 frontal face images with various illumination conditions.

Figure 10 Example of Face images of Yale-B

5.1.3 Walking of Human Video Dataset

The proposed system evaluated the walking category of the Recognition of Human Actions video dataset [46]. The walking video dataset comprises 25 people in four scenarios, as shown in Figure 5.3. The first one is (d1) outdoors situation, the second is (d2) outdoors with scale variation, the d3 is a person walking with different clothes, and the fourth one is a person walking in the (d4) indoors environment. There are 100 videos of 25 people in total.

(a) Outdoors (d1) (b) Outdoors with scale variation (d2) (c) Outdoors with different clothes (d3) (d) Indoors(d4)

Figure 11 Example of input videos of a person in four scenarios

We developed the proposed network model by TensorFlow framework with python. They ran on a GPU Tesla V100-SXM2-16GB Google Colab Cloud Platform with 50 GB RAM.

5.2 Evaluation Measure

This section describes the evaluation metrics we used to evaluate the proposed network model's performance, including unimodal recognition and the multi-modal recognition process. The evaluation metrics that we used in our study are shown as follows:

5.2.1 Accuracy

The performance analysis of classification on multi-class balanced datasets used recognition accuracy as the most relevant metric because class distribution did not need to be considered. The accuracy returned an overall measurement of the correct prediction of individual class among the total number of classes. The actual positive value [47] of the whole system in total is obtained through the equation:

$$TTP_{all} = \sum_{j=1}^{n} x_{jj} \tag{5.1}$$

where n is the number of classes in a system and j means the row or column size of a confusion matrix [47]. The overall accuracy of the model is computed using the following equation:

$$Accuracy = \frac{TTP_{all}}{Total\ of\ test\ samples} \tag{5.2}$$

5.2.2 Confusion Matrix

Confusion matrix evaluated the multi-class classification performance. The diagonal components indicate the counts in which the predicted label's value measures equal the actual in the confusion matrix. The higher the diagonal numbers, the better, revealing numerous correct predictions. Many evaluation metrics occurred on the confusion matrix that made the records of the number of counts of actual and

the predicted classification. Distinct from the confusion matrix used in the binary classification task, there are no positive or negative classes in the multi-class classification matrix.

We have shown the classification report of the predicted model as well. The report describes Precision, Recall and F1-score for each class. These calculations are computed based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) values for classes individually.

5.2.3 Precision

Precision describes the ratio of accurately predicted actual results to the total expected positive outcomes. The more increased the precision associated, the lower the false positive rate.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.3}$$

5.2.4 Recall

A recall means the ratio of definitely predicted true results to all observations in the actual class. It is also known as True Positive Rate (TPR) or Sensitivity in machine learning.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.4}$$

5.2.5 Sensitivity

Precision and Recall metrics can use to measure model performance for doing binary classification or multiclass classification although Sensitivity is specifically for binary classification.

5.2.6 F1-score

F1-score is the weighted norm of Precision and Recall utilized in all classification algorithms; the score value gives equal weight to Precision and Recall. F1-score is used when False Negative and False Positive are more critical than True Positive and True Negative for evaluation measure. When the model scores high, the Precision and Recall metrics are increased. An F1-score value becomes one that is considered perfect.

$$F1_{\text{score}} = 2 * \frac{(precision * recall)}{(precison + recall)} \tag{5.5}$$

5.2.7 Equal Error Rate

One of the evaluation metrics for measuring and comparing the biometric system's performance is Equal error rate (EER). It can be obtained by the error value calculation of the false acceptance rate ($\boldsymbol{FAR}$) and false rejection rate ($\boldsymbol{FRR}$) equal. The Equal Error Rate *(EER)* is the value at the point of intersection of these two functions. The *FAR*, a synonym of false positive rate *(FPR)*, and $\boldsymbol{FRR}$, known as false-negative rate *(FNR)*, are calculated [48] by the following equations:

$$FPR\ (FAR) = \frac{FP}{(FP + TN)} \tag{5.6}$$

$$FNR\ (FRR) = \frac{FN}{(TP + FN)} \tag{5.7}$$

5.2.8 AUC-ROC curve

AUC- Area Under the Curve *ROC*-Receiver Operating Characteristics curve is used for measuring the multi-class classification performance at various threshold values. ROC curve shows a probability by a plot with *TPR* on the y-axis as opposed to the *FPR* on the x-axis, and AUC describes the separability degree, which means the curve displays the distinguishable performance of the classification model among classes. The higher the AUC, the better the model predicts the categories correctly.

5.3 Results

In this section, we describe the experimental results of the proposed model that include unimodal recognition and multi-modal recognition process.

5.3.1 Experiment I

Firstly, we divided a publicly available CASIA-B gait dataset into training and testing parts to perform gait recognition. In the training dataset, we split it into five sub-datasets for a five-fold cross-validation operation. It is a resampling method to secure the model is good sufficiently and can endure data variations. The main objective of utilising resampling is to facilitate the proposed model to learn as much as possible. The training losses and accuracies of the folds are shown in Table 5.1. The average accuracy score for all folds is 97.19 percent, with 0.72 standard deviations, and the average loss value is 0.13.

Table 3 Accuracy and loss of five folds cross-validation on the CASIA-B training dataset

Fold	Loss	Accuracy
1	0.16	96.28 percent
2	0.15	96.61 percent
3	0.11	98.00 percent
4	0.11	98.04 percent
5	0.13	97.02 percent
Average	0.13	97.19 percent

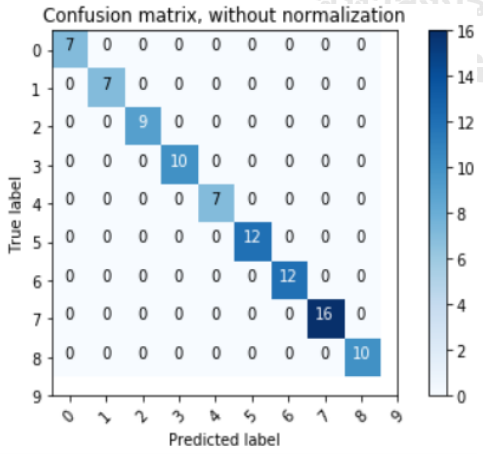
Next, this experiment followed the same methodology for face recognition on the publicly available face dataset Yale-B. It is also split into five for five-fold validations. The loss and accuracy of each fold are described in Table 5.2. The average

accuracy score for all folds is 92.71 percent, with 3.31 standard deviations, and the average loss value is 0.29.

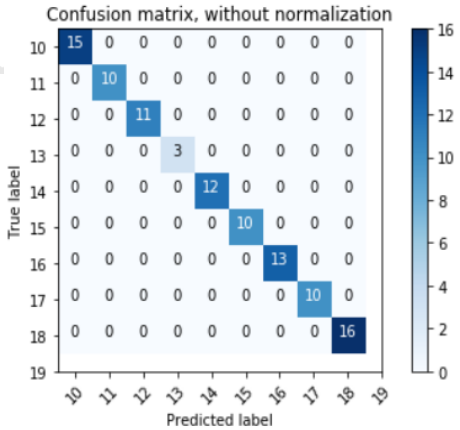
Table 4 Accuracy and loss of five folds cross-validation on the Yale-Face training dataset:

Fold	Loss	Accuracy
1	0.21	94.40 percent
2	0.21	96.15 percent
3	0.29	94.23 percent
4	0.23	92.30 percent
5	0.49	86.54 percent
Average	0.29	92.17 percent

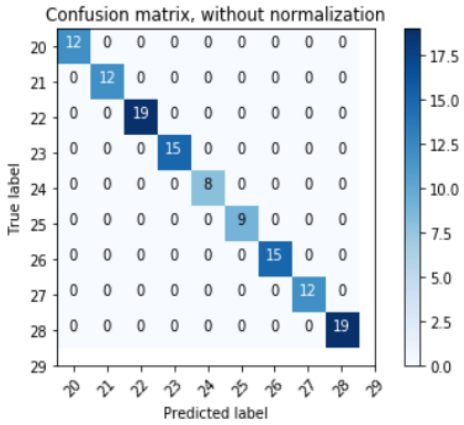
We evaluate the proposed model performance on the gait test data with 1,360 images. We described the performance of a multi-class classification on a set of test data by a confusion matrix. The matrix result of tests class 0 to class 124 is shown in Figure 12 (i) and 12 (ii).



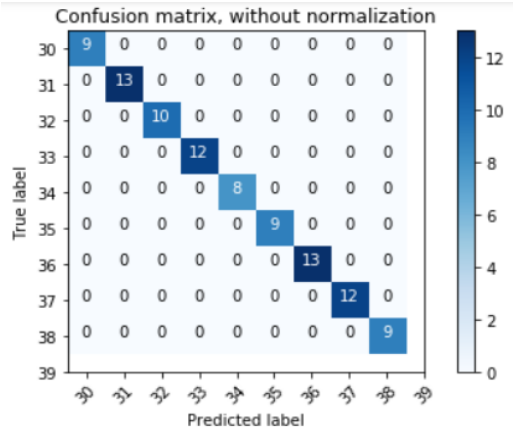
a. *class 0 to class 9*



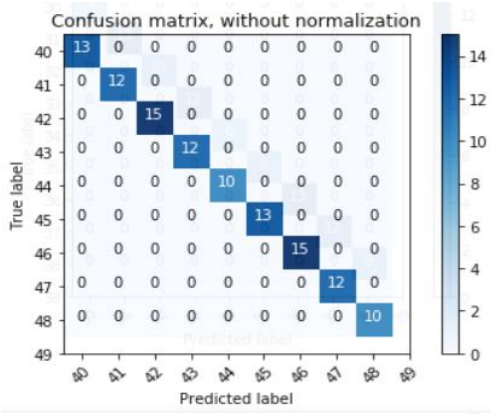
b. *class 10 to class 19*



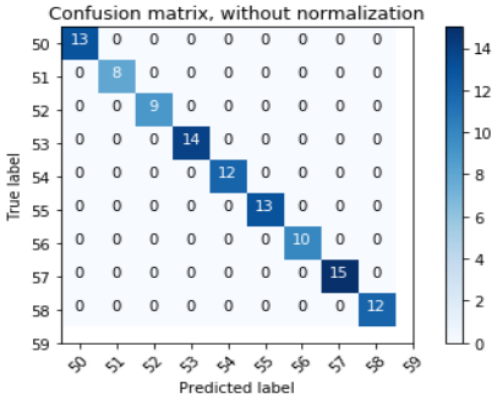
c. *class 20 to class 29*



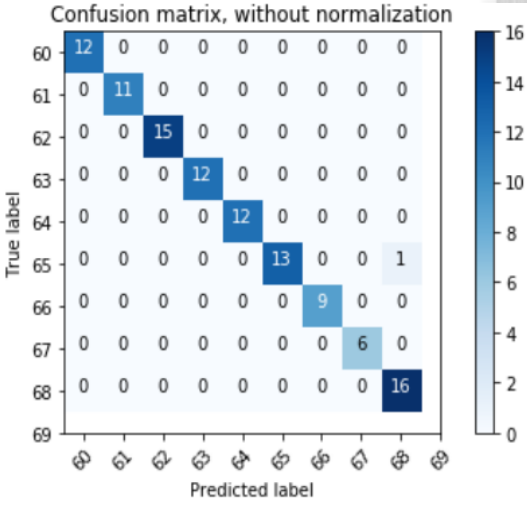
d. *class 30 to class 39*



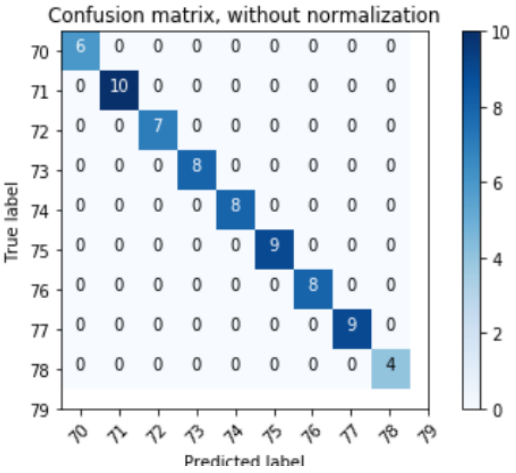
e. *class 40 to class 49*



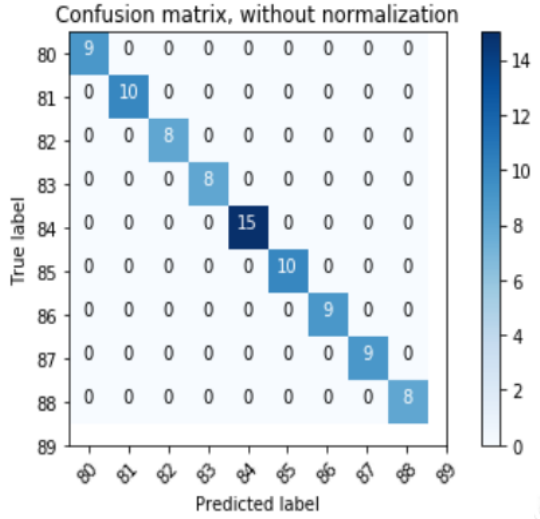
f. *class 50 to class 59*



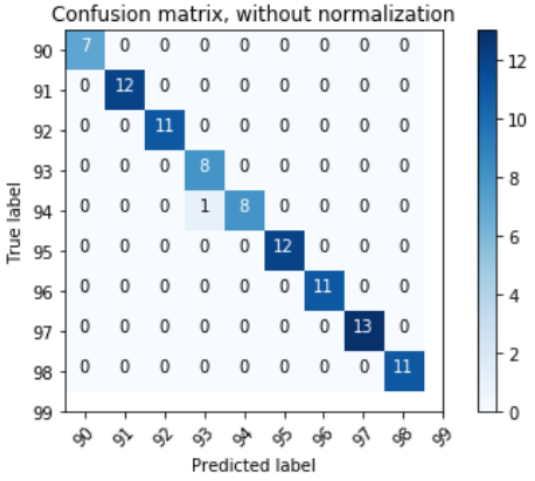
g. *class 60 to class 69*



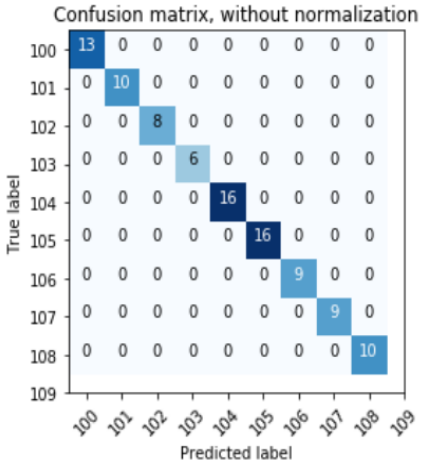
h. *class 70 to class 79*



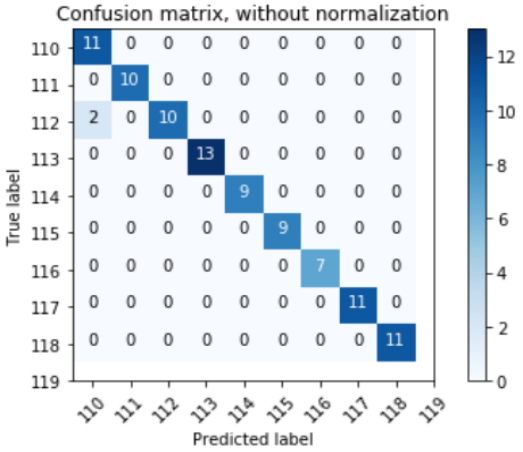
i. class 80 to class



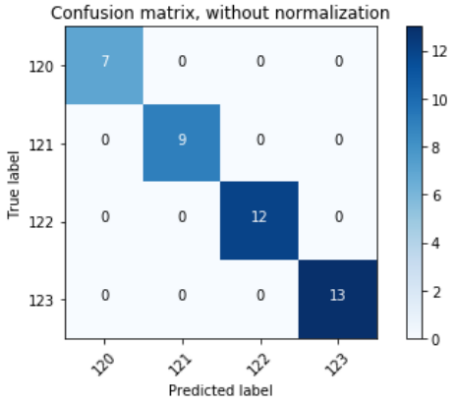
j. *class 90 to class 99*



k. *class 100 to class109*



l. *class 11 0 to class 119*



m. *class 120 to class 123*

Figure 12 Non-normalized confusion matrix of gait test classes 124

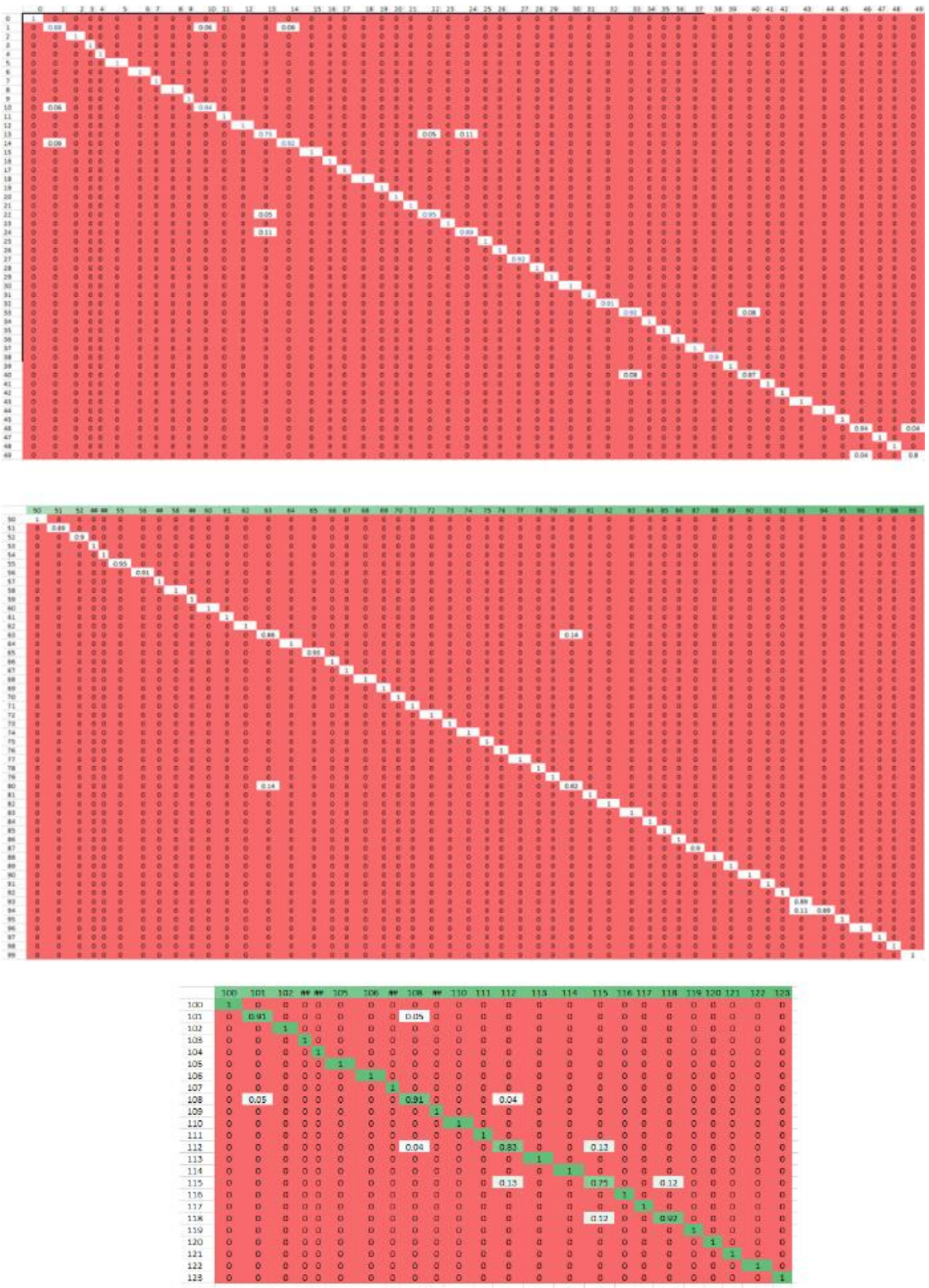


Figure 13 Normalized confusion matrix of gait test classes 124

The predicted model classification report on the 124 classes of the testing dataset is as follows.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	7
1	1.00	0.88	0.93	8
2	1.00	1.00	1.00	9
3	0.83	1.00	0.91	10
4	1.00	1.00	1.00	7
5	1.00	1.00	1.00	12
6	1.00	1.00	1.00	12
7	1.00	1.00	1.00	16
8	0.83	1.00	0.91	10
9	1.00	1.00	1.00	13
10	1.00	0.94	0.97	16
11	1.00	1.00	1.00	10
12	0.92	1.00	0.96	11
13	0.75	0.75	0.75	4
14	1.00	0.92	0.96	13
15	1.00	1.00	1.00	10
16	1.00	1.00	1.00	13
17	1.00	1.00	1.00	10
18	1.00	1.00	1.00	16
19	1.00	1.00	1.00	16
20	1.00	1.00	1.00	12
21	1.00	1.00	1.00	12
22	0.95	0.95	0.95	20
23	1.00	1.00	1.00	15
24	1.00	0.89	0.94	9
25	1.00	1.00	1.00	9

26	1.00	1.00	1.00	15
27	1.00	0.92	0.96	13
28	1.00	1.00	1.00	19
29	1.00	1.00	1.00	15
30	1.00	1.00	1.00	9
31	1.00	1.00	1.00	13
32	1.00	0.91	0.95	11
33	1.00	0.92	0.96	13
34	0.89	1.00	0.94	8
35	1.00	1.00	1.00	9
36	1.00	1.00	1.00	13
37	0.92	1.00	0.96	12
38	1.00	0.90	0.95	10
39	0.92	1.00	0.96	12
40	1.00	0.87	0.93	15
41	1.00	1.00	1.00	12
42	0.94	1.00	0.97	15
43	1.00	1.00	1.00	12
44	0.91	1.00	0.95	10
45	0.93	1.00	0.96	13
46	0.94	0.94	0.94	16
47	1.00	1.00	1.00	12
48	1.00	1.00	1.00	10
49	1.00	0.80	0.89	10
50	1.00	1.00	1.00	13
51	1.00	0.89	0.94	9
52	1.00	0.90	0.95	10
53	0.93	1.00	0.97	14
54	1.00	1.00	1.00	12
55	1.00	0.93	0.96	14
56	1.00	0.91	0.95	11

57	1.00	1.00	1.00	15
58	1.00	1.00	1.00	12
59	1.00	1.00	1.00	9
60	1.00	1.00	1.00	12
61	1.00	1.00	1.00	11
62	1.00	1.00	1.00	15
63	1.00	0.86	0.92	14
64	1.00	1.00	1.00	12
65	1.00	0.93	0.96	14
66	1.00	1.00	1.00	9
67	1.00	1.00	1.00	6
68	0.89	1.00	0.94	16
69	0.88	1.00	0.93	7
70	0.86	1.00	0.92	6
71	1.00	1.00	1.00	10
72	1.00	1.00	1.00	7
73	1.00	1.00	1.00	8
74	1.00	1.00	1.00	8
75	1.00	1.00	1.00	9
76	1.00	1.00	1.00	8
77	1.00	1.00	1.00	9
78	1.00	1.00	1.00	4
79	1.00	1.00	1.00	9
80	1.00	0.82	0.90	11
81	1.00	1.00	1.00	10
82	1.00	1.00	1.00	8
83	1.00	1.00	1.00	8
84	1.00	1.00	1.00	15
85	1.00	1.00	1.00	10
86	1.00	1.00	1.00	9
87	1.00	0.90	0.95	10

88	1.00	1.00	1.00	8
89	1.00	1.00	1.00	8
90	0.88	1.00	0.93	7
91	0.92	1.00	0.96	12
92	1.00	1.00	1.00	11
93	0.89	0.89	0.89	9
94	1.00	0.89	0.94	9
95	1.00	1.00	1.00	12
96	0.79	1.00	0.88	11
97	1.00	1.00	1.00	13
98	1.00	1.00	1.00	11
99	0.88	1.00	0.93	7
100	0.93	1.00	0.96	13
101	1.00	0.91	0.95	11
102	0.89	1.00	0.94	8
103	1.00	1.00	1.00	6
104	1.00	1.00	1.00	16
105	1.00	1.00	1.00	16
106	1.00	1.00	1.00	9
107	1.00	1.00	1.00	9
108	1.00	0.91	0.95	11
109	1.00	1.00	1.00	6
110	0.85	1.00	0.92	11
111	1.00	1.00	1.00	10
112	0.91	0.83	0.87	12
113	1.00	1.00	1.00	13
114	0.90	1.00	0.95	9
115	1.00	0.75	0.86	12
116	1.00	1.00	1.00	7
117	1.00	1.00	1.00	11
118	1.00	0.92	0.96	12

119	0.80	1.00	0.89	8
120	1.00	1.00	1.00	7
121	0.90	1.00	0.95	9
122	1.00	1.00	1.00	12
123	1.00	1.00	1.00	13
accuracy			0.97	1360
macro avg	0.97	0.97	0.97	1360
weighted avg	0.98	0.97	0.97	1360

The support column indicates the number of samples for each class. The macro average is used to average the unweighted mean of measures, and the weighted average is used to average the support-weighted mean of measurements.

We also evaluate the proposed deep learning CNN model on the face test data with 10 classes. The confusion matrix result of face classes is shown as below:

Figure 14 Non-normalized and normalized confusion matrix of face test classes

The classification report of the predicted model on face test classes as follows:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2
1	0.83	1.00	0.91	5
2	1.00	1.00	1.00	5
3	0.67	1.00	0.80	2
4	1.00	1.00	1.00	1
5	1.00	1.00	1.00	1
6	1.00	1.00	1.00	4
7	1.00	1.00	1.00	3
8	1.00	1.00	1.00	5
9	1.00	1.00	1.00	1
accuracy			0.93	29
macro avg	0.85	0.90	0.87	29
weighted avg	0.88	0.93	0.90	29

All training samples of previous experiments are not greater 13,000 images in all subjects and not exceed 100 image per subjects. According to experimental results, we conclude the proposed deep CNN architecture could recognize biometric with small amount of training data.

5.3.2 Experiment II

This experiment was intended to design an effective multi-modal biometrics recognition system in surveillance environments where the object is far from the cameras. In this experiment, the system evaluated the performance of unimodal recognition systems of LR face and GE images. After that, the system assessed the proposed multi-modal biometrics recognition system by combining features of two modalities using feature-level fusion.

The study used the LR face region images, as shown in Figure 14, corresponding to 25 subjects with 45 images per person from the dataset described in session 5.1.3. The human silhouette images are extracted by detecting foreground objects from 100 video sequences corresponding to 25 subjects. The proposed method extracted the four sequences of silhouette images for each person in a scenario. The range of 20 – 25 silhouette frames from the videos is covered one walk cycle, and is used for an average GE image, as shown in Figure 15, explained in Section 4.2 for each person.

Figure 15 Result Samples of LR face region images for one person from video sequences:

Figure 16 Result Samples of GE images for one person from video sequences:

5.3.2.1 Comparison of Architecture and parameters of CNN models

The architecture and essential parameters of the proposed network model and other state-of-the-art (SOTA) models are compared in Table 5. The proposed model had nine depth layers of 33.73 MB in size, and the input image is 224 pixels in height and width with a channel.

Table 5 Comparison of Architectures and Parameters of CNN models

Model	Depth Layers	Input Image Size (h,w,c)	Parameter Size (MB)
AlexNet [49]	8	227x227x3	219
DenseNet121[50]	121	224x224x3	26.92
InceptionV3[51]	48	299x299x3	96.58
VGG16[52]	16	224x224x3	513.73
Proposed	**9**	**224x224x1**	**33.73**

5.3.2.2 Comparison of Recognition Accuracy

First, the proposed network model evaluation is performed in unimodality face and gait recognition. Classification tasks with more than two class labels are called multi-class classification. The multi-class classification categorized the testing data into multiple classes in trained data as a model prediction.

For the performance measurement of the proposed model and SOTA models of the biometric system, the Yale-B face is used to train the proposed and Alexnet model from scratch and train the popular SOTA plain ImageNet-train network models as the first experiment. Table 6 compares the popular SOTA network models and the proposed model on the extended Yale-B dataset. It contained an average of 60 images per person of 38 people.

Table 6 Comparison of the proposed model and SOTA models

Models	Total Parameters	Trainable Parameters	Accuracy	Training Time (seconds)
Alexnet	59,345,446	59,342,694	73.6 percent	44
DenseNet121	19,892,582	12,855,078	95.6 percent	718
InceptionV3	55,367,238	33,564,454	95.7 percent	1513
VGG16	21,147,238	6,432,550	83.7 percent	101
Proposed	**3,570,950**	**3,570,950**	**95.8 percent**	**14**

The proposed model had the lower learned layers with the lower trainable parameters. The total size of the models is composed of input size, forward/backwards pass size and parameter size. The proposed model has not only less parameter size but also the whole model size is less than other SOTA models. The larger the trainable parameters lead the network to consume highly computing resources. The proposed model achieved the highest accuracy with less training time.

5.3.3 Experiment III

According to the previous experimental result, the proposed network can be defined as small network due to that has small number of learned layers with trainable parameters.

In real world biometric system, there is not possible to get large number of samples to train the whole CNN but also many training images for each class. In the following experiment, we used the proposed network as pre-trained network feature extractor by transferring knowledge from the base model to the classification model for multi-modal recognition. We extracted the face features from LR face images and gait features from GE images shown in Figure 14 and Figure 15.

Firstly, we performed unimodal recognition, we compared the recognition accuracies of popular multi-class classification algorithms on the extracted features of testing data by categorizing these data into multiple classes. Binary classification methods such as Logistic Regression and SVM (Support Vector Machine) with One-vs-Rest (OvR) and One-vs-One (OvO) strategies were also used as multi-class classifiers. OvR splits the dataset of more than two classes into multiple binary classification tasks, trains them separately, and makes predictions. OvO divides the multi-class dataset into one for each class versus every other.

Table 7 Gait Recognition Accuracy Comparison on Different Classifiers:

Classification Algorithms	Accuracy (unit: percent)
K-Nearest Neighbors	64
Decision Tree	42
Naïve Bayes	59
Random Forest	73
Gradient Boosting	60
Logistic Regression (OvR)	**89**
Logistic Regression (OvO)	88
SVM (OvR)	84
SVM (OvO)	84

The proposed model achieved higher accuracy with 89 percent in the OvR Logistic Regression classifier for gait recognition.

Table 8 Face Recognition Accuracy Comparison on Different Classifiers:

Classification Algorithms	**Accuracy (unit: percent)**
K-Nearest Neighbors	74
Decision Tree	49
Naïve Bayes	73
Random Forest	73
Gradient Boosting	69
Logistic Regression (OvR)	80
Logistic Regression (OvO)	**86**
SVM (OvR)	85
SVM (OvO)	**86**

Logistic Regression and SVM with OvO strategy gained 86percent more accuracy than other classifiers for face unimodal.

For the multi-modal biometrics recognition, the proposed system fused the extracted features of two biometrics face and gait from the same person. The fusion conducted all probable combinations of face features and gait features to generate the maximum number of the concatenated feature vectors. The integrated feature vector is passed through the classifier layer for the multi-modal recognition process. The recognition accuracies comparison of different classifiers is shown in Table 9.

Table 9 Multi-modal Recognition Accuracy Comparison on Different Classifiers

Classification Algorithms	**Accuracy (unit: percent)**
K-Nearest Neighbors	65
Decision Tree	51
Naïve Bayes	79
Random Forest	88
Gradient Boosting	75
Logistic Regression (OvR)	**97**
Logistic Regression (OvO)	96
SVM (OvR)	88
SVM (OvO)	94

The highest accuracy is 97 percent on the OvR Logistic Regression classifier.

Logistic Regression and SVM classifiers-based models get higher accuracy to result in both unimodal and multi-modal recognition. After these two classification

algorithms, Random Forest accurately classified the gait and face traits with 73 percent and 88 percent in multi-modal biometrics recognition. The Decision Tree classifier provided a less accurate result for all our recognition systems. Although the classification using Naive Bayes and Gradient Boosting methods worked better on the face than gait classification, K-Nearest Neighbors-based gait recognition is higher than face recognition. Multi-modal recognition using various classifiers except K-Nearest is more effective than the unimodal system.

As seen in Table 10, the proposed multi-modal recognition method is compared with other existing approaches. Our multi-modal system showed comparable accuracy as opposed to other multi-modal systems.

Table 10 Comparison of Multi-modal recognition systems

Papers	**Modalities**	**Method**	**Recognition Accuracy (percent)**
Zhou et al. [14]	face(side) + gait	MDA, PCA	87
Hossain et al.[15]	face(side) + gait	PCA, LDA, Holistic or Hierarchical Fusion	75
Zhang et al.[16]	face + gait	LBP, CCA	90
Xing et al.[17]	face + gait	Projection of Heterogeneous features, KNN	94
Our Proposed Method	**face + gait**	**Deep CNN with Transfer Learning**	**97.3**

F1-score comparison is also shown due to false negative and false positive are also important for the biometric system. The F1 score merges the Precision and Recall metrics into new metrics. The F1-scores for gait, face and multi-modal recognition are described in Table 11. The resulting score of Logistic Regression with OvR provided the highest F1-score value for multi-modal biometrics. The score of the first two classifiers in the face modality had a greater result than the other two modalities. However, other classifiers got a higher F1-score in multi-modal than unimodal.

Table 11 F1-score Comparison on Different Classifiers of Three Modalities

Classification Algorithms	**Gait**	**Face**	**Multi-modal**
	F1-score	**F1-score**	**F1-score**
K-Nearest Neighbours	0.61	0.72	0.63
Decision Tree	0.41	0.50	0.49
Naïve Bayes	0.58	0.72	0.79
Random Forest	0.71	0.72	0.87
Gradient Boosting	0.59	0.69	0.72
Logistic Regression (OvR)	0.85	0.79	**0.97**
Logistic Regression (OvO)	0.87	0.85	0.96
SVM (OvR)	0.84	0.84	0.87
SVM (OvO)	0.80	0.86	0.94

Multi-modal recognition with OVR Logistic Regression achieved highest accuracy and F1 –score value.

Equal Error Rate is the inverse of accuracy which means the descending the EER value, the more acceptable the accuracy of the biometric system. Table 12 describes the EER values of the proposed biometrics system based on the best classifier. The proposed multi-modal system EER was smaller than the unimodal biometrics system.

Table 12 : Equal error rates for the proposed biometric recognitions method

Biometrics	EER
Gait	0.032
Face	0.055
Multi-modal	0.004

According to the previous experimental result, the proposed pre-trained network CNN based feature extractor identify good feature representations and the multi-modal biometrics system got the better performance than unimodal.

5.3.4 Experiment IV

In this experiment, we intended to investigate the performance of a transfer learning-based similarity metric learning SNN model for multi-modal recognition. We have taken on the video dataset in Section 5.1.3. The face images and GE images of 25 subjects with normal walking conditions were collected. To train the proposed SNN network, combine two types of triplet mining easy and Semi-hard triplets mining with positive loss for input triplets' selection. The process generated triplets with all samples randomly. In the training stage, the three images (anchor, positive and negative) as a triplet image input separately enter the three parallel CNNs, passes them to extract 75- Dimensional embeddings, and then compute their loss function and optimize a ranking loss to fine-tune the model. In the testing part, test image is sent into only one CNN and then perform the feedforward network computation to extract the features in the testing stage. Finally, the proposed model presents the similarity score as an output for person recognition.

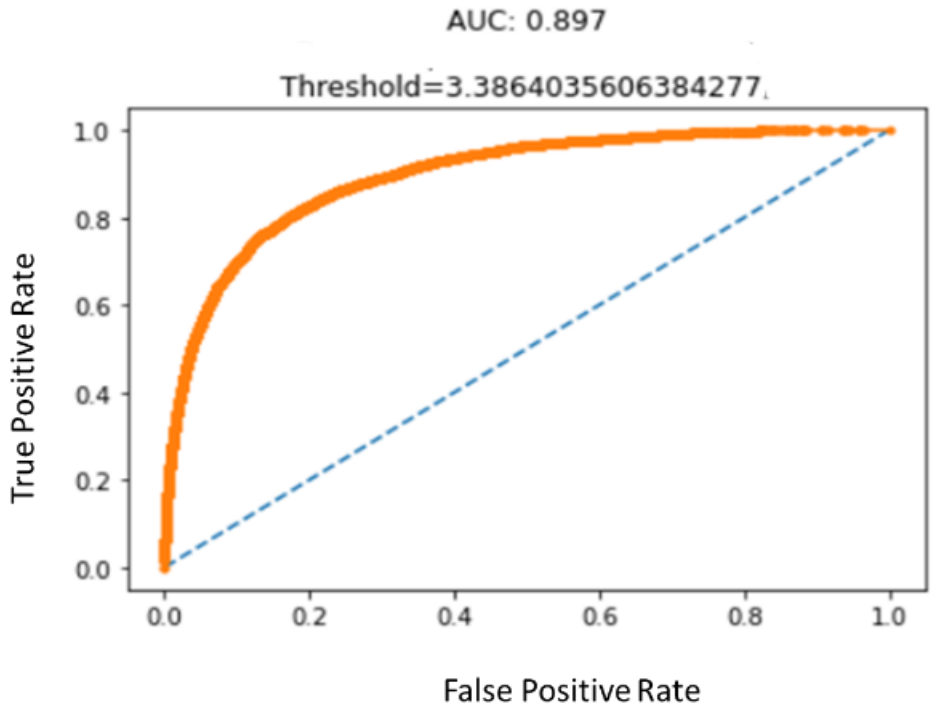
We developed the proposed SNN model utilizing by TensorFlow framework with python. The optimized different hyper-parameters such as learning rate, optimizer and the number of epochs selection conducted.

Table 13 Comparison of proposed SNN model training time on different modalities

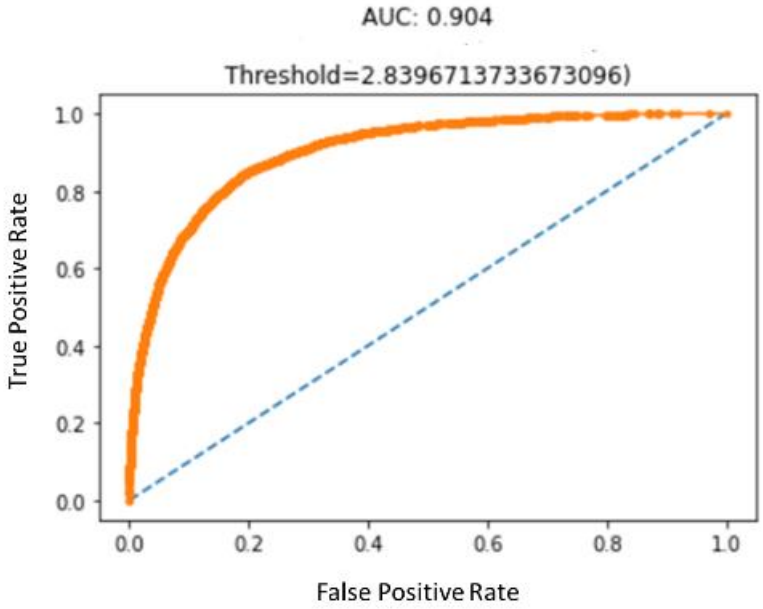
Modalities	Training Time (s)
Face	26199.46 seconds
Gait	20987.44 seconds
Face + Gait	20457.77 seconds

In this experiment, we selected an Adam optimizer with a learning rate of 6 e-04 and 600 epochs with a batch size of 50. Table 14 shows the training time for each modality.

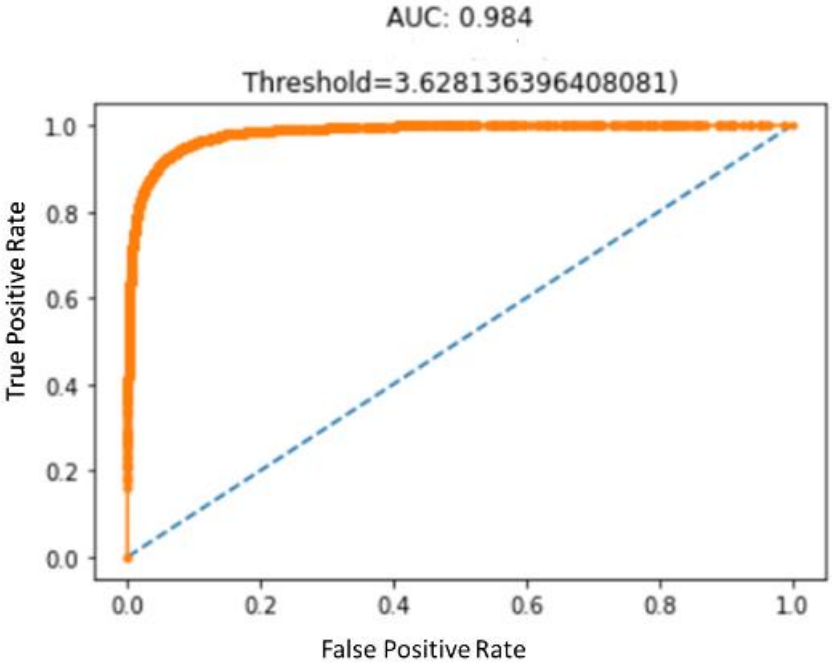
Siamese Neural Network based recognition performance is described by AUC-ROC curve. It measures the multi-class classification performance at various threshold values. ROC curve shows a probability by a plot with TPR on the y-axis as opposed to the FPR on the x-axis, and AUC describes the separability degree, which means the curve displays the distinguishable performance of the classification model among classes. The higher the AUC, the better the model predicts the categories correctly.



*(a) AUC-ROC curve of **face** biometric recognition*



*(b) AUC-ROC curve of **gait** biometric recognition*



(c) AUC-ROC curve of multi-modal biometrics recognition

Figure 17 AUC - ROC curves of SNN

The AUC-ROC curves shown in Figure 16 (a) (b) and (c) evaluated the unimodal and multi-modal system's performance across different threshold settings. In this experiment, a 3.62 threshold value was used for the proposed model, reaching the point where the TPR of the model got the highest value, 0.984 for multi-modal. For the unimodal system, 0.897 AUC on face biometric with a threshold value of 3.386 and 0.904 AUC on gait with 2.839 thresholds were achieved. According to the experiment, multi-modal biometrics recognition is the highest AUC which means the proposed SNN model can better predict people correctly on multi-modal biometrics than the unimodal biometric.

Due to low FN is more important than high TP in a biometric system, we also showed the Sensitivity result of our proposed model of biometrics. Sensitivity indicates the number of positive records correctly predicted.

Table 14 Comparison of proposed SNN model Sensitivity on different modalities

Modalities	**Sensitivity (percent)**
Face	69.8
Gait	69.7
Face + Gait	95.7

SNN model based Multi-modal achieved highest Sensitivity.

5.3.4.1 N-way One-shot Classification

Next, we performed an N-way one-shot classification task to evaluate the SNN model performance. In this classification task, the two inputs, a query image and a labelled support set are used to predict a query image label from the support set labels. N-way means N different classes in support set with one shot means one example per class.

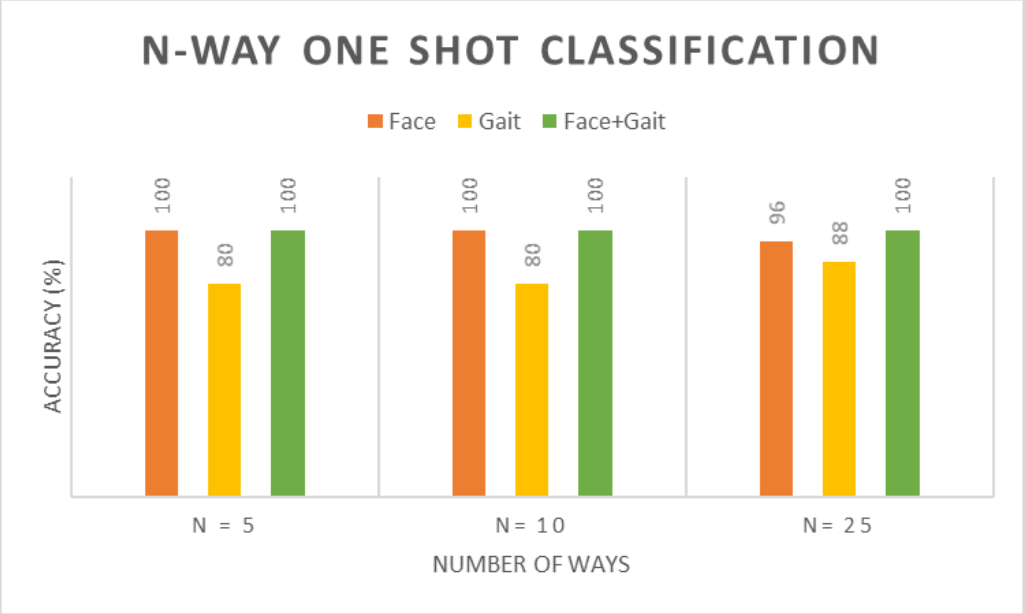


Figure 18 Comparison of N-way One-Shot Classification Accuracy (unit percent)

Although several experiments with varying values of N, the proposed model performed slightly better than with three different N values (5, 10, 25) for one-shot classification tasks. The detailed result of the N-way one-shot classification is described in Figure 17. the proposed model achieved slightly better accuracy in multi-modal than unimodal with three different N values.

5.3.5 Discussion

This research proposed a deep CNN model with small trainable parameters for biometrics recognition. The model trained with a small amount of data provided high recognition accuracy on two biometrics. And then, the proposed deep CNN model was performed as a feature extractor by extracting face, and gait features from LR face images and GE images acquired from low-quality surveillance videos as inputs to the classification model for biometrics recognition on the small amount of data. Although GE-based methods have been generally used in gait recognition, the derivation process of GE from raw gait sequences has many challenges. Due to the human body region segmentation and alignment steps being inaccurate and the segmented human body region noise, the extraction of gait movement information and consequent recognition can increasingly cause errors. Our proposed system used the simplified SOBS method based on moving object detection and silhouette image extraction. Then GE images are abstracted from the horizontally centre-aligned and size-normalized silhouette images. Most previous studies used HR face images for LR face recognition by synthetically generating the corresponding LR image and mapping function. This study used the RetinaFace detection technique to directly detect the LR face region as an input image to the recognition process. It decreases the processing time and complexity. The proposed multi-modal recognition system applied a feature fusion method to integrate learned features that provided more discriminative information about input biometrics. The deep CNN-based feature extractor performed dimensionality reduction to overcome the curse of dimensionality and effectively transferred learned knowledge from one task to other related tasks. Therefore, the proposed system got better performance on the small amount of data available for tasks. The research also performed a Siamese Neural Network-based multi-modal recognition system (multiSNN) by combining two different biometrics traits, face and gait. The multiSNN was developed to learn discriminative features from face and GE images using CNNs models. Sensor level fusion was performed in the early stage. Triplet loss-based distance measures are used to calculate the similarity among image pairs. The performance of face or gait just unimodal systems and the multi-modal system was examined using ROC-AUC curves, displaying that multiSNN exceeded gait recognition by increasing TPR from 89.7 percent

of the face and 90.4 percent of gait to 98.4percent of multi-modal. Although the facial features and gait features are almost identical effects in the proposed method, there has a problem when the face regions or human-bodies regions are entirely occlusions at a distance. As a result, combining two traits can improve the accuracy performance than unimodal when the unimodal trait is challenging to acquire. The proposed multiSNN achieved a 100 percent classification accuracy in 5, 10 and 25-way one-shot classification from the experiments conducted.

CHAPTER 6

CONCLUSIONS

In this study, a multi-modal biometrics recognition framework with a deep learning approach was proposed, along with the various classifiers, to achieve better person recognition accuracy. In the experiment, the extracted LR face and GE images from a publicly available videos database are utilized as input to the proposed CNN model trained on public datasets for learning and extracting practical features. And then, the extracted features were transferred to the classification model to classify a person. Seven kinds of classifiers were evaluated, namely, K-Nearest Neighbors, Decision Tree, Naïve Bayes, Random Forest, Gradient Boosting, Logistic Regression and SVM on three modalities. The experiment indicated that Logistic Regression was the most suitable classifier for human identification because it provided the average highest accuracy near 97 percent with multi-modal biometrics that is compatible in comparison to other multi-modal recognition methods. Furthermore, the experiment was conducted on the public dataset to compare the recognition performance and the architecture and parameters of the proposed model and other SOTA, namely VGG Net-16, Inception_v3, Densenet, and Alexnet. The proposed network has the smallest number of trainable parameters and achieves higher accuracy with lower training time than other SOTA. Moreover, In *EER* metric-based comparison, the proposed method has 0.055 *EER* in face modality, and the gait modality has 0.032 *EER* and 0.004 *EER* in a multi-modal system. To investigate the proposed Siamese Neural Network (SNN) performance that used similarity metric learning, we experimented on triplet input images of LR faces and GE images of 25 people by combining simple triplets mining and semi-hard triplets mining. The experimental results indicated that the proposed model gave the higher AUC 0.984 for multi-modal than the results of unimodal system, 0.897 AUC on face biometric and 0.904 AUC on gait. The proposed SNN model predicted people accurately near 98 percent on multi-modal biometrics. In addition, the Sensitivity results also described that the multi-modal SNN was 95 percent. The

experiment also investigated the n-way one-shot classification using 5, 10 and 25 classes in support set with one sample per class. The proposed SNN model on multi-modal achieved 100 percent accuracy in all three ways one-shot classification. Thus, the proposed approach is promising for human identification with high performance to be applied in the multi-modal biometric system.

Future work for this research is to consider GE image extraction from front view image sequences with various kinds of walking conditions. And next concern is to exploit the fore-ground objects detection and extraction in in-door environment that have lots of glasses and illuminations. More fusion methods can be applied to features integration.

REFERENCES

[1] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of fingerprint recognition*. Springer Science & Business Media, 2009.

[2] A. S. Buddharpawar and S. Subbaraman, "Iris recognition based on pca for person identification," *International Journal of Computer Applications,* vol. 975, p. 8887, 2015.

[3] B. A. Rosdi, C. W. Shing, and S. A. Suandi, "Finger vein recognition using local line binary pattern," *Sensors,* vol. 11, no. 12, pp. 11357-11371, 2011.

[4] A. Joshi, A. K. Gangwar, and Z. Saquib, "Person recognition based on fusion of iris and periocular biometrics," in *2012 12th international conference on Hybrid Intelligent Systems (HIS),* 2012: IEEE, pp. 57-62.

[5] S. Maity, M. Abdel-Mottaleb, and S. S. Asfour, "Multimodal low resolution face and frontal gait recognition from surveillance video," *Electronics,* vol. 10, no. 9, p. 1013, 2021.

[6] J. Han, B. J. I. t. o. p. a. Bhanu, and m. intelligence, "Individual recognition using gait energy image," vol. 28, no. 2, pp. 316-322, 2005.

[7] Y. G. Kim, K. Y. Shin, E. C. Lee, and K. R. Park, "Multimodal biometric system based on the recognition of face and both irises," *International Journal of Advanced Robotic Systems,* vol. 9, no. 3, p. 65, 2012.

[8] M. Gomez-Barrero, J. Galbally, and J. Fierrez, "Efficient software attack to multimodal biometric systems and its application to face and iris fusion," *Pattern Recognition Letters,* vol. 36, pp. 243-253, 2014.

[9] H. F. Liau and D. Isa, "Feature selection for support vector machine-based face-iris multimodal biometric system," *Expert Systems with Applications,* vol. 38, no. 9, pp. 11105-11111, 2011.

[10] B. Bhanu and J. Han, *Human recognition at a distance in video*. Springer Science & Business Media, 2010.

[11] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," *Pattern Recognition Letters,* vol. 141, pp. 61-67, 2021.

[12] D. Chicco, "Siamese neural networks: An overview," *Artificial Neural Networks,* pp. 73-94, 2021.

[13] M. O. Oloyede and G. P. Hancke, "Unimodal and multimodal biometric sensing systems: a review," *IEEE access,* vol. 4, pp. 7532-7555, 2016.

[14] X. Zhou and B. Bhanu, "Feature fusion of face and gait for human recognition at a distance in video," in *18th international conference on pattern recognition (ICPR'06)*, 2006, vol. 4: IEEE, pp. 529-532.

[15] E. Hossain and G. Chetty, "Multimodal identity verification based on learning face and gait cues," in *International Conference on Neural Information Processing*, 2011: Springer, pp. 1-8

[16] D. Zhang, Y. Wang, Z. Zhang, and M. Hu, "Ethnicity classification based on fusion of face and gait," in *2012 5th IAPR International Conference on Biometrics (ICB)*, 2012: IEEE, pp. 384-389.

[17] X. Xing, K. Wang, and Z. Lv, "Fusion of gait and facial features using coupled projections for people identification at a distance," *IEEE Signal Processing Letters,* vol. 22, no. 12, pp. 2349-2353, 2015.

[18] Ghalleb, A.E.K.; Amara, N.E.B. Remote person authentication in different scenarios based on gait and face in front view. In Proceedings of the 2017 14th International Multi-Conference on Systems, Signals & Devices (SSD), 2017; pp. 486-491.

[19] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Transactions on Multimedia,* vol. 17, no. 11, pp. 2049-2058, 2015.

[20] S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, and N. M. Nasrabadi, "Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018: IEEE, pp. 3469-3476.

[21] A. S. Al-Waisy, R. Qahwaji, S. Ipson, and S. Al-Fahdawi, "A multimodal biometrie system for personal identification based on deep learning approaches," in *2017 Seventh international conference on emerging security technologies (EST)*, 2017: IEEE, pp. 163-168.

[22] K. Gunasekaran, J. Raja, and R. Pitchai, "Deep multimodal biometric recognition using contourlet derivative weighted rank fusion with human face, fingerprint and iris images," *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije,* vol. 60, no. 3, pp. 253-265, 2019.

[23] W. Kim, J. M. Song, and K. R. Park, "Multimodal biometric recognition based on convolutional neural network by the fusion of finger-vein and finger shape using near-infrared (NIR) camera sensor," *Sensors,* vol. 18, no. 7, p. 2296, 2018.

[24] I. Boucherit, M. O. Zmirli, H. Hentabli, and B. A. Rosdi, "Finger vein identification using deeply-fused Convolutional Neural Network," *Journal of King Saud University-Computer and Information Sciences,* 2020.

[25] Silva, Pedro H., et al. "Multimodal feature level fusion based on particle swarm optimization with deep transfer learning." 2018 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2018.

[26] Lumini, Alessandra, and Loris Nanni. "Deep learning and transfer learning features for plankton classification." Ecological informatics 51 (2019): 33-43.

[27] Tao, Zhiyong, et al. "Finger-vein recognition using bidirectional feature extraction and transfer learning." Mathematical Problems in Engineering 2021 (2021).

[28] Therar, Huda Moyasar, Lect Dr Emad Ahmed Mohammed, and Ahmed Jadaan Ali. "Multibiometric system for iris recognition based convolutional neural network and transfer learning." IOP Conference Series: Materials Science and Engineering. Vol. 1105. No. 1. IOP Publishing, 2021.

[29] Daas, Sara, et al. "Multimodal biometric recognition systems using deep learning based on the finger vein and finger knuckle print fusion." IET Image Processing 14.15 (2020): 3859-3868.

[30] Zhu, Hongyi, et al. "Human identification for activities of daily living: A deep transfer learning approach." Journal of Management Information Systems 37.2 (2020): 457-483.

[31] Derbel, A., D. Vivet, and Bc Emile. "Access control based on gait analysis and face recognition." Electronics Letters 51.10 (2015): 751-752.

[32] Kurban, Onur Can, Tülay Yildirim, and Ahmet Bilgic. "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). IEEE, 2017.

[33] Koo, Ja Hyung, et al. "CNN-based multimodal human recognition in surveillance environments." Sensors 18.9 (2018): 3040.

[34] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815-823.

[35] Zhang, Cheng, et al. "Siamese neural network based gait recognition for human identification." 2016 ieee international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016.

[36] Hayale, Wassan, Pooran Negi, and Mohammad Mahoor. "Facial expression recognition using deep siamese neural networks with a supervised loss

function." 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019.

[37] Song, Lingxue, et al. "Occlusion robust face recognition based on mask learning with pairwise differential siamese network." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[38] Yu, Jun, et al. "Deep fusion siamese network for automatic kinship verification." 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020.

[39] Chakladar, Debashis Das, et al. "A multimodal-Siamese Neural Network (mSNN) for person verification using signatures and EEG." Information Fusion 71 (2021): 17-27.

[40] Bashar, A. Survey on evolving deep learning neural network architectures. Journal of Artificial Intelligence 2019, 1, 73-82.

[41] Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; Zafeiriou, S. Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641 2019.

[42] Maddalena, L.; Petrosino, A. A self-organizing approach to background subtraction for visual surveillance applications. IEEE Transactions on Image Processing 2008, 17, 1168-1177.

[43] Selvarasu, N.; Nachiappan, A.; Nandhitha, N. Euclidean distance based color image segmentation of abnormality detection from pseudo color thermographs. International Journal of Computer Theory and Engineering 2010, 2, 514.

[44] Yu, S.; Tan, D.; Tan, T. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August; pp. 441-444.

[45] A, S, Georghiades., P,N, Belhumeur., D, J, Kriegman.: 'From few to many: Illumination cone models for face recognition under variable lighting and pose', IEEE Trans. Pattern Analysis and Machine Intelligence, 23, (6), pp. 643–660, 2001

[46] Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: a local SVM approach. In Proceedings of the Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 26–26 August 2004; pp. 32-36.

[47] C. Manliguez, "Generalized confusion matrix for multiple classes," *URL* https://www. *researchgate. net/publication/310799885_Generalized_Confusion_Matrix_for_Multiple_Classes, DOI,* vol. 10, 2016.

[48] S. Marcel, M. S. Nixon, and S. Z. Li, *Handbook of biometric anti-spoofing*. Springer, 2014.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, pp. 1097-1105, 2012.

[50] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818-2826.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

List of Publications

International Journal

1. Aung, Hsu Mon Lei, and Charnchai Pluempitiwiriyawej*. Kazuhiko Hamamoto[3] and Somkiat Wangsiripitak[4] “Multimodal Biometrics Recognition using a Deep Convolutional Neural Network with Transfer Learning in Surveillance Videos”. Computation(ISSN 2079-3197) on 06 June 2022.(Accepted)

International Conference Proceedings

2. Aung, Hsu Mon Lei, and Charnchai Pluempitiwiriyawej. "Gait Biometric-based Human Recognition System Using Deep Convolutional Neural Network in Surveillance System." In *2020 Asia Conference on Computers and Communications (ACCC)*, Singapore, Singapore, pp. 47-51. IEEE, 2020.
3. Aung, Hsu Mon Lei, Charnchai Pluempitiwiriyawej , Somkiat Wangsiripitak and Kazuhiko Hamamoto, “Siamese Neural Network based Multi-modal Biometrics Recognition”, *Regional Conference on Electrical and Electronic Engineering (RCEEE) 2021*, Bangkok, Thailand, 6-7 Jan

VITA

NAME	Hsu Mon Lei Aung
DATE OF BIRTH	11 Aug 1987
PLACE OF BIRTH	Myanmar
INSTITUTIONS ATTENDED	Yangon Technological University
HOME ADDRESS	Room 509, Rajatavee Apartment, Phetchaburi Soi 18, Ratchathewi. Bangkok.