CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

# A Comprehensive Comparative Analysis of State-of-the-Art (SOTA) Face Recognition Algorithms under Diverse Degradation Conditions

by

Amir Hamza

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Engineering
Department of Electrical Engineering

2022

*With love to Nazakat Hussain, Nagina Rani, Ghulam-e-Fatima (Late), Ghulam Nabi (Late), Farkhanda Yasmin, Shazia Naseem, Asif Iqbal, Khalida Jabeen, AbuBakar, Umar Farooq, Jazib and two little Zaynab & Khadija*

## CERTIFICATE OF APPROVAL

# A Comprehensive Comparative Analysis of State-of-the-Art (SOTA) Face Recognition Algorithms under Diverse Degradation Conditions

by

Amir Hamza

(MEE-183008)

## THESIS EXAMINING COMMITTEE

| S. No. | Examiner | Name | Organization |
|---|---|---|---|
| (a) | External Examiner | Dr. Zahid Halim | GIKI, Swabi |
| (b) | Internal Examiner | Dr. Nadeem Anjum | CUST, Islamabad |
| (c) | Supervisor | Dr. Imtiaz Ahmad Taj | CUST, Islamabad |

Dr. Imtiaz Ahmad Taj
Thesis Supervisor
May, 2022

Dr. Noor Muhammad Khan
Head
Dept. of Electrical Engineering
May, 2022

Dr. Imtiaz Ahmad Taj
Dean
Faculty of Engineering
May, 2022

# *Author's Declaration*

I, **Amir Hamza** hereby state that my MS thesis titled "**A Comprehensive Comparative Analysis of State-of-the-Art (SOTA) Face Recognition Algorithms under Diverse Degradation Conditions**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

**Amir Hamza**
Registration No: MEE-183008

# *Plagiarism Undertaking*

I solemnly declare that research work presented in this thesis titled "**A Comprehensive Comparative Analysis of State-of-the-Art (SOTA) Face Recognition Algorithms under Diverse Degradation Conditions**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

**Amir Hamza**

Registration No: MEE-183008

# *List of Publications*

It is certified that following publication(s) have been made out of the research work that has been carried out for this thesis:-

1. M. Ullah, **A. Hamza**, I. A. Taj and M. Tahir, “Low Resolution Face Recognition using Enhanced SRGAN Generated Images" *16th International Conference on Emerging Technologies (ICET)*, IEEE, 2021.
2. **A. Hamza**, M. Ullah, I. A. Taj and M.T. Awan, “An Improved Multispectral Face Recognition System based on Deep Learning Models" *Journal of Electronic Imaging*, I.F = 1.006 [Under Submission].

**(Amir Hamza)**<br>
Registration No: MEE-183008

# *Acknowledgement*

Then which of the Blessings of your Lord will you deny. (Al-Quran).

First and foremost to the creator, the most gracious, the most beneficent, the Almighty **ALLAH S.W.T**, I owe it all to you, Thank you!

There have been many people who have walked alongside me, who have guided me through all these efforts. I would like to outstretch gratitude to each one of them. Topping the list is my supervisor **Dr. Imtiaz Ahmad Taj** to whom i owe my deepest gratitude for providing his valuable guidance to complete this research. Beside that i am also very grateful to my lab-mate **Mr. Mohsin Ullah** for his unconditional help as well as technical & motivational support thorough out the research journey. Moreover, I would like to thank each member of the VisPRS research group for their kindness.

Furthermore, I owe a great deal to my teachers and parents who shaped me into the person I am today. Their continuous support and encouragement made this work possible. Nevertheless, I also want to acknowledge my grandmother's unconditional love and unending prayers for me.

I also acknowledge partial support from National Center of Big Data and Cloud Computing (NCBC) and HEC of Pakistan for conducting this research.

**(Amir Hamza)**

# *Abstract*

Due to technological advancements and the transfer of huge amounts of sensitive data every day, biometric authentication has recently dominated the market. In past access to certain data or services is typically gained via documents or a password. However, these methods have proven unreliable over time. As an alternative, biometric systems based on fingerprint, iris, voice, face recognition, or a combination of these can be used. A facial recognition algorithm identifies or verifies a person in a still image or video by using a database of stored facial images. There have been several advances in face recognition over the last two decades. Consequently, face recognition systems have now achieved satisfactory performance under controlled conditions. The systems are, however, hampered by varying illumination, pose and expression.

In this study, we investigate how different face recognition and verification algorithms based on deep learning techniques perform under a variety of adverse conditions, such as pose effects, aging effects, resolution effects, cross-spectral matching and ethnicity effects. Five pre-trained deep learning models including FaceNET, VGGFace2, SphereFace, CosFace and ArcFace are evaluated. ArcFace trained using angular margins, can be seen clearly outperforming the counterparts in all of the scenarios. In addition to that a novel technique for direct cross spectral matching has also been proposed and have shown some promising results by increasing the recognition accuracy upto 7% to 8%.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **CFR** | Cross-Spectral Face Recognition |
| **CNN** | Convolutional Neural Network |
| **DL** | Deep Learning |
| **FR** | Face Recognition |
| **HR** | High Resolution |
| **LFW** | Labelled Faces in Wild |
| **LR** | Low Resolution |
| **LMCL** | Large Margin Cosine Loss |
| **LDML** | Logistic discriminant metric learning |
| **ML** | Machine Learning |
| **MLP** | Multi-layer Perceptron |
| **MTCNN** | Multi-task Cascaded Convolutional Networks |
| **NSL** | Normalised Softmax Loss |
| **SENet** | Squeeze and Excitation Network |
| **SGD** | Stochastic Gradient Descent |
| **SOM** | Self Organizing Map |
| **YTF** | Youtube Faces |

# Symbols

| | |
|---|---|
| $W$ | Weight vector |
| $x$ | Feature vector |
| $\theta$ | Angle between weight and feature vector |
| $m$ | margin penalty |
| $p$ | Posterior probability |
| $f$ | Activation function |
| $C$ | Total no. of classes |

# Chapter 1

# Introduction

## 1.1 Background

In human beings, biometrics relate to physiological and behavioural characteristics that are used to identify them automatically. During the last several decades, biometric identification systems based on biometric techniques such as the face, iris, fingerprint, and palm print have been admired in the industry. One of the most vastly used biometric modalities is the human face, which is famous for its contactless acquisition, social acceptability, and suitability for usage in non-cooperative circumstances.The field of machine-based face recognition has garnered a great deal of interest in recent decades, particularly in the areas of biometrics, pattern recognition, and computer vision research. In addition to primary and demanding challenges in this domain, researchers are driven by everyday applications such as those in financial services, forensics, authentication, and video surveillance, among other areas. Now a days many commercial face recognition systems are in place, and they are capable of meeting a wide range of requirements while also making a positive contribution to society.

Every face, like a fingerprint, is distinctive, even when identical twins appearances are compared [1]. This means that a facial recognition system's accuracy should be comparable to a fingerprint scanner. Finding a proper balance between a facial recognition technique's computing speed and accuracy is a big issue that requires more research. The system must be precise and accurate while being rapid enough to be inconvenient.

Many commercial companies employ face recognition, such as Facebook, which uses facial recognition to tag individuals in photos automatically. Selfie Pay is a payment system developed by Mastercard that uses face recognition. Facial recognition has also been used to take school attendance automatically. Specialized facial recognition systems for video surveillance are designed to identify the presence of certain persons across a dispersed network of video cameras under uncontrolled capture situations. Thus, detecting the faces of target people in such an environment is a difficult task since the look of faces fluctuates due to changes in pose, size, lighting, occlusion, and blur, among other factors. This is a major hurdle for modern computer systems when it comes to facial recognition.

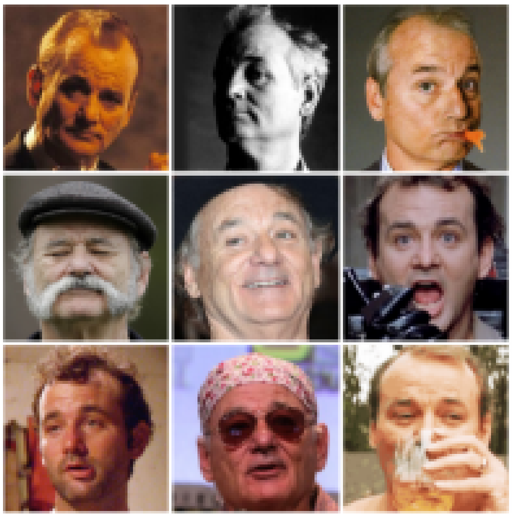


Figure 1.1: 9 Different images of same individual taken at different ages

To illustrate how difficult this challenge might be, Figure 1.1 shows multiple images of the same subject. Regardless if they all belong to the same individual, even a human is unlikely to recognise them as such. Computational complexity is also a factor to consider, given the increasing number of cameras and the processing time of cutting-edge face identification, tracking, and matching algorithms. Meanwhile, another field of research - artificial neural networks - was developing. This is inspired by the human brain structure and has shown to be a game-changer for several technical issues. Nowadays, one of the most frequently investigated ways for solving the face recognition issue is the deep neural network approach. Deep Learning is at the frontier of what computers are capable of. It seems to perform the best of all face recognition systems.

## 1.2 Identification vs Verification

Generally speaking, the word "facial recognition" relates to two basic scenarios: one for verification or authentication and another for identity or recognition. One's biometric template is checked against the claimed identity solely in the context of the verification task. However, in the context of the identification task, it is matched against every template registered in the gallery as depicted in Figure 1.2.

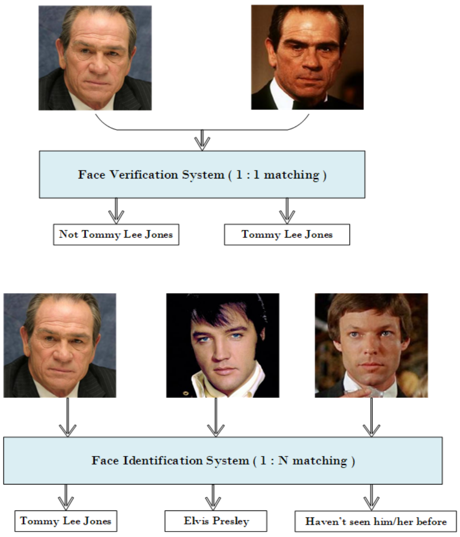

Face Verification System ( 1 : 1 matching )
Not Tommy Lee Jones
Tommy Lee Jones
Face Identification System ( 1 : N matching )
Tommy Lee Jones
Elvis Presley
Haven't seen him/her before

FIGURE 1.2: An Example of Verification vs Identification

### 1.2.1 Performance Metric : ROC vs CMC Curve

It is possible to examine the outcomes of verification trials in terms of the Receiver Operating Characteristic (ROC) Curve, which illustrates the Verification or True Acceptance Rate (TAR) as a trade-off against the False Acceptance Rate (FAR) (FAR). The Verification Rate is the proportion of a set of probe face images that is correctly accepted. At the same time, the False Acceptance Rate reflects the percentage of a group of probe face photos that is erroneously accepted. The

Verification Rate (TAR) of 0.1 % is the most usually quoted single value from the ROC curve.

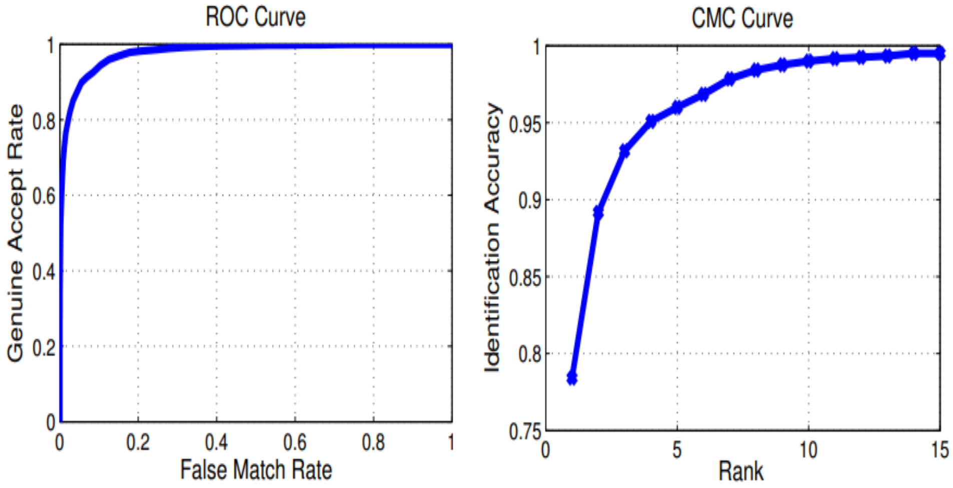


FIGURE 1.3: An Example Illustrating ROC vs CMC Curve [2]

The efficiency of face recognition algorithms can be perceived using the Cumulative Match Characteristic (CMC) curve. The Cumulative Match Curve (CMC) is a performance indicator for 1:N identification systems that is used to compare two sample facial images. ROC curves of verification systems, on the other hand, are used to indicate the quality of a 1:1 matching systems. The Rank-1 recognition/identification rate or accuracy is the most usually quoted single metric from the CMC curve since it is the most straightforward to calculate. Figure 1.3 shows an example of both a ROC and a CMC curve.

## 1.3 Common Terminologies in Face Recognition

As the field of Face Recognition has developed recently, Some new terms are introduced to gauge the performance of FR algorithms. Following are some of the important terminologies one must understand in detail while working with face recognition algorithms:

1. HR vs LR FR Problem

2. Gallery vs Probe Set

3. Open vs Close Set

### 1.3.1 High Resolution vs Low Resolution Face Recognition Problem

The literature lacks a clear definition of what constitutes a high-resolution image and what defines a low-resolution image. Typically, images recorded with still high-resolution cameras in constrained scenarios are referred to as high resolution images, and face recognition tasks involving both gallery and probe images with a high resolution are classified as high resolution face recognition problems. Numerous studies have been conducted on HR human face recognition, and state-of-the-art algorithms now outperform humans in terms of recognition accuracy.

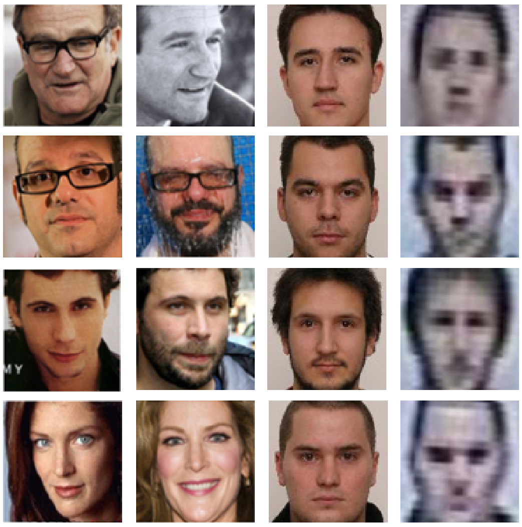


FIGURE 1.4: Coloumn (1-2) : High Resolution Gallery and Probe images from MegaFace Database Coloumn (3-4) : Low Resolution Gallery and Probe images from SCFace Database

Facial images that are relatively low in resolution i.e below 32x32 pixels, are recognized as low resolution images and are the possible cause of degradation in the performance of face recognition algorithms. Challenges that specifically a low resolution in facial images brings include degradation because of camera noise, Occlusion, Scale variation, motion blur and out of focus blur. So, with the aforementioned degradation's in facial images it is a challenge to match a low resolution probe image with a high resolution gallery image. LR face recognition is an under researched topic as compared to its counter part. Some of the most popular approaches in LR face recognition involves resolution invariant facial models and super-resolution based techniques. An Example of Set of facial images involved in both high resolution and low resolution face recognition is depicted in Figure 1.4

### 1.3.2 Gallery Set vs Probe Set

In literature gallery set is the collection of high definition still frontal images against whom the test images are matched. These images are the representative of template images that would be enrolled in a real-world facial recognition system for its deployment. Generally gallery images with neutral expression, frontal pose and even illumination levels are considered as the best gallery image.

A probe set is a collection of probe/under-test images of unknown individuals that need to be recognized or matched against the gallery or template image.For example in real world surveillance scenarios, the stream of images via cameras are captured and frame by frame facial images (probe images) are detected and matched against the gallery set.

### 1.3.3 Open-Set vs Closed-Set Problem

Face recognition systems can be assessed in closed-set or open-set environments, as seen in Figure 1.5 All testing identities are specified in the training set for the closed-set approach. It is natural for testing facial image to be classified according to their allotted labels. In these circumstances, Face verification is equal to perform identification on a pair of facial images as seen in the left hand side of Figure 1.5. As a result, closed set FR may be effectively treated as a classification problem having separable features.

In Contrast the test images identities are often isolated from the training identity in open-set protocols, which makes FR more demanding but close to practise. Due to the impracticality of classifying faces according to their known identities in the

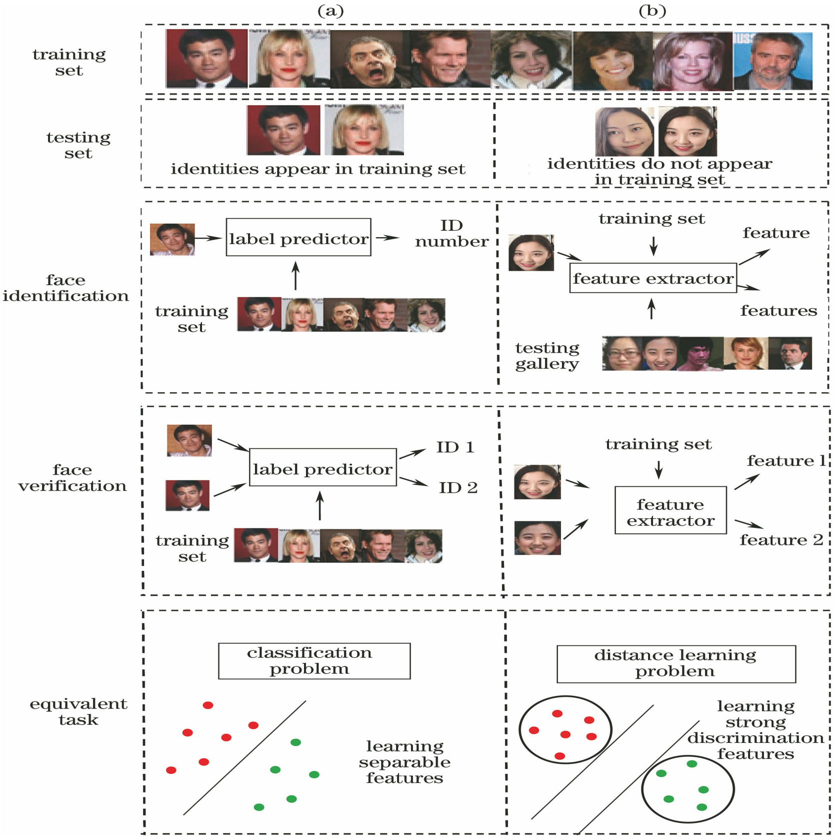


FIGURE 1.5: An Example of (a) Openset vs (b) Close-set Face Recognition

training set, we must map facial features to a discriminative feature space. Face identification in this case may be thought of as doing face verification between the probe facial image and each identity in the gallery. Face identification in this case may be thought of as doing face verification between the probe facial image and each identity in the gallery .

## 1.4 Thesis Structure

This thesis is structured as Follows :

1. Chapter 1 includes the details about the background and history of biometric systems with focus on FR systems and some commonly used terminologies in FR systems.

2. Chapter 2 is about the literature review of the FR technologies including some brief details about the face detection systems, parts based, holistic and

hybrid approaches, insights of deep learning technology and FR methods using this.

3. Chapter 3 includes the details about the FR models used for the evaluation.

4. Chapter 4 presents the details about the five scenarios i.e Age, Pose, Resolution, Cross Spectral matching and the effect of ethnicity under which all of the models are tested and the databases used for these scenarios.

5. Chapter 5 presents the results of the evaluation done under all of the selected scenarios.

6. Chapter 6 presents the details of the novel approach utilized for improving the Cross-Spectral Matching (RGB to Thermal).

7. Chapter 7 will be about the Conclusions drawn and the Future Work related to this research.

## 1.5 Summary

The purpose of this chapter is to provide an overview of the background history, the applications and need for biometrics in everyday life. It also includes the explanation about some of the common terminologies i.e high resolution vs low resolution, gallery set vs probe set and openset vs closeset face recognition problem, used in face recognition and the performance metrics employed for the evaluation. Additionally, it summarizes the general organization of the thesis.

# Chapter 2

# Literature Review, Problem Statement & Research Contribution

## 2.1 Introduction

Since the 1970s, face recognition has become one of the most actively researched problems in the fields of computer vision and biometrics. Deep neural networks trained on very large datasets have recently surpassed conventional methods relying on hand-crafted features as well as traditional machine learning techniques. Face recognition may be structured as a classification challenge, which allows for the use of a variety of machine learning techniques for the sake of developing a robust approach that infers the person's identification automatically. Each machine learning algorithm starts with a dataset and learns from it. Following the encoding of each instance/sample (in this example, an image) with a feature vector, a learning algorithm traverses the data and identifies patterns. Due to the intrusive aspects of posture, expression, and lighting, it is critical to choose a suitable (rich, with many within-class variations) dataset if we desire for our system to be resilient to these changes. Additionally, two additional difficulties are critical; the first is determining which traits to utilize to represent a face in order to be as resistant as feasible to all of these differences. The second is how to use the selected representation to categories a fresh facial image. This chapter will begin with the details about face detection followed a review of the literature on face recognition methods covering both classical ( holistic, feature-based, and hybrid) and deep learning approaches are presented, as well as the details about the basic preprocessing processes that should be performed prior to feature extraction and classification algorithms. Then we'll look at several strategies that have been extensively utilized and shown success.

## 2.2 Face Detection

Face detection is one of the important and critical applications of computer vision. Many methods for detecting multiple facial features have been introduced in the last decade. Convolutional neural networks (CNN) and deep learning have, however, recently shown great success. A face detector detects a human face in a digital image based on its location and size. It depends on determining whether any faces are present in a given image and resulting the bounding box of each detected face. All facial analysis algorithms need to detect faces before they can perform alignment, recognition or verification operations. Due to the dynamic nature of human faces and their high degree of variability, it is difficult to detect them.

A facial part can be detected by means of two methods i.e feature-based and image-based techniques. Methods based on features attempt to find features that are invariant across faces. The main idea is inspired by human vision system that is capable of detecting faces in different poses and lighting conditions without much effort. Therefore, it must possess certain properties or features regardless of those variations in pose and lighting conditions. To detect the presence of a face, many different algorithms have been proposed in the past. Although feature-based approaches are easy to implement, there is a major problem with feature-based algorithms in the sense that illumination, noise, and occlusion can severely corrupt the image features. Additionally, the edges of features on faces can be weakened when shadows are present, which renders perceptual grouping algorithms ineffective. Methods based on images are aimed at learning templates from examples.

Appearance based methods analyze images of "face" and "no-face" relying on machine learning and statistical techniques. These characteristics are either expressed as distribution models or discriminant functions, which are then applied to the task of detecting faces. The most common image-based approaches include CNNs [3], SVMs [4], and Adaboosts [5]. Although several studies had been conducted before 2000 but prior to the groundbreaking research proposed by Viola and Jones [6], there was no satisfactory FR method evolved. A deep learning based SOTA face detector is can detect faces with upto 800 faces out of 1000 reported in the World's Largest Selfie [7] as seen in Figure 2.1

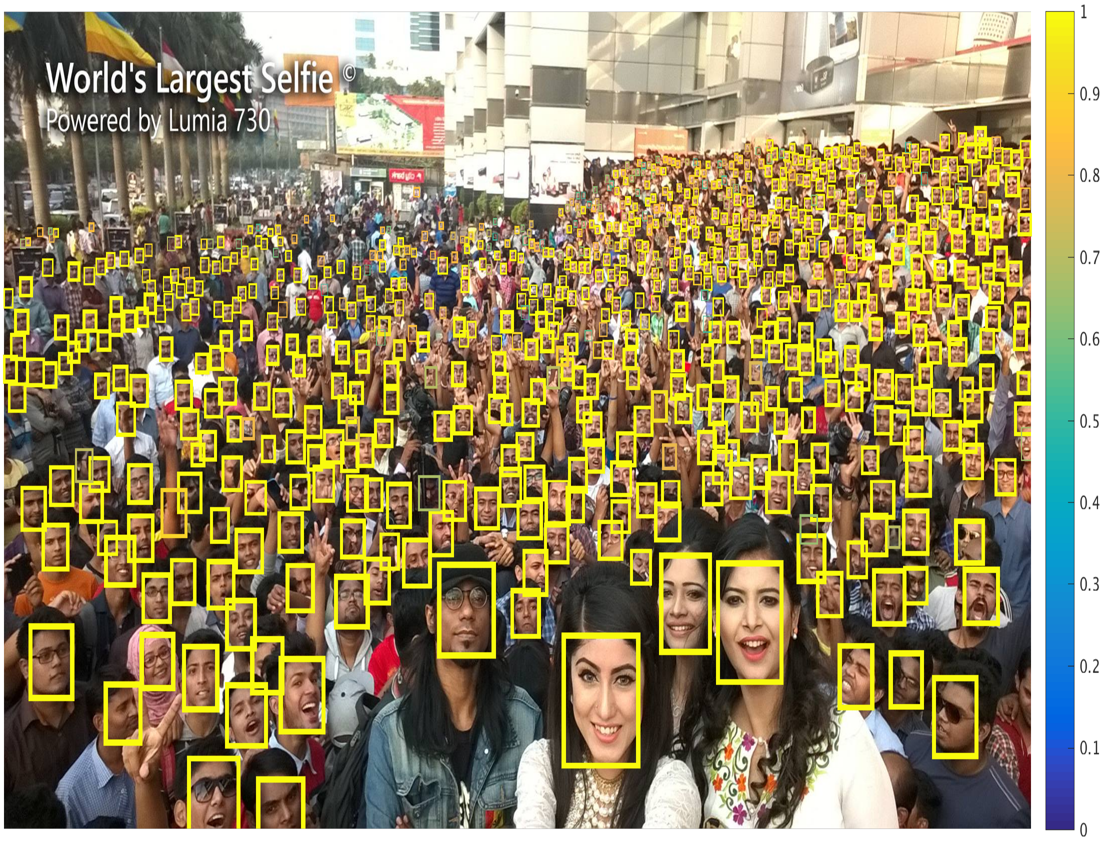


FIGURE 2.1: Face detector in action on World's Largest Selfie [7]

The face detection field has made great progress since the innovative work by Viola and Jones. Through training a detector using Haar features and AdaBoost, they were able to detect faces more accurately, leading to the development of several approaches over time.. Despite this, the detector has some critical drawbacks. For example, the size of its features was relatively large. Additionally, it struggles to handle faces in the wild or non-frontal faces. Traditional machine learning algorithms were mostly used for training classifiers for detection based on handcrafted features extracted by domain experts in computer vision. It is often impossible to find the effective features within the face using these methods, Separate optimizations are applied to every component, resulting in the inefficiency of the detection pipeline.

In order to deal with the primary difficulty, researchers have devised further complex features such as HOG [8], SIFT [9] and SURF [10]. An enhanced level of detection has been achieved by merging multiple detectors that have been trained separately for different views or poses. However, training and testing such models was often more time-consuming, with relatively limited results in terms of improved detection performance.

A key advantage of deep learning methods over traditional computer vision approaches is that they avoid the handcrafted design process, and they have dominated many well-known benchmarks like the ILSVRC [11]. In recent years, scientist have applied one of the most popular generic object detectors, the Faster R-CNN [12], with promising outcomes. Furthermore, joint trainings conducted using CNN cascade, region proposal networks (RPNs), and Faster R-CNNs have contributed to end-to-end optimization. With the combined use of hard negative mining and ResNet, a faster R-CNN face detection algorithm was developed that achieved significant improvements in recognition performance on benchmarks such as FDDB [13]. Another popular approach is multi-task cascaded convolutional neural network, or MTCNN for short.

### 2.2.1 MTCNN

MTCNN [14] employed a cascade structure as seen in Figure 2.2, three networks are employed; first, an image is resized (called an image pyramid), next, a proposal network (P-Net) proposes areas of interest, then a refine network (R-Net) refines bounding boxes, and the output network (O-Net) provides landmarks for facial identification. As opposed to being directly connected, the outputs of each stage are fed into the previous stage. In this way, it is possible to perform additional processing between stages as well; for example, a non-maximum suppression (NMS)

[15] filter can be applied to candidate bounding boxes as they are provided by the P-Net in the first stage before being introduced into the second stage R-Net.

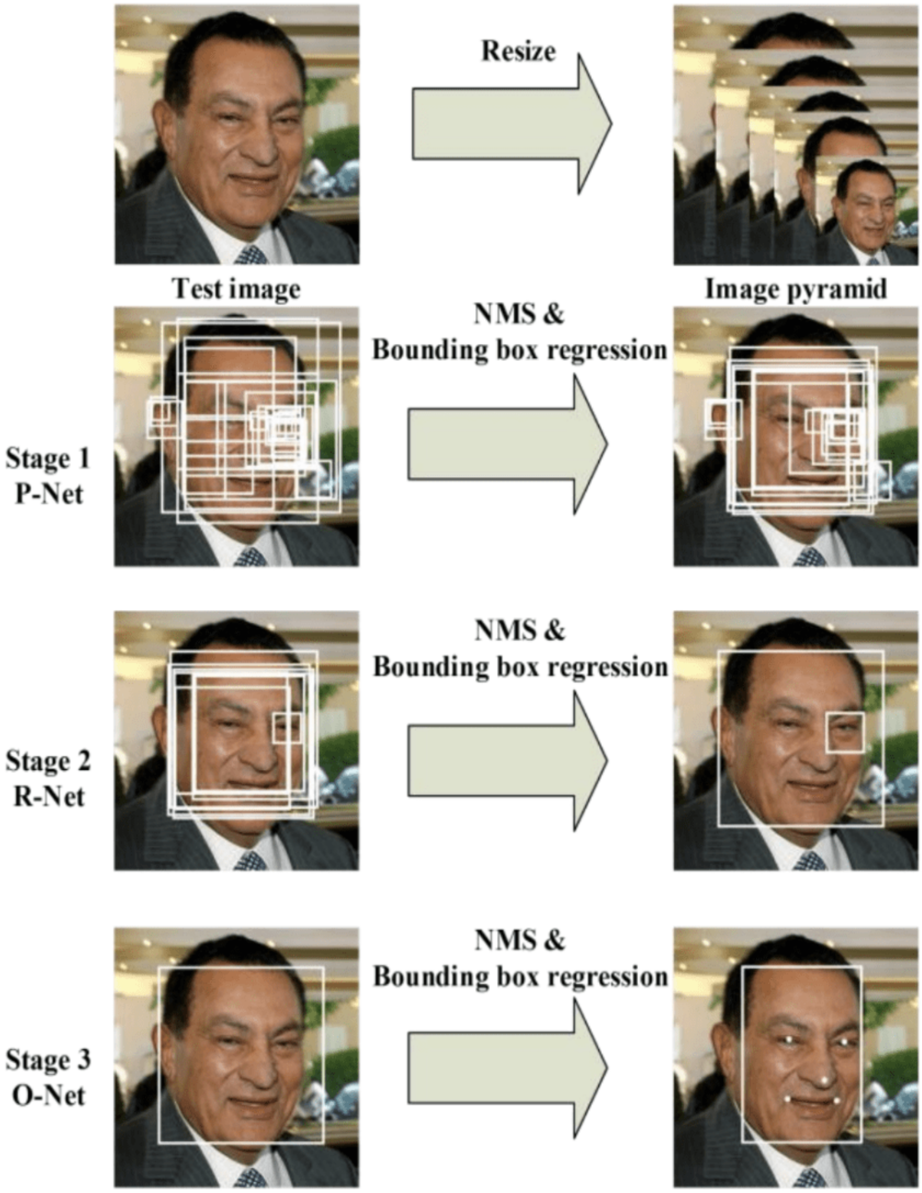


Figure 2.2: Multi-Task Cascaded Convolutional Neural Network (MTCNN) pipeline [14]

Implementing the MTCNN architecture is relatively complex. The architecture is open source, so there are implementations of it that can be trained on new datasets as well as pre-trained models you can use for face detection directly. We have used the Caffe [16] official implementation of MTCNN in this research to detect faces among different images utilized for the comparative analysis, It worked well for all the scenarios except the thermal images, where it is unable to detect faces, multiple hyperparameters for controlling the sensitivity of the detectors were also tuned, but this did not work. So we used RetinaFace detector for detecting the faces in Thermal images.

### 2.2.2 RetinaFace

Like MTCNN, there are many othre top of the line face detectors like TinyFace [17], SSH [18], PCN [19] and RetinaFace [20]. With RetineFace, you can perform three different tasks, including face detection, 2D facial alignment, and 3D facial reconstruction based on a single shot. Three different targets are solved keeping only one thing in mind: that all points in the regressed data for all three tasks should be on an image plane. Three main components make up the model including Feature Pyramid Network, Context Head Module and Cascade Multi Task Loss

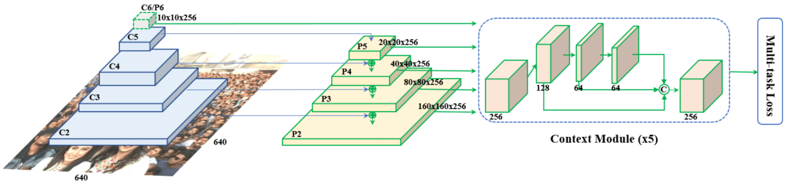


FIGURE 2.3: RetinaFace model architecture [20]

The Feature Pyramid Network produces five feature maps of varying scales based on the input image. The first four feature maps in the Figure 2.3 are calculated by ResNet [21] architecture, which was pre-trained on an Imagenet [22] dataset of 11k images. A convolution of 3x3 and stride 2 was applied to C5 to create the top most feature map.

A deformation convolutional network (DCN) [23] is used instead of a normal 3x3 convolution in this module to enhance the context modelling capability. Cascade regression is used along with multi-task loss to improve face localization. In the first context module, regular anchors are used to predict the bounding box, and the subsequent modules use regressed anchors for more precise predictions. The first context head module matches ground truth boxes to anchors if their IoU is more than 0.7 and background to anchors if it is less than 0.3. The second context head module matches anchors to ground truth boxes if their IoU is more than

0.5 and background to anchors if it is less than 0.4. Training examples are both positive and negative using Online Hard Example Mining(OHEM) [24].

## 2.3 Traditional Face Recognition Methods: A Review

A person's distinctive facial features are not the first step in face identification, but they are a useful starting point. To begin, all of the faces in an image must be identified and extracted.

The first phase in a face recognition system is termed face detection as seen in Figure 2.4

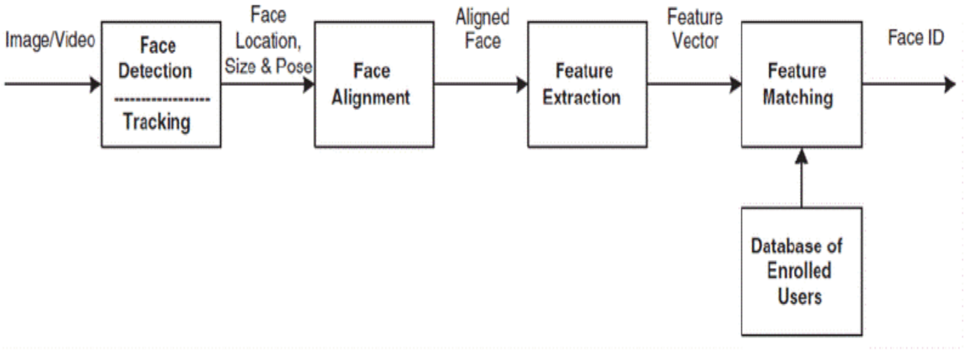


FIGURE 2.4: A Generic Face Recognition System Pipeline

Viola and Jones proposed a cascade of Adaboost classifiers in 2004 by extracting Haar features, which quickly became a popular approach for face recognition that could be done in real time. Their method involved the extraction of fast Haar features from integral images, using cascade to rapidly remove non-face areas using simple checks and the use of boosting for choosing out the most unique and important features. In the domain of face detection another effective method is the Histogram of Oriented Gradients (HOG) technique . Concatenating the HOGs of a given image patch's sub regions creates its feature vector, these feature vectors are then input into a linear SVM classification model, that decides whether an image patch is facial or not. Because a HOG-SVM face detector produces fewer false positives than a Haar cascade because it is more accurate and quicker than Haar cascades.

Local Binary Patterns (LBP) can be used in place of Haar features in conjunction with cascade boosted classifiers [8]. As a result of deep learning, we are now able to achieve ground breaking results for FR. YOLO (You only look once) [25] is one of many deep convolutional neural networks (DNN) used for object identification, also certain face detectors have been developed based on its design. Such detectors, on the other hand, may be rather sluggish. Another step of the preprocessing process that may increase the efficacy of face recognition is face alignment, which implies that critical facial features like the mouth and eyes are centralized also frequently in the same location on the image as each other. They function by detecting certain facial characteristics, including the mouth, eye brows, nose and jaw and then on the rotation, translation and scaling of those landmarks to produce an idealised depiction of the face, as shown in Figure 2.4. In [26], it was suggested to employ a distinctive facial landmark detector that effectively recognises 68 facial feature points, otherwise known as landmarks and can therefore be used for face alignment. After being provided with some training, facial image labelled with the facial landmarks position and prior knowledge about the distances between them, by training an ensemble of regression trees with gradient boosting ensemble only on the pixel intensity, the landmark positions can be estimated. Gradient boosting [27] is a practical application of the boosting concept to regression problems. An alternative alignment method is presented [28], which is based on the idea of congealing. This method aims to minimise the entropy of the empirical density function field at every single pixel.

Following the geometric normalisation stage illumination is also normalised. Due to the fact that facial images of the same individual seem significantly differently under various lighting circumstances, it becomes necessary to adjust for these differences and make the images more comparable. For example, histogram equalisation, which is used to increase contrast or intensity levels while also making the histogram more uniform, and the normalising approach described by [29], which is composed of three steps: first gamma correction second differences between

Gaussian filters and finally the contrast equalisation. Following the detection and extraction of faces, the process of normalised feature extraction is carried out. Feature-based [30], [31], holistic [32], and hybrid [33] face recognition techniques are the three types of face recognition algorithms that may be classified [34] depending on the characteristics that they collect. Feature-based techniques identify and extract facial landmarks for example the eyes, nose, and mouth and quantify their geometrical features as well as the distances between them in order to improve face recognition accuracy. These characteristics remain constant regardless of changes in lighting conditions or pose. However, the identification of such landmarks is not as accurate as it should be, lowering the efficacy of such systems. Furthermore, they do not take into consideration facial texture, which might be utilised to discriminate between people. Holistic approaches look at the facial image as a complete unit and aims to extract valuable statistical insights from the data that is provided by the input. In the early days of face recognition, holistic approaches predominated, followed by hybrid techniques which included local-based face descriptors into a single global feature vector. Eigenfaces [35], Fisherfaces [36], Independent Component Analysis (ICA) [37], and kernel based approaches [38]. They use pixel intensity characteristics as well as dimensionality-reduction approaches for both Eigenfaces and Fisherfaces. Principal Component Analysis (PCA) [39] defined as non-supervised dimensionality reduction methodology that keeps as much variability as possible while still correlating features. A Linear Discriminant Analysis (LDA) [40] is a supervised approach to reducing dimensionality which uses Fisher's discriminant ratio [41] to discover the projection that best differentiates the classes.

In unconstrained situations, Eigenfaces is particularly sensitive to pose, expression, and light fluctuations; Fisherfaces, on the other hand, is more resilient to pose and lighting variations since it additionally uses class label information. Face recognition may be improved by using ICA for dimensionality reduction as done by [42], that aims to capture a projection which not only makes features only uncorrelated but also independent. It is shown in [43] that using the kernel approach in PCA and LDA transforms results in superior nonlinear transforms than Eigenfaces, Fisherfaces, and ICA.

At the same time, hybrid approaches take advantage of local block-based characteristics, making them more resilient to pose and lighting issues. Binary Patterns at the local level face descriptors include the histogram, Gabor-based features [44], and two-dimensional DCT coefficients [45]. Because there are many micropatterns in faces, an operator known as Local Binary Pattern (LBP) [46] provides a texture descriptor which could be used for identifying faces. It was initially intended for a fixed 4x4 scale, as shown in Figure 2.5. However, various expansions were suggested in [47]. Concatenating histograms of such patterns from diverse locations results in the formation of feature vectors.

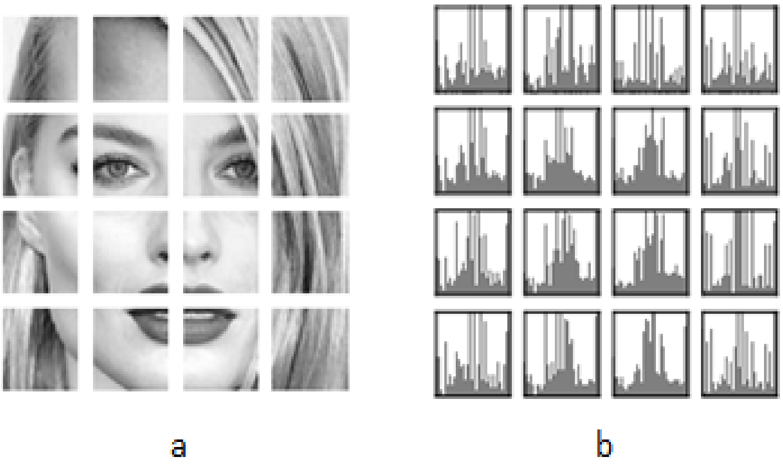


FIGURE 2.5: (a) Facial Image sectioned into 4x4 local regions. (b) An illustration of LBP histogram from each local area [46]

Simple classifiers like the K-nearest neighbour classifier, KNNC [48] (or nearest mean classifier, NMC [49]) could be used for classification. These classifiers (primarily based on distance) may be used with Euclidean distance [50], [51], (weighted) X2 distances, or cosine similarity [42] measures when histogram-based features are used. Another more complicated approach is the Support Vector Machine (SVM) classifier [52]. Now, we will examine other hybrid approaches that disobeys the pipeline depicted in Figure 2.4. In [53], low-level local features like image intensity levels in RGB and HSV colour spaces, edge magnitudes and gradient orientations, were often used to train features and simile binary SVM classifiers to calculate high-level visual features. Attribute classifiers identify characteristics of faces i.e. gender, ethnicity, and age. In addition, these classifiers discover indescribable characteristics by comparing various features of one face with a small selection of reference subjects. In order to compare two facial images, SVM classifiers are used to classify the outputs of images based upon attribute and simile classifiers. A method similar to the simile classification system described in [53] was presented in [54]. Primary disparities between [53] and [54] includes that [54] employed a long series of small 1-to-1 classifiers rather than a sophisticated 1-to-all classifiers in [53], and also that SIFT descriptors served as low-level features. [55] offered two techniques based on metric learning regarding face recognition. Hybrid approaches combine the benefits of holistic and feature-based approaches. However the primary constraint is a lack of good features capable of extracting all of the information required to identify a face.

Two more local-based face descriptors that have shown excellent results in the domain of face recognition are Gabor-based features and 2D DCT coefficients that will be discussed in the upcoming Section 2.4

## 2.4 Parts-based Face Recognition

### 2.4.1 Gabor features

In a wide variety of image processing applications, Gabor filters [56] have proven effective, including image smoothing, texture analysis, edge detection, iris and fingerprint identification, and face recognition. These have demonstrated to provide the best results in the time (spatial) and frequency domains equally. Gabor filter primarily is a Gaussian that has been modified by a complex exponential in two dimensions given by :

$$G(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \tag{2.1}$$

here

$x' = x\cos\theta + y\sin\theta, y' = -x\sin\theta + y\cos\theta, \theta$ represents the orientation of the perpendicular to the parallel stripes of the Gabor function, $\sigma$ is the standard deviation of the Gaussian envelope, $\gamma$ is the spatial aspect ratio that controls the ellipticity of the ellipses, $\lambda$ is the wavelength of the sinusoidal factor and $\psi$ is the phase offset.

In [57], feature vectors of different orientations and frequencies were obtained using a filterbank of odd-symmetric Gabor filters (the complex exponential in formula (2.1) reduces to a sine), and classification was conducted using a simple nearest mean classifier. The authors of [58] attempted to merge magnitude and phase information from Gabor-filtered pictures in order to generate a more detailed feature representation of the facial images than prior research efforts that focused only on magnitude information. PCA was used to minimise the amount of features before utilising SVM to do the classification job.

The two-dimensional Discrete Cosine Transforms (2D-DCT) were also employed in the application of face recognition and is preferred in comparison with the Discrete Fourier Transform (DFT) due to its exceptional compression qualities. When paired with polynomial coefficients, often referred as deltas, derived from neighbor blocks, Features of DCT are simpler to derive than the other Gabor feature format. After detecting the face and possibly normalising it geometrically and illumination (using a preprocessing approach such as Tan and Triggs normalisation or histogram equalisation), feature extraction is done via block-based DCT. Then facial image is then boken down into blocks of overlapping dimensions i.e MxN and feature vectors are rediscovered in every section by retaining a limited number of low-frequency DCT-II coefficients via a zig-zag approach, as illustrated in Figure 2.6.

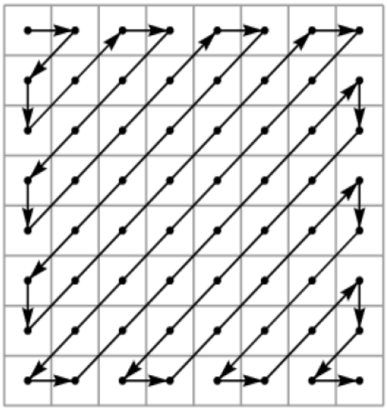


Figure 2.6: Method of extracting DCT coefficients using zig-zag scanning

## 2.5 Face Recognition using Deep Learning

The concept of deploying neural networks to identify faces is not new. In 1997, On the basis of a probabilistic decision-based neural network (PBDNN) [59], a technique for detecting faces, locating eyes, and identifying them has been developed. To limit the amount of hidden units and avoid overfitting, the face recognition PDBNN was separated into one fully connected subnet per training subject. Two PBDNNs were trained separately on intensity and edge characteristics and the outputs of both is merged in order to provide a final classification decision.

One of another early approaches [60] proposes using a self-organizing map (SOM) in conjunction with a CNN. SOM [61] is a sort of unsupervised neural network method for transforming data into a lower-dimensional space while preserving its input space's topological characteristics (i.e. When inputs are closely spaced in the original space, they are also closely spaced in the output space).

However, none of these two early approaches was trained end-to-end [59] used edge features and [60] used SOM), and the proposed neural network architectures were shallow. In [62], an end-to-end CNN for facial recognition was presented. This technique made advantage of a siamese architecture that had been trained using a contrastive loss function [63]. This algorithm provides a metric learning technique which seeks to minimise the distance between pairs of feature vectors pertaining to the same subject while maximizing the intra pair distance between feature vectors in accordance to dissimilar subjects. This approach also employed a shallow CNN architecture that was trained on relatively small datasets.

Because of the limited capability of the networks deployed and the limited amount of data available, none of the approaches outlined above achieved ground-breaking results. Not until these models were scaled up and trained on big datasets , Facebook's DeepFace [64], among the first techniques to use CNN for facial recognition, it took use of more deep model, scored 97.35% accuracy on LFW [65] database, a reduction in the error is more than 27% over prior state-of-the-art techniques. The newly proposed CNN is trained on a database of 4.4 million facial images from 4,030 participants using softmax loss. This study makes two unique contributions, One is the development of a successful facial alignment system based on explicit 3D modelling of faces, and a CNN architecture with locally connected layers [66], [67] that, unlike conventional convolutional layers, learns more distinctive features from every small region of an image. Simultaneously, another approach DeepID [68] reached at the comparable level of accuracy by training 60 unique CNNs on patches with 10 regions and 3 scales. 160 bottleneck features from every patch and its horizontally flipped counterpart were extracted in the testing phase, resulting in a 19,200-D feature vector (160x2x60).

Just as is the case with [64], the suggested CNN design use of locally linked layers. The verification result was achieved by training a joint Bayesian classifier [69] on the resultant 19,200-D vectors generated by the CNNs.

In order to train the algorithm, 202,599 facial images of 10,177 celebrities were used [68]. CNN architectures utilized for face recognition were influenced by those that achieve the highest level of accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). For instance, [70] employed VGG-16 architecture [71], while [72] used a comparable but smaller network. [73] investigated two distinct varieties of CNN architectures: VGG-style networks [71] and GoogleNet-style [74] networks. While both types of networks achieved equivalent accuracy, the GoogleNet-style networks had a factor of twenty times less parameters. Recently, residual networks (ResNets) [21] have established themselves as the dominant architecture for a variety of object identification tasks, including face recognition [75], [76], [77], [78], [79], [80], [81]. The primary innovation of ResNets is the addition of a building block that learns a residual mapping through a shortcut connection, as seen in Figure 2.7.

Due to the fact that shortcut connections enhance the flow of information between layers, they enable the training of much deeper architectures. [81] conducted a comprehensive examination of several CNN designs. The optimal trade-off between accuracy, inference time, and size of the model was found using a ResNet-100 [82] architecture backbone with a residual block identical with that of described in [83].

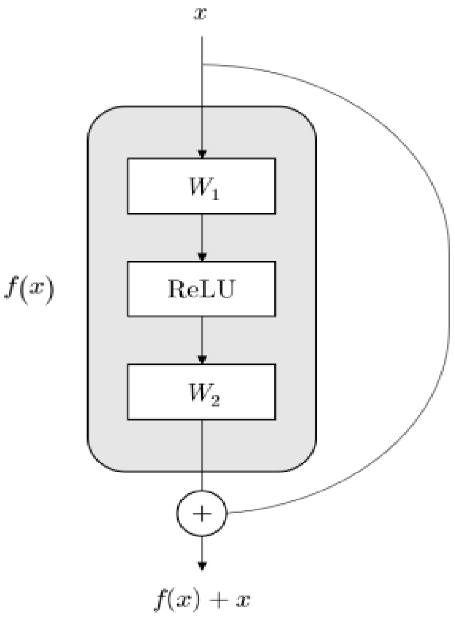


FIGURE 2.7: Residual block as proposed in [21]

### 2.5.1 Public Large Scale Datasets for training FR Models

Three major elements that influence the accuracy of CNN-based FR methods are the training data, CNN architecture, and loss function utilized for the training. As is the case with the majority of deep learning applications, extensive training sets are required to avoid overfitting. The accuracy of CNNs trained for classification increases with the number of samples per class.

It's because when the CNN model is subjected to greater intra-class variance, it is capable of learning more robust features. In face recognition, on the other hand, one is concerned for retrieving the generalized features of individuals not included in training set. Therefore, for face recognition a large number of individuals must be included in the dataset, so the model can be exposed to more interclass variance.

TABLE 2.1: Public large-scale available face databases

| Databases | No. of images | Subjects | Images per subject |
|---|---|---|---|
| CelebFaces+ [84] | 202,599 | 10,177 | 19.9 |
| UMDFaces [85] | 367,920 | 8,501 | 43.3 |
| CASIA-WebFace [86] | 494,414 | 10,575 | 46.8 |
| VGGFace [70] | 2.6M | 2,622 | 1,000 |
| VGGFace2 [87] | 3.31M | 9,131 | 362.6 |
| MegaFace [88] | 4.7M | 672,057 | 7 |
| MS-Celeb-1M [89] | 10M | 100,000 | 100 |

## 2.6 Comparative Analysis of FR Algorithms: A Review

Face Recognition has recently become a very popular research field because of some promising results of machine learning and deep learning-based techniques. Several algorithms have been proposed for face recognition. A few evaluation and comparative studies have also been conducted in order to evaluate the performance and effectiveness of FR systems. The following section and the Table 2.2 gives a brief overview of the comparative studies conducted in the past, brief detail and their shortcoming.

A total of 15 studies will be discussed along with their main contribution and limitation. Major studies were found to be old school and classic, as they compared only classical approaches, only few/limited studies were found to be related with Deep Learning Face Recognition Algorithms.

## 2.7 Gap Analysis and Problem Statement

Following are some of the gaps that are identified from the reviewed literature which relates to form a problem statement as follow:

- After a detailed study conducted in Section 2.60, it has been identified that there is a lack of comprehensive, impartial and unbiased comparative analysis of face recognition algorithms. Another important finding is that most of the techniques compared in Section 2.60 are conventional and old school. However a very limited amount of studies are seen involving state of the art machine learning and deep learning models.
- DCCNs have achieved excellent face verification and identification results on high resolution benchmark datasets. However their performance is still effected when the images have wide variations among Age, Pose, Resolution and Expression.

Table 2.2: Comparative Analysis of FR Algorithms done so far.

| **Study** | **Contribution/ Details** | **Limitations** |
|---|---|---|
| A Comparative Study of Baseline Algorithms of Face Recognition by Z. Mahmood et al [90].<br>Published Year : 2014 | This study presents the comparison of two classical approaches used for FR i.e. PCA and AdaBoost with LDA.<br>Two experiments involving pose and resolution variations were performed using PIE database. | Major limitations of this study includes the lack of variety of different experiments and the models used for performing the comparison are also limited. |
| Comparative analysis of advanced Face Recognition Technique by Kannan. et al. [91]<br>Published Year : 2014 | This study focuses utilizes Fuzzy C-Means clustering and parallel neural networks techniques to evaluate the performance for FR applications.<br>Recognition accuracy and Inference time is also calculated. A private database is used to evaluate the performance of the selected techniques. | Only two classical approaches are selected for the evaluation.<br>The database used for this study is very limited in numbers and in the variety of degradation present in the images. |
| Comparative study of some FR Algorithms by Lang et al. [92]<br>Published Year : 2008 | 2DPCA, SVD and fusion of both the classifiers are used for the evaluation on ORL and Yale face database. | Only facial expressions are used with limited illumination variation. Just 40 subjects are present in ORL and 165 images are there in Yale face database. |

Continuation of Table 2.2

| Study | Contribution/ Details | Limitations |
|---|---|---|
| Comparative Study of Face Recognition Classifier Algorithm by Zhan et al. [93]<br>Published Year : 2015 | This paper compares classical approaches as PCA, FLDA, SVM and Bayes Classifier. Recognition accuracy and Classification time are reported. AT&T facial database is used for the evaluation. | Only limited techniques and classical approaches for FR are tested. The database used is very limited i.e. 40 subjects and 400 images with limited degree of variation among faces. |
| A Multifaceted Independent Performance Analysis of Facial Subspace Recognition Algorithms by Usama et al. [94]<br>Published Year : 2013 | Six appearance based FR methods are used i.e. PCA, 2DPCA, A2DPCA, (2D)2PCA, LPP and 2DLPP to perform the independent evaluation. Three databases like FERET, ORL and YALE are used in this evaluation with expression, illumination and Ageing modalities. | Only classical approaches were evaluated in this study and the limited amount of facial modalities were under test. |
| A Comparative Study on Facial Recognition Algorithms by Sanmoy et al. [95]<br>Published Year : 2020 | Facial Recognition accuracy of Eigen faces with PCA, SVM, KNN, and CNN are selected for the comparison of performance. | Private database with limited number of individuals and facial modalities is used for the testing. |

Continuation of Table 2.2

| Study | Contribution/ Details | Limitations |
|---|---|---|
| A Comparative Study of Facial Recognition Techniques With focus on low computational power by Timmy et al. [96]<br>Published Year : 2019 | This study focuses on the evaluation of FaceNET, Eigen Faces and Fisher Faces with KNN.<br>The study reports accuracy, Recall, precision, F-score and Fall out. Training time and prediction time is also reported. | Limited portion of LFW database in used in this study.<br>A deep learning model is compared with the conventional technique of FR, which is not a fair comparison. |
| Comparative analysis of FR algorithms and investigation on the significance of colour by Behnam et al. [97]<br>Published Year : 2006 | This study presents the comparison of PCA, FLD, Laplacian faces and Gabor filters used in the application of Face recognition.<br>CVL Database and Georgia Tech Face database is used for the evaluation. | Selected amount of facial modalities are used for the comparison. Limited and classical FR techniques were compared. |
| A Comparative Analysis of Face Recognition Algorithms in Solving the Problem of Visual Identification by Gorbunov et al. [98]<br>Published Year : 2017 | This study compares the recognition results of Fisher Face, LBP and Eigen Face algorithms using facial images captured at three different distance i.e. 0.4m, 0.5m and 0.6m respectively. | A very brief study with insufficient number of facial modalities and FR models under test. |

Continuation of Table 2.2

| Study | Contribution/ Details | Limitations |
|---|---|---|
| A Comparative Study of Face Recognition Algorithms under Facial Expression and Illumination by Ali. et al. [99]<br>Published Year : 2019 | This study features LBPH, PCA and LDA for the analysis. The database used is Yale Face database and JAFFEE which contains 213 images of 10 individuals with 6 basic facial expressions.<br>Yale database contains 165 images of 15 individuals with varying illumination conditions. | Limited facial modalities were consider to test. Only few and classical methods are used to perform the analysis. |
| A Comparative Analysis of Face Recognition Algorithms by Gagan et al. [100]<br>Published Year : 2016 | Authors compared only PCA and LDA techniques. Only pose and illumination variations are tested. | Limited facial modalities were consider to test.<br>Only few and classical methods are used to perform the analysis. |
| A comparative study on face recognition techniques and neural network by Meftah. et al. [101]<br>Published Year : 2012 | PCA, MPCA and a Backpropagation Neural Network are compared in this study. AT&T database is used for the evaluation. | Experiments were conducted using only the subset of dataset. Also the database used contains limited amount of facial modalities. i.e Illumination and Facial Expression. |

Continuation of Table 2.2

| **Study** | **Contribution/ Details** | **Limitations** |
|---|---|---|
| A Comparative Study of Deep Learning Based Face Recognition Algorithms for Video Under Adverse Conditions by GALİP et al. [102]<br>Published Year : 2019 | This Study performs the comparison of 3 Deep learning models used for facial recognition i.e. FaceNET, VGGFace2 and ARCFace.<br>UvA-NEMO database is used to perform the evaluation, it contains HR frontal videos with different expressions. The Author then applied some noise to input video frames including Gaussian Blur, Gaussian Noise, Salt and Pepper Noise. | Only 3 Deep Learning models are studied under limited amount of facial modalities. |
| Face Recognition Comparative Analysis Using Different Machine Learning Approaches by Nisar et al. [103]<br>Published Year : 2021 | Four machine learning based techniques were considered for the evaluation including KNN, LDA, SVM and PCA.<br>The database used is gathered by the Olivetti Research Laboratory in Cambridge, UK and is publically available for testing and benchmarking the performance of FR algorithms | Limited facial modalities were consider to test. Only few and classical methods are used to perform the analysis. |

Continuation of Table 2.2

| Study | Contribution/ Details | Limitations |
|---|---|---|
| Comparative Analysis of Face Recognition Approaches: A Survey by Ripal et al. [104]<br>Published Year : 2012 | This study involves the scenario like Illumination, pose, expression, ageing, occlusion and Low resolution for the evaluation of FR techniques. A handful amount of models are used for the testing including Gabor + ICA , Kernel associated Memory Mode (KAMM), Kullback-Leibler divergence (KLD)-based, local Gabor binary patterns (LGBP), Hybrid Colour and Frequency Features (CFF), Gabor Image Representation (GIR) and 3D Morphable Model (3DMM) . Comparative analysis of the graph show that Gabor + ICA, Kernel associated Memory Model. Their study also reports the computational efficiency of each algorithm studied in their research | This study is old and does not compare modern FR techniques based on Machine Learning and Deep Learning yielding state of the art results. |

End of Table

## 2.8 Research Contribution

With an aim of bridging the gap discussed in previous section, this research thesis has made the following cardinal and novel contributions:

- A detailed and comprehensive comparative analysis is done using the five top of the line deep learning models i.e FaceNET, VGGFace2, SphereFace, CosFace and ArcFace yielding state of the art results on benchmark datasets tested in five different facial images degradation i.e Ageing, Pose, Resolution, Cross Spectral and Ethnicity. A total of 7 benchmark databases are used in this regard.
- A novel cross spectral face recognition (CFR) technique is introduced, by which deep learning models trained on RGB images are capable of recognizing the Thermal images without fine-tuning with acceptable results. However the results became more reliable after fine tuning.

## 2.9 Summary

This chapter discusses the work done so far on face recognition technologies. There is a brief introduction to the topic of face detection and then a discussion of the methods applied from classical to deep learning including details of MTCNN [14] and RetinaFace [20]. An overview of the most widely used FR methods i.e. parts-based, feature-based, holistic, and hybrid techniques followed by an introduction to deep learning technologies that can be applied to computer vision problems and also deals with the most recently used deep face recognition methods. A detailed review of the comparisons of FR systems done so far was also explained in this Chapter, followed by the Problem statement/Gap Analysis and Research Contribution of this Thesis.

# Chapter 3

# Face Recognition Models under Test

## 3.1 Introduction

In order to compare and test the performance of face recognition algorithms, five state-of-the-art deep learning based FR models were brought under test, with the latest model being published in 2021 and bottom most in 2015. In this Chapter we will discuss each one in detail.

## 3.2 FaceNet

In 2015, Google researchers developed a FR system called FaceNet [73]. It exhibit excellent performance on a variety of benchmark FR databases i.e LFW [65] and YTF [105]. They developed a technique that utilises DL backbone models like ZF-Net [106] and Inception [107] to produce more accurate representations of facial images. Then, as a loss function, it employed a technique called triplet loss to train this architecture.

FaceNet's architecture is based on end-to-end learning as depicted in Figure 3.1. Its underlying architecture is either ZF-Net or Inception. Additionally, it incorporates many 1x1 convolution in order to reduce number of parameters.

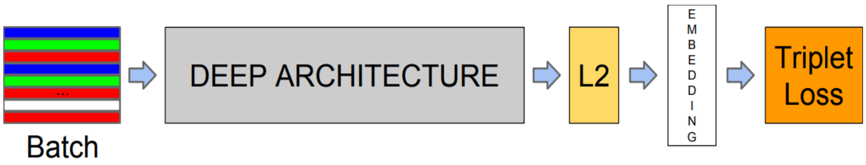


FIGURE 3.1: FaceNet model architecture [73]

These DL based models generate embedding(unique representation) of a facial image $f(x)$ that has been L2 normalised. The loss function is then used to generate these embeddings. The primary objective of loss function is to minimise a distance (i.e euclidean or square) in-between two facial images embeddings that are identical in terms of image condition and facial posture, while increasing the distance (squared) between images of different identities.In this regard the Triplet Loss [108] came into being with the objective of imposing a margin in between faces representing two different individuals.

### 3.2.1 Triplet Loss

Embedding generated by the DL model is given as $f(x)$, for instance, $x \in \mathbb{R}$. This embedding is actually of a vector of shape 128-dimensional that has been normalised so that :

$$\|f(x)\|_2^2 = 1$$

The objective is to make sure that the distance between anchor image and the positive image (image of the same individual) is lesser as compared to the acho a negative image(image of another person) as depicted in Figure 3.2, so that:

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2 \tag{3.1}$$

$$\forall \left(f(x_i^a), f(x_i^p), f(x_i^n)\right) \in \top \tag{3.2}$$

Here $\alpha$ represents a margin term, enforced to differentiate positive and negative pairs and $\top$ are the image space. Hence the loss function will be represented as following :

$$L = \sum_{i}^{N} \left[ \| f\left(x_i^a\right) - f\left(x_i^p\right) \|_2^2 - \| f\left(x_i^a\right) - f\left(x_i^n\right) \|_2^2 + \alpha \right] \tag{3.3}$$

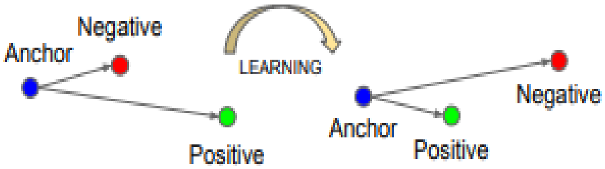


FIGURE 3.2: The Triplet Loss Learning [73]

The Equation 3.6 would not be helpful if triplets easily satisfy it, so it is vital to have triplets that violate it. This means for given $x_i^a$ triplets should be selected as $\| f\left(x_i^a\right) - f\left(x_i^p\right) \|_2^2$ is maximum and $\| f\left(x_i^a\right) - f\left(x_i^n\right) \|_2^2$ is minimum. The generation of triplets using the entire training set is computationally expensive. Two methods can be employed to do so:

- On each step, compute the min & max for subset of a data based on the previous checkpoints.
- Selecting hard positive $(x_i^p)$ and hard negative $(x_i^p)$ by using min & max on a mini batch.

Here authors claims to achieves the classification accuracy of 98.87%( 0.15% of standard error) on unrestricted protocol of LFW database. However, the model used in this research is provided by David Sandberg on GitHub [109] with Inception ResNet v1 as backbone. It is trained in VGGFace2 large scale database consisting of 3.3M facial images of 9000 different classes. This model yields the accuracy of 99.65% on LFW dataset.

## 3.3 VGGFace2

VGGFace2 [87] itself is a dataset proposed by the members of Visual Geometry Group at Oxford University to train sophisticated and modern the CNN's used for face recognition. This dataset comprise of 3.31 million images from 9131 people, with average of 362.6 images per subject. All of the images were retrieved from Google Image Search and vary significantly in terms of pose, age, lighting, ethnic origin, and profession (i.e. actors, athletes, politicians). Entire database is divided into a training set (consisting of 8631 unique identities) and a test set (including 500 identities). Name VGGFace2, on the other hand, has become associated with the pre-trained face recognition models trained on this dataset. The legacy of VGGFace Models dates back in 2015 [110], when the first model was presented with famous VGG-16 architecture as backbone, later on the model is trained using ResNet-50 architecture and SENet [111] architecture in 2017.

Originally VGGFace architecture is made up of 13 convolutional layers, they each have its own unique hyper parameter values. Each convolutional layer has 15 rectified linear units (ReLUs) [112], along with maxpooling layers. The third and fourth layers on top of this are the fully connected (FC) layers, designated FC6 and FC7, respectively. FC8 has 2622 channels, while FC6 and FC7 have 4096 channels each as seen in Figure 3.3. The final layer is a softmax layer that is used for classification of images in accordance with their class.

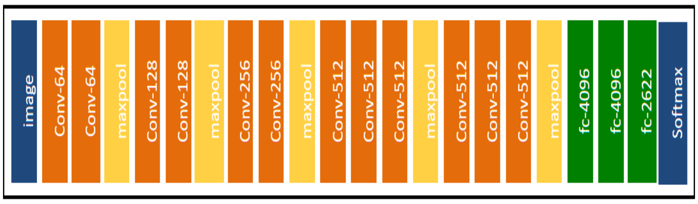


FIGURE 3.3: VGGFace Architecture [87]

The model utilized for this research is pre-trained using MS1M [113] database and later finetuning was done using VGGFace2 database with SENet architecture. Next section will discuss the SENet architecture in detail.

### 3.3.1 Squeeze-and-Excitation Networks (SENets)

Squeeze-and-Excitation Networks (SENets) [111] provide a low-cost building block for CNNs that enhances interdependence among channels. They were employed in last year's ImageNet competition and contributed to a 25% improvement over previous year's outcome. Apart from providing a significant performance improvement, they can be simply integrated into existing architectures.

The Convolutional Neural Network (CNN) extracts hierarchical information from the images using the convolution operator. The bottom layer detects lines, edges, and other discrete objects, but the top layer detects whole objects such as a human face, cat, or dog. All of this is accomplished by integrating spatial and channel-specific information at each layer. Convolution builds a feature map with a variable number of channels, where each channel is treated equally. This implies that each channel is equally significant, which may not be the ideal approach. The Squeeze and Excitation attention technique adds a scaling parameter to each channel. The Squeeze and Excitation functions essentially as a content-aware technique that adaptively reweights each channel that can be seen in Figure 3.4.

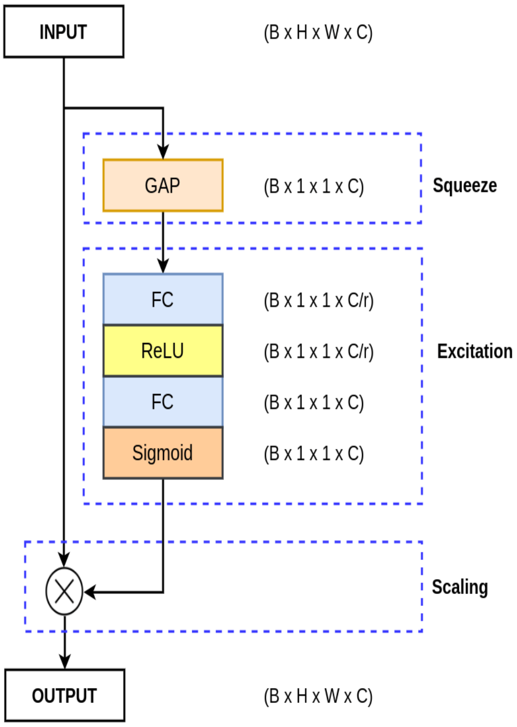


FIGURE 3.4: A detailed diagram of the Squeeze and Excitation Network [111]

The squeeze function is mostly used to obtain global information from the feature map's channels. The Convolutional layer outputs the feature map, which is a $B \times H \times W \times C$ dimensional 4D tensor.

Here:

- B: denotes the batch size.
- H: denotes the elevation of each feature map.
- W: Width of each feature map.
- C: represents the channel count in the feature map.

As we know, convolution is a local operation, as it only sees a subset of the input image. As a result, it is critical to have a holistic knowledge of the feature map. As here we are coping with a four-dimensional tensor that has a large number of parameters means that one must deal with numerous parameters when the number of channels in a CNN rises dramatically. As a result, a method for reducing each feature channel to a single numeric value is required. This decomposition would result in a reduction in the number of parameters, which would result in a reduction in computing complexity. Pooling techniques are employed in modern convolutional neural networks to minimise the spatial dimensions of the feature maps. The two most often utilised pooling operations are as follows:

- Max Pooling: This operation extracts the maximum pixel value from a specified window.
- Average Pooling: This technique calculates the average pixel values for a specified timeframe.

The author conducts a series of experiments to check the performance of two different pooling operations: Global Max Pooling (GMP) and Global Average Pooling (GAP). Global Average Pooling (GAP) performs better than the Global Max Pooling (GMP). Thus in the squeeze operation, the Global Average Pooling (GAP) is used to reduce the $B \times H \times W \times C$ feature map to $B \times 1 \times 1 \times C$.

There are now only four dimensions on the feature map i.e $B \times 1 \times 1 \times C$, effectively to a single vector for each channel of size $H \times W$. A fully connected MLP having a bottlenecked shape is then employed for excitation operation. Each feature map channel is scaled adaptively using weights generated from the MLP.

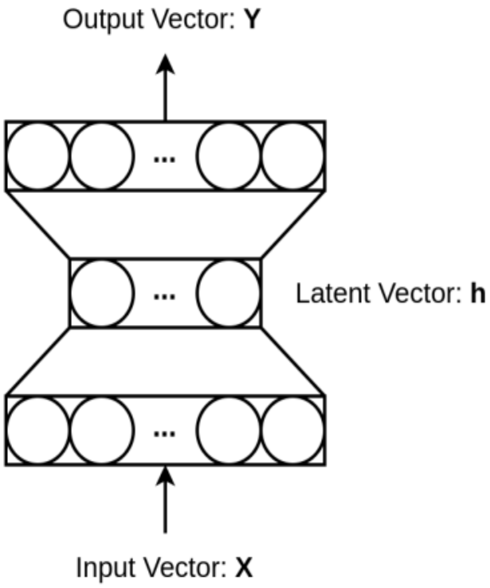


FIGURE 3.5: A Multilayer Perceptron (MLP) with bottleneck structure

The MLP as seen in Figure 3.5 is composed of three layers, the first one is tasked to minimise features by a factor of $r$. The dimensions of the feature maps comprised inside the layers are as follows: The input is of the form $B \times 1 \times 1$, which is reduced to the format $B$. Thus, the input layer contains $C$ neurons. The hidden layer significantly lowers the number of neurons in the network by a factor of $r$. Thus, the hidden layer contains $C/r$ neurons. Finally, the number of neurons in the output layer was increased to $C$. In general, the MLP accepts an input dimension of $B \times 1 \times 1 \times C$ and returns an output dimension of the same size.

The "excited" tensor is sent through the excitation procedure then sigmoid activation function is applied. It maps tensor values between 0 and 1 to tensor values. The output of the sigmoid activation function is then multiplied by the input feature map element by element. If the value is close to 0, the channel is considered less important, and thus the values of the feature channel are reduced; if the value is close to 1, the channel is considered important. To further analyse the scaling process, the author conducted an ablation study using non-linear activation functions in place of the sigmoid. Multiplication (element-wise) between the original feature map and output of sigmoid activation function takes place during the scaling process. The sigmoid activation function returns a value between 0 and 1, which is then multiplied by each channel. Therefore, consider multiplying a channel by a number close to 0. It will decrease the pixel values in that feature map, as these pixel values are considered irrelevant by the SE-block. When the channel is multiplied by a value close to 1, the pixel values are not suppressed nearly as much as in the prior situation. As a result, we can verify that the Squeeze and Excitation Networks effectively scale the information contained in each channel. It minimises irrelevant channel information while leaving the important channels mostly unchanged. Thus, at the end of the process, the feature map contains only the necessary information, thereby increasing the network's representational ability.

## 3.4 SphereFace

SphereFace [76] is joint effort of the researchers of Georgia Institute of Technology, Carnegie Mellon University and Sun Yat-Sen University in 2018. By not depending on a euclidean margin, the research greatly separates itself from earlier explored losses by employing the angular margin. This is seen to be extremely successful in tasks requiring facial recognition. The loss name provides information on how the features are transformed during the computation of the loss. The features are projected onto a manifold of hyperspheres.

SphereFace originates from the softmax [114] loss which is mostly employed in general classification tasks. It is defined as:

$$\mathcal{L}_S = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j-1}^{n} e^{W_j^T x_i + b_j}} \tag{3.4}$$

"where $x_i \in \mathbb{R}^d$ denotes the feature vector of the $i$-th sample belonging to the $y_i$-th class. $W_j \in \mathbb{R}^d$ is the $j$-th column of the weight matrix $W \in \mathbb{R}^d$ and $b$ corresponds to the bias term. $N$ is batch size and $n$ denoted corresponding class"

There is one significant disadvantage of softmax loss. It has no effect on classes cluster compactness. In other words, it does not ensure that samples within a category are comparable. As a result, the learnt features are not sufficiently discriminative for the open-set face recognition task. Another concern is the output weight matrix's dimension, which rises linearly in size as the number of identities in the training set increases. As a result, softmax loss is unsuitable for large-scale implementation. To derive SphereFace from softmax, we first incorporate the angle into the softmax equation using the dot product definition:

$$(a \cdot b = \|a\| \|b\| \cos\theta)$$

$$\begin{aligned}
\mathcal{L}_S &= -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j-1}^{n} e^{W_j^T x_i + b_j}} \\
&= -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\left\|W_{y_i}\right\| \|x_i\| \cos\left(\theta_{y_i,i}\right) + b_{y_i}}}{\sum_{j-1}^{n} e^{\|W_j\| \|x_i\| \cos(\theta_{j,i}) + b_j}}
\end{aligned} \tag{3.5}$$

> "where $\theta_{j,i}$ is the angle between vector $W_j$ and $x_i$. The remaining parameters are same as to those in the Softmax Equation 3.4. Next we normalize $\|W_j\| = 1$, $\forall j$ and set the bias term to 0." The modified loss function can be written as :

$$\mathcal{L}_{\text{modified}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\|x_i\| \cos\left(\theta_{y_i,i}\right)}}{\sum_{j-1}^{n} e^{\|x_i\| \cos(\theta_{j,i})}} \tag{3.6}$$

While it is feasible to train features using the modified loss function, but the resulting feature set would not be sufficiently discriminative. The researchers addressed this issue by incorporating an angular margin as represented in Equation 3.7:

$$\mathcal{L}_{ang} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\|x_i\| \cos\left(m\theta_{y_i,i}\right)}}{e^{\|x_i\| \cos\left(m\theta_{y_i,i}\right)} + \sum_{j \neq y_i} e^{\|x_i\| \cos(\theta_{j,i})}} \tag{3.7}$$

where $\theta_{y_i,i}$ lies in $\left[0, \frac{\pi}{m}\right]$. The decision boundary for a binary case is defined by:

$$\cos m\theta_1 = \cos\theta_2$$

“where $\theta_i$ is the angle between the feature and weight of class $i$. To make the loss in Equation 3.6 optimizable for CNNs the definition range of $\cos\left(\theta_{y_i}, i\right)$ is expanded. This is achieved by replacing the cosine term with monotonically decreasing angle function $\Psi\left(\theta_{y_i}, i\right)$ resulting in Equation 3.8:”

$$\mathcal{L}_{ang} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\|x_i\|\Psi\left(m\theta_{y_i,i}\right)}}{e^{\|x_i\|\Psi\left(m\theta_{y_i,i}\right)} + \sum_{j\neq y_i} e^{\|x_i\|\Psi(\theta_{j,i})}} \tag{3.8}$$

The angle function has the following definition:

$$\Psi\left(\theta_{y_i,i}\right) = (-1)^k \cos\left(\theta_{y_i,i}\right) - 2k \tag{3.9}$$

where $k \in [0, m1]$. The parameter $m \geq 1$ gives us control over the size of angular margin.

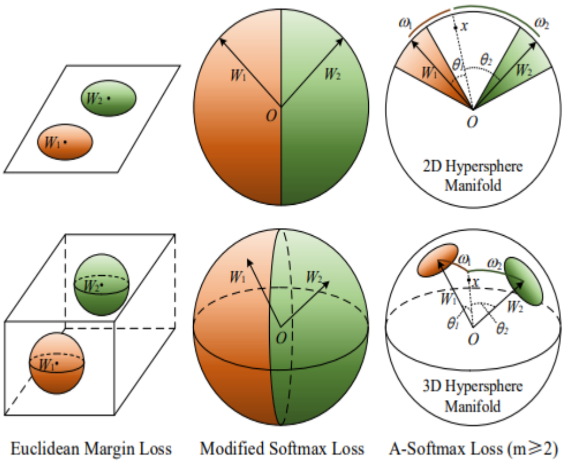
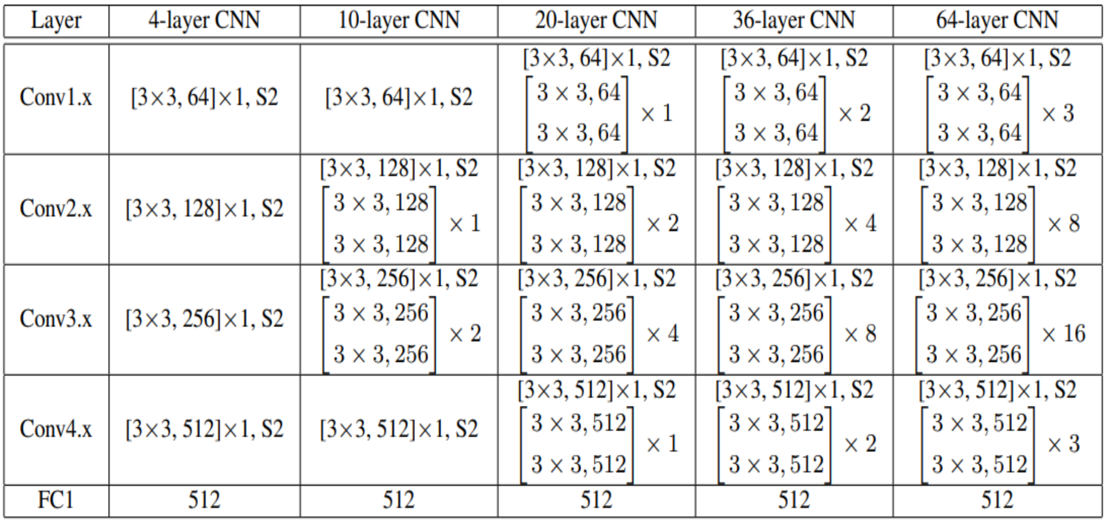


FIGURE 3.6: Comparison of various loss functions with angular loss [76]

The Figure 3.6 represents a geometry interpretection of Euclidean margin loss (e.g contrasive loss, triplet loss etc), modified and angular softmax loss. The first row represents a 2D feature constraint, second row represents a 3D feature constraint. The orange and green regions represents the discriminative constraints for class 1 and class 2.

Figure 3.7, illustrate the CNN architectures with varying convolutional layers. Units of convolution containing more than one layer are labeled as Conv1.x,



| Layer | 4-layer CNN | 10-layer CNN | 20-layer CNN | 36-layer CNN | 64-layer CNN |
|---|---|---|---|---|---|
| Conv1.x | [3×3, 64]×1, S2 | [3×3, 64]×1, S2 | [3×3, 64]×1, S2<br>$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times1$ | [3×3, 64]×1, S2<br>$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$ | [3×3, 64]×1, S2<br>$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$ |
| Conv2.x | [3×3, 128]×1, S2 | [3×3, 128]×1, S2<br>$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times1$ | [3×3, 128]×1, S2<br>$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$ | [3×3, 128]×1, S2<br>$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$ | [3×3, 128]×1, S2<br>$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times8$ |
| Conv3.x | [3×3, 256]×1, S2 | [3×3, 256]×1, S2<br>$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$ | [3×3, 256]×1, S2<br>$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times4$ | [3×3, 256]×1, S2<br>$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times8$ | [3×3, 256]×1, S2<br>$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times16$ |
| Conv4.x | [3×3, 512]×1, S2 | [3×3, 512]×1, S2 | [3×3, 512]×1, S2<br>$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times1$ | [3×3, 512]×1, S2<br>$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$ | [3×3, 512]×1, S2<br>$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$ |
| FC1 | 512 | 512 | 512 | 512 | 512 |

FIGURE 3.7: Different Architectures proposed by authors in SphereFace [76]

Conv2.x, and Conv3.x, while residual units are shown in double-column brackets. i.e. 4 cascaded convolution layers with 64 filters of size 3×3 is denoted by [3×3, 64]×4, S2 represents a stride of 2.

The SphereFace model utilized in this research is pretrained on CASIA-Webface [86] database which contains 494,414 facial images belonging to 10,575 different individuals with standard 64-layer CNN as backbone, presenting a verfication accuracy of 99.2% on LFW benchmark.

## 3.5 CosFace

The Large Margin Cosine Loss (LMCL) [79], also known as CosFace is another loss using margin with means of improving discrimination of softmax, restructures the traditional softmax loss by making weight as well as feature vectors L2 normalised to eliminate radial variations. Further maximising the decision margin in the angular space is then achieved by introducing $m$ as a cosine margin term. As an outcome, the lowest intra-class margin and the highest inter-class margin possible for reliable face verification is obtained.

As depicted in Figure 3.8, During the training stage, the discriminating facial features between various classes will be learnt with a large margin. During the

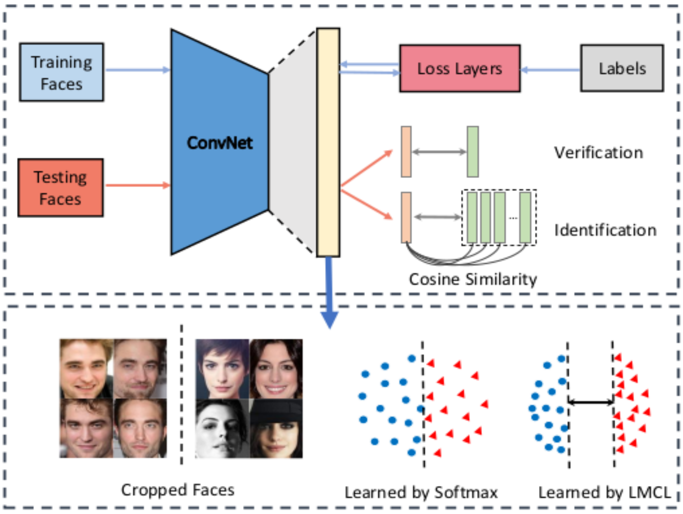


FIGURE 3.8: Overview of CosFace Framework as presented by authors [79]

testing stage, test data is fed to CosFace model, which extracts facial features and then utilizes them to calculate the cosine similarity aiding face verification and identification. High separability among classes can be visualized in the Figure 3.8.

By maximising posterior probability of a true class, softmax loss splits features of several classes. Softmax loss may be expressed as in Equation 3.4 inputs a feature vector $x$ and corresponding label $y$:

$$L_s = \frac{1}{N}\sum_{i=1}^{N} -\log p_i = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_{y_i}}}{\sum_{j=1}^{C} e^{f_j}} \tag{3.10}$$

here $p$ denoted posterior probability that $x$ is categorised correctly. $N$ represents total number of training samples, whereas $C$ denotes the total classes. $f$ is the activation fucntion used for fully connected layer with the weight vector as $W$. Keeping bias equal to zero for simplicity's sake. As a result, $f$ is defined as follows:

$$f_j = W_j^T x = \|W_j\| \, \|x\| \cos\theta_j \tag{3.11}$$

where $\theta$ denotes the angle created by W and $x$. According to the formula, a posterior probability is determined by both the norm and angle of vectors. In order to ensure proper feature learning, $\|W\|$ must be constant, which is why we set $\|W\| = 1$ using L2 normalisation. As we compare the two face feature vectors using cosine similarity during the testing, we can conclude that the feature vector's norm has no effect on the scoring function. Thus, we may set $\|x\| = s$ during training. As a result, the posterior probability is only determined by the cosine of the angle, and so the loss may be expressed as:

$$L_{ns} = \frac{1}{N} \sum_i -\log \frac{e^{s\cos(\theta_{y_i,i})}}{\sum_j e^{s\cos(\theta_{j,i})}} \tag{3.12}$$

In this case, we set the $\|x\|$ to $s$, so the model only learns features that can be separated in the angular space (NSL). However, NSL (Normalized Version of Softmax Loss) is insufficient since it only focuses on proper classification. To resolve the issue, the loss function is modified to include a cosine margin. Considering an example of binary classification, Hence LMCL is formulated as :

$$L_{lmc} = \frac{1}{N} \sum_i -\log \frac{e^{s\left(\cos\left(\theta_{y_i,i}\right)-m\right)}}{e^{s\left(\cos\left(\theta_{y_i,i}\right)-m\right)} + \sum_{j\neq y_i} e^{s\cos(\theta_{j,i})}} \tag{3.13}$$

Here $N$ represents the number of training samples, $x_i$ is a feature vector with label $y_i$, $W_j$ is weight vector and $\theta_j$ represents the angle among $W_j$ and $x_i$.

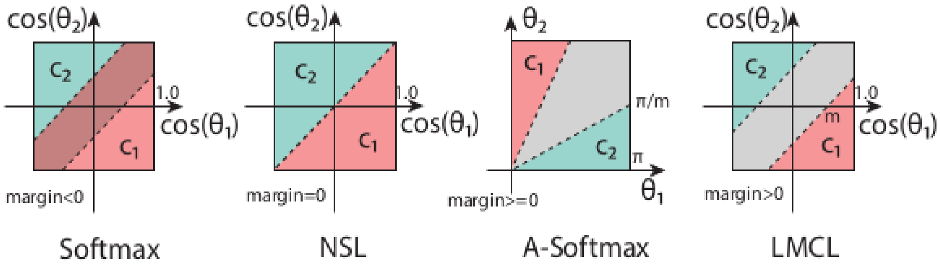


FIGURE 3.9: Comparison of different decision Margins [79]

As illustrated in Figure 3.9, different decision margins are used for binary-class scenarios with different loss functions. Grey areas represent decision margins and dashed lines are decision boundaries.

The softmax loss creates a decision boundary as $\|W_1 \times \cos(\theta_1)\| = \|W_2 \times \cos(\theta_2)\|$ Because the decision boundary is dependent on both the magnitude and angle of the weight vectors, the decision margin overlaps in the cosine space. The weight vector is normalised to have magnitude 1 by NSL, and hence the decision boundary is defined as $\cos(\theta_1) = \cos(\theta_2)$. As seen in the Figure 3.9, by reducing radial variation, it is capable of completely classifying samples with $margin = 0$. However, it is not resistant to noise.

A-Softmax (Angular softmax) minimizes softmax loss by inserting an additional margin, resulting in the following decision boundary :

$$\begin{aligned} C_1 &: \cos\left(m\theta_1\right) \geq \cos\left(\theta_2\right) \\ C_2 &: \cos\left(m\theta_2\right) \geq \cos\left(\theta_1\right) \end{aligned} \tag{3.14}$$

The third plot in Figure 3.9 illustrates the decision area, with the grey region representing the decision margin. Because CosFace's decision boundary is not specified across angular space, the loss is easier to optimise than SphereFace's. Because of the non-monotonic property of cosine function, optimization in angular space is more complex. Another improvement over SphereFace is that not only the weight vector $W_j$, but also the feature vectors $x_i$, are normalised. As a result of the emphasis on the angle during training, the intraclass variability of the learnt features is significantly smaller. The decision margin was specified in a cosine-space rather than anglular-space by LMCL as follows:

$$\begin{aligned} C_1 &: \cos\left(\theta_1\right) \geq \cos\left(\theta_2\right) + m \\ C_2 &: \cos\left(\theta_2\right) \geq \cos\left(\theta_1\right) + m \end{aligned} \tag{3.15}$$

In order to perform the large-margin classification $\cos\left(\theta_1\right)$ is maximized while $\cos\left(\theta_2\right)$ being minimized for $C_1$ (similarly for $C_2$) . It is possible to observe a clear margin in the produced distribution of the cosine of angle in Figure 3.9, which is the decision boundary of LMCL in the cosine space. This shows that LMCL has a greater degree of robustness than NSL. Both the weight vector and feature vector are normalised to obtain the formulation of cosine loss and to eliminate radial variation. As an outcome, feature vectors are dispersed on the hypersphere, with the radius controlled by the scaling parameter $s$. Without feature normalisation, the original softmax loss automatically learns both the Euclidean norm (L2 -norm) of feature vectors and the angle's cosine value. The L2 -norm is adaptively learnt with the goal of reducing the total loss, which results in the relatively weak cosine constraint. On the other hand, LMCL demands that the whole collection of feature vectors to have the same L2 -norm, such that learning is based exclusively on cosine values. On the hypersphere's surface, feature vectors belonging to the same class are grouped together, while those belonging to other classes are separated. The total number of classes as $C$, given the normalised learned feature vector $x$ and the unit weight vector $W$. Assume that the learnt feature vectors each reside on the hypersphere's surface and are centred on the corresponding weight vector. Let $P_W$ be the predicted lowest posterior probability of the class centre (i.e., $W$), which serves as the lower bound for $s$ :

$$s \geq \frac{C-1}{C} \log \frac{(C-1)P_W}{1-P_W} \tag{3.16}$$

On the basis of this constraint, we may conclude that $s$ must be constantly increased If we anticipate that $P(w)$ is optimal for classifying a set of given classes. The ideal $s$ must be larger in order to accommodate more classes, as the growing number of classes makes classification more difficult. Thus, for features with a low intra-class distance, but a high inter-class distance, a hypersphere with a large radius $s$ is required.

Choosing the ideal value of $m$ may result in more promising learning of highly discriminative face features. Bigger is a better choice for $m \in \left[0, \frac{C}{C-1}\right)$ as it will improve learning of highly discriminative features. As all feature vectors are centred on the associated class's weight vector. Indeed, Too large $m$ can causes the model failure to converge, since the cosine constraint becomes more stringent and difficult to satisfy (i.e. $\cos\theta_1 - m > \cos\theta_2$ or $\cos\theta_2 - m > \cos\theta_1$ for binary classification). Additionally, the cosine constraint with an excessively high m makes the training process more vulnerable to noisy input. At some point, due to its inability to converge, increasing $m$ value begins to degrade overall performance.

The CosFace model used for this research is pre-trained on CASIA WebFace Database [86] with sphere 64 layer CNN as backbone and reports the accuracy of 99.23% on LFW Database.

## 3.6 ArcFACE

ArcFace [81] is a freely released research that produced record breaking results on LFW database in 2018. The majority of the concepts underlying ArcFace were previously discussed in the Section 3.4 SphereFace and Section 3.5 CosFace. These concepts are normalisation of class weights and feature vectors, as well as the addition of margin term $m$ in the loss function equation. These two concepts reduce intra-class variance in angular space, resulting in a model with enhanced discriminative ability for facial recognition tasks.

There are two primary lines of research for training CNNs for face recognition: one uses softmax classifiers for training a multiple class classifier, as well as the other uses embeddings like the triplet loss to learn the embeddings. Each, however, has its disadvantages. Softmax loss requires more parameters as the classes increase in recognition problem. Similarly, the number of face triplets grows logically with the size of the dataset, resulting in a large number of iterations for the triplet loss. In ArcFace [81], by introducing an additive angular margin loss, the face recognition model's discriminative ability can be further improved and the training process is stabilised. The angle between the existing feature and the desired weight is calculated using the arc-cosine function. ArcFace optimises the geodesic distance margin directly due to the exact correspondence between angle and arc in the normalised hypersphere (which contains the face features). ArcFace, like SphereFace and CosFace, originates in the softmax loss Equation 3.4. The difference between the original and the upgraded version is five steps. The first four are identical to their counterparts in CosFace:

1. fix the bias $b_j = 0$.
2. transform the logit using the dot product definition $W_j^T x_i = \|W_j\| \, \|x_j\| \cos\theta_j$ ( $\theta$ is the angle between the weight $W_j$ and the feature $x_i$ ).
3. fix the individual weights $\|W_j\| = 1$ by $l_2$ normalization.
4. do the same for feature $x_i$ and re-scale it to a predetermined feature scale $s$.

The two normalization steps make the prediction depend only on the angle $\theta$. The embeddings are scattered on the hypersphere having radius $s$.

At start the loss function equation will be given as (Softmax):

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{3.17}$$

where $x$ indicates the $i - th$ sample's feature vector and $W$ and $b$ signify the weight and bias, respectively. There is no explicit optimization in Softmax loss for feature embedding to ensure that samples of the same class are more similar, while those of different classes have a great degree of diversity, Due to large intra-class appearance variations (i.e age gap and pose variation), such a performance gap has been observed for deep face recognition.

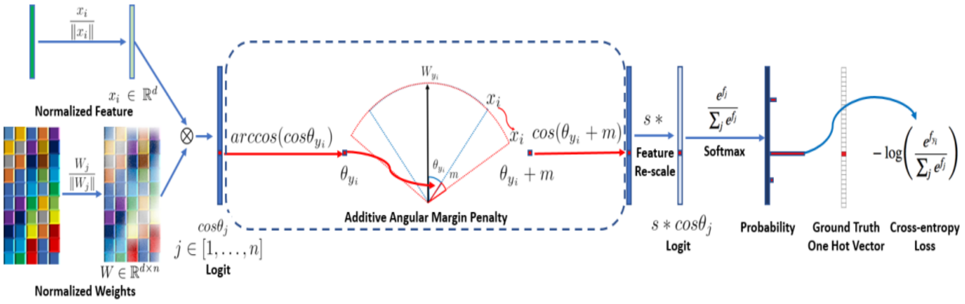


FIGURE 3.10: A CNN is trained by using the ArcFace loss function [81]

For simplicity, we set the bias to zero in the softmax loss and then transform the logit functions as follows:

$$W_j^T x_i = \|W_j\| \, \|x_i\| \cos\theta_j \tag{3.18}$$

where $\theta$ is the angle formed by the weight $W$ and the feature $x$. By applying the $L2$ normalization, the weight is normalised to one. Also, the feature is $L2$ normalised and rescaled to $s$.

The normalizing procedures enable predictions dependent only on the angle ($\theta$). The learnt embedding is distributed in the following order on a hypersphere of radius $s$:

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}} \tag{3.19}$$

Between weight and feature, an additive angular margin penalty m is introduced to increase intra-class compactness and inter-class discrepancy. Due to the fact that the suggested additive angular margin penalty equals the geodesic distance margin penalty in the normalised hypersphere, it is referred to as ArcFace. As a result, the final loss function is as follows:

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}} \tag{3.20}$$

To train the ARCFace model as proposed by authors in Figure 3.10 following algorithm is used:

1. After feature $x_i$ and weight $W$ normalisation, we get the $\cos\theta_j$ (logit) for each class as $W_j^T x_i$
2. We calculate the $\arccos\theta_{y_i}$ and get the angle between the feature $x_i$ and the ground truth weight $W_{y_i}$.
3. We add an angular margin penalty $m$ on the target (ground truth) angle $\theta_{y_i}$.
4. We calculate $\cos\left(\theta_{y_i}+m\right)$ and multiply all logits by the feature scale $s$.
5. As a result, the logits then contribute to the CE loss by applying softmax.

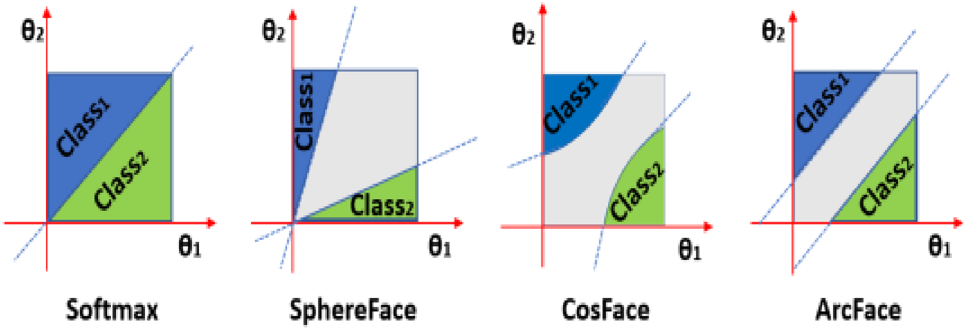


FIGURE 3.11: Comparison of angular decision Margins[81]

Figure 3.11 shows a comparison of classification boundaries in the case of binary classification. ArcFace throughout maintains a constant linear angular margin. SphereFace and CosFace, on the other hand, presents a nonlinear angular margin. The authors then used a set of parameters to compare various network architectures. Each of these networks have a similar architecture.

All of these models are trained on 112x112 input images in RGB domain after cropping, resizing, and normalising images. The first variation is in the output size, which may be set to 7x7 (denoted by a L at the network's commencement) or 3x3. Then, five distinct embedding options are evaluated. Following that, two distinct versions of the ResNet network are evaluated. The second form is generated by using an enhanced version of the residual units termed 'IR' (for improved residual). Finally, they examine a variety of alternative network topologies, including MobileNet [115], Inception-ResNet-V2 [116], DenseNet [117], Squeeze and excitation networks [111], and Dual Path Network [118]. According to their findings, the best result is obtained by utilising a ResNet with the L output, a BN-dropout-FC-BN layer after the convolution layer as an embedding setup, and IR residual units. Authors also specify that the second class of losses are superior because they include "discriminative constraints on a hypersphere manifold, which intrinsically fits the assumption that the human face is on a manifold." In comparison to its predecessors, the authors state that ArcFace, or additive angular margin, has a "superior geometrical interpretation" that enables the acquisition of "more discriminative deep features." The ARCFace pretrained model used in this research posses a backbone architecture of ResNet-100 [119], trained on MS1Mv2 Dataset [113] and using Apache MXNet [120] framework. The model yields an accuracy of 99.77% on LFW Dataset.

## 3.7 Summary

This chapter include the details of five deep learning models i.e. FaceNET, VGGFace 2, SphereFace, CosFace and ArcFace models are used for evaluation used for the evaluation. It also describe the training database, the network architecture and the loss function involved in the training of these models. The last three models utilize different modalities of angular loss function to attain the state of the art results on LFW benchmark database.

# Chapter 4

# Evaluation Datasets and Protocols

This chapter presents a details about the scenarios and the databases used to evaluate the performance of the selected deep learning based FR models. The scenarios selected are :

1. Aging Effect
2. Facial pose Effect
3. Resolution Effect
4. Cross spectral facial matching
5. Ethnicity Effect

## 4.1 Aging Effect

Aging has a noticeable influence on the automated face recognition process, increasing the FAR and FRR consequently degrading the performance. These rates might have a severe impact on operations in high-volume areas (e.g. airports), resulting in security concerns and unanticipated delays. Understanding the impact caused by aging on the performance of FR systems is critical, turning study on aging effects an interesting research topic. Face recognition is a traditional technique for authentication [121]. FR faces many challenges related to aging, which have drawn the attention of researchers [122], [123]. The development of various datasets and methodologies has aided in the study of ageing effects on automated face recognition.

These developments have enabled algorithms to cope with the impacts of ageing on the identification process while still improving overall performance [124]. Despite all of the advancements over the years, several factors still continue to impact the effectiveness and accuracy of the identification process as people age. A best face recognition algorithm is the one which must performs good at facial images of the corresponding individual captured at different ages. A performance evaluation of the selected FR models was conducted over aging effect by utilizing two databases .i.e AgeDB [125] and CALFW [126].

### 4.1.1 AgeDB

AgeDB [125] has 16,488 images of notable individuals from numerous fields of life, including politicians, writers, and actors/actresses. Each image is labelled along the information on the subject's identity, age, and gender. There are 568 unique subjects. Each individual gets an average of 29 images. The minimum and maximum ages are 1 and 101 year, respectively. Each subject has an average age of 50.3 years.

Due to the rising interest in age estimation 'in-the-wild' and the rise of new databases, to address this issue in recent years there is no manual collection of year-age information from the 'wild' . To address this gap in the literature, the authors provide the first manually gathered 'in-the-wild' age database, named AgeDB. AgeDB features images of a variety of subjects labeled with year-accurate age labels.

Figure 4.1 depicts the scatter plot of Age distribution of all the samples present in the AgeDB dataset. Where as Figure 4.2 shows various samples and their labels

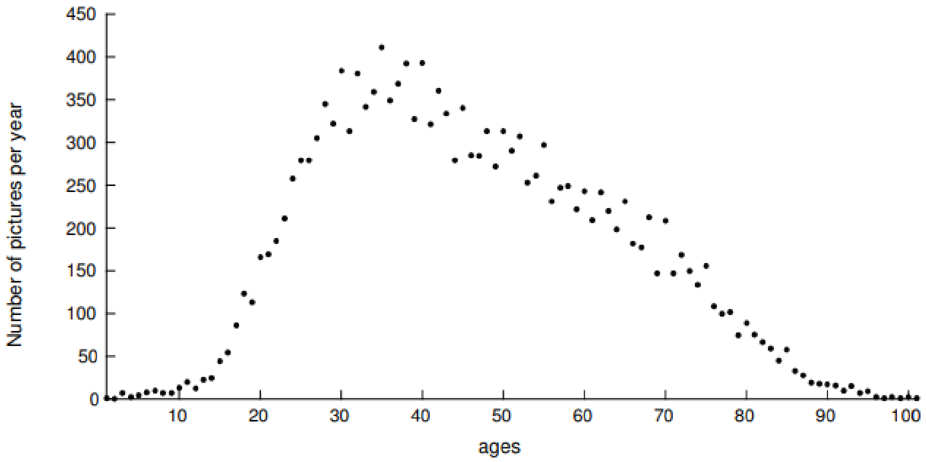
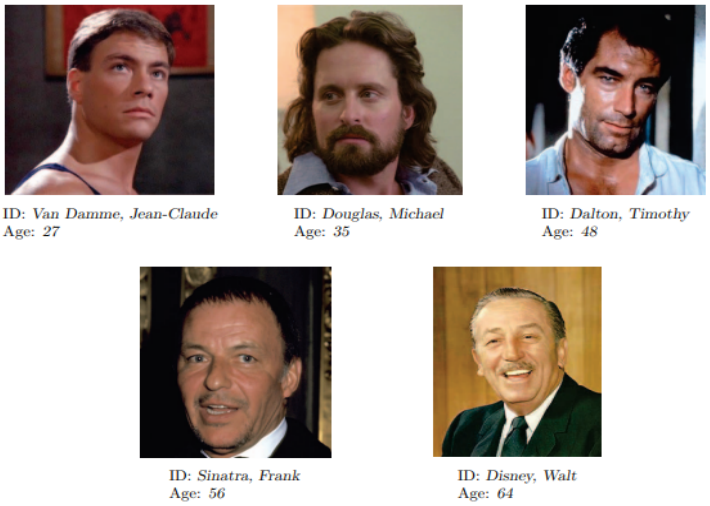


FIGURE 4.1: Scatter plot of Age distribution in AgeDB [125]



ID: *Van Damme, Jean-Claude*
Age: *27*

ID: *Douglas, Michael*
Age: *35*

ID: *Dalton, Timothy*
Age: *48*

ID: *Sinatra, Frank*
Age: *56*

ID: *Disney, Walt*
Age: *64*

FIGURE 4.2: Sample images in AgeDB [125]

presents in the AgeDB dataset. The fact that AgeDB is gathered manually ensures the correctness of the age labels in various ways:

1. AgeDB will be used to conduct age-invariant face verification studies i.e., Sensitivity of FR algorithm can be determined when the age difference between instances (images) of the same person rises. Due to the absence of noise in the age labels, AgeDB enables the fair assessment of various face recognition systems.

2. AgeDB can be utilised in "in-the-wild" age estimate studies. AgeDB may be used as a benchmark database for such tasks due to its accurate age labels.

3. AgeDB may be used in "in-the-wild" tests on face age progression, as it is a manually gathered database with a wide range of ages for each individual. This trait makes AgeDB extremely advantageous when it comes to training models for age progression investigations.

In this research, AgeDB is used to conduct age-invariance studies of the selected top of the line face recognition models.

#### 4.1.1.1 Evaluation Protocol

For evaluation of different models, we followed the same protocol as suggested by authors to report the accuracy.

The authors divided AgeDB into ten folds for each protocol, with each fold containing 300 intraclass and 300 intraclass pairs. The primary difference between the protocols is that each protocol has a specific, pre-defined age difference between the faces of each pair, i.e., 5, 10, 20, and 30 years. We are using the AgeDB-30 protocol because it contains age gaps of more than 30 years as seen in Figure 4.3 and is the most commonly reported and challenging one by AgeDB.

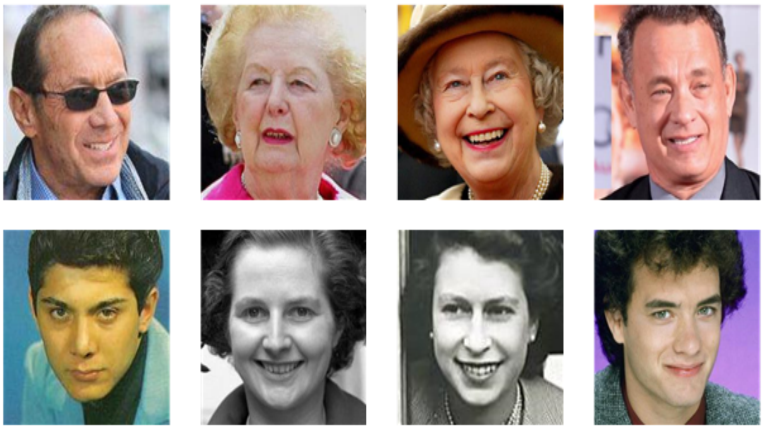


FIGURE 4.3: Image pairs with over 30 years of age gap in AgeDB [125]

As depicted in Figure 4.4 and Equation 4.1, after performing 10-fold cross validation, accuracy was computed after each fold and the average accuracy, is reported in this case.

$$E = \frac{1}{K} \sum_{i=1}^{K} E_i \tag{4.1}$$

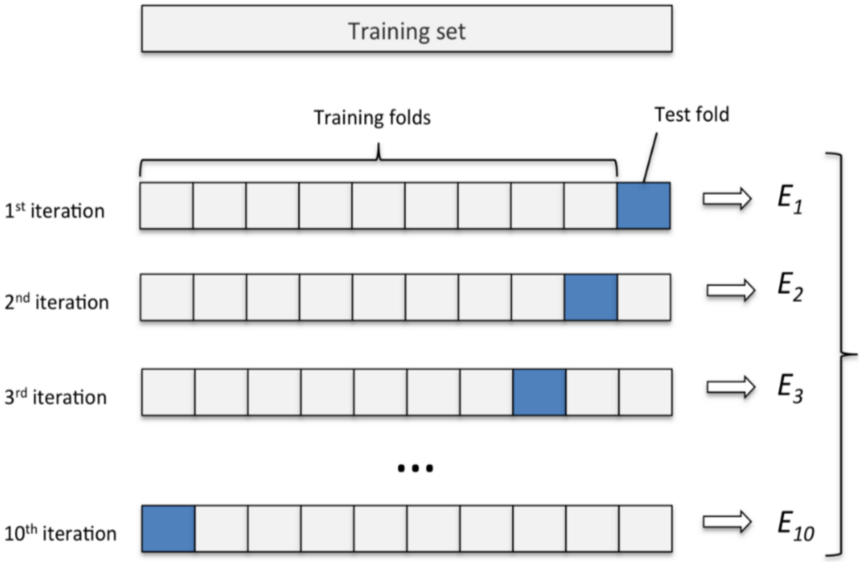


FIGURE 4.4: N-fold cross validation topology

### 4.1.2 CALFW

CALFW [126] is another benchmark dataset to evaluate the face recognition models under aging effect. CALFW (Cross-age LFW) originates from the famous LFW(Label faces in Wild) [65] Database which has been widely used as benchmark to study face verification. It is created by using images taken as part of the Berkeley Faces in the Wild project [127], [128]. The images in the project were gathered from Yahoo News between 2002 and 2003 and were taken in natural settings with a variety of settings, poses, expressions, and lighting. These images were popular for research purposes, but due to the presence of more than 10% noisy labels and a high number of duplicates, they could not be used as a benchmark. As a result, the dataset was cleaned manually,new protocols were developed, and released the dataset termed as 'Labeled Faces in the Wild'. The LFW database contains 13,233 face images of 5,749 individuals and two views of LFW for experiments: view 1 for development purposes and view 2 for fair comparison. For cross validation in view 2, the dataset was divided into ten non-repeating subsets of image pairs. Each subset comprises 300 pairs of positive images (images of the same person) and 300 pairs of negative images (images from different people).

When the database is utilised exclusively for testing, all pairings (3000 positive and 3000 negative) are included to acquire performance results.

The new database, called Cross-Age LFW (CALFW), was compiled through crowdsourcing attempts to collect images of people wearing LFW with the largest age gap possible on the Internet, in order to supplement the original LFW with age intra-class diversity. Following the search, an age estimation technique [129] is used to determine the ages of all selected images, and the pairs with the highest age disparities are chosen as positive pairs in View 2. The comparison of the same individual in LFW and CALFW is given in Figure 4.5, and as can be seen from the image, the ageing process is more visible in CALFW.

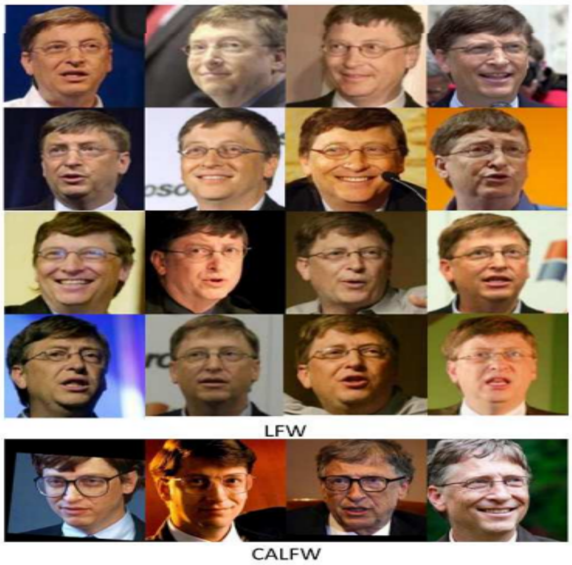


FIGURE 4.5: Image pairs with significant age gap in LFW and CALFW [126]

Figure 4.6 shows the improved age distribution in CALFW in comparison with LFW Database.

#### 4.1.2.1 Evaluation Protocol

CALFW dataset was partitioned into ten distinct folds using the same identities as the LFW ten folds.

The CALFW dataset comprises 4,025 persons, each with two, three, or four images. To determine the age of each image, we utilise Dex [129], the winner of

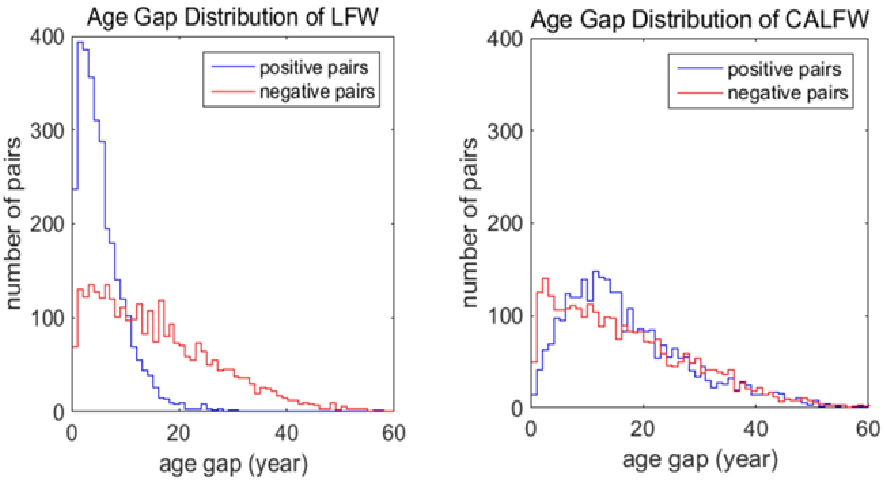


FIGURE 4.6: Image pairs with over 30 years of age gap in AgeDB [125]

the ChaLearn LAP 2015 [130] age estimation competition, and the names of the images are as : $name0001.jpg, name0002.jpg,$

the number ”0001”, ”0002” reflects the rank of age estimation result. ”0001” is the youngest image and ”0002” is the oldest image of a subject.

We utilized *pairs_CALFW.txt* from the testing section of CALFW, which contains 10 sets of 300 matched and 300 mismatched pairs respectively. So 12,000 pairs in total, half of which are matched, the other half of which are mismatched. We set thresholds range from 0 to 1, gap 0.001 for example, which produces 1,000 thresholds.

Calculated the distance $d$ over all pairs. For a threshold $t$, if $d \leq t$, pairs are predicted as matched. Otherwise, pairs are predicted as mismatched.

For each threshold t: For each matched pair,

- If it is predicted as matched, TP+=1;
- If it is predicted as mismatched, FN+=1.

For each mismatched pair,

- If it is predicted as matched, FP+=1;

- If it is predicted as mismatched, TN+=1.

Verification Accuracy is computed as Follows :

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

## 4.2 Pose Variation

One of the key issues of the face recognition system is the distribution of different facial poses. To match the profile face with the gallery face, frontal face reconstruction is required [131]. This reconstruction is essential because a database image with a frontal view and a non-frontal profile face might produce incorrect results. Various ways described by researchers to transform the non-frontal face to the frontal face may improve recognition accuracy [131], [132]. Researchers in the suggested methodologies [133], [134] describing how pose variation significantly impairs the algorithm's performance. For analyzing the performance of models described in Chapter 3 two benchmark datasets i.e CPLFW [126] and CFP-FP [135] are used to evaluate the performance.

### 4.2.1 CPLFW

CPLFW [126] is a database derived from LFW benchmark database, by breaking down two of its limitations factors. Firstly the authors look for images with a lot of pose changes and use identities/labels to find positive matches.

Second, they choose negative pairs consisting of people of the same gender and race i.e Identities that do not match vary exclusively in identity. CPLFW database was created via crowd sourcing efforts. Figure 4.7 highlights the pose effect present in CPLFW as compared to LFW dataset.

The following are the three motivations for the development of the CPLFW benchmark:

1. Creating a more challenging and unbiased database in order to perform a real world evaluation of FR algorithms so that their efficacy and efficiency can be completely justified.

2. Fostering research on cross-position face verification in unconstrained situations while continuing extensive research on LFW with a more realistic view on pose intra-class variance. The CPLFW challenge promotes pose difference to increase intra-class variation. Furthermore, negative pairs are purposefully chosen to prevent opposite genders or races. CPLFW takes into account both the substantial intra-class variance and the low inter-class variance at the same time.

3. Keeping the data size constant, the face verification protocol that offers a 'same/different' benchmark, and the same identities in LFW, so that CPLFW can be simply applied to assess the performance of face verification.

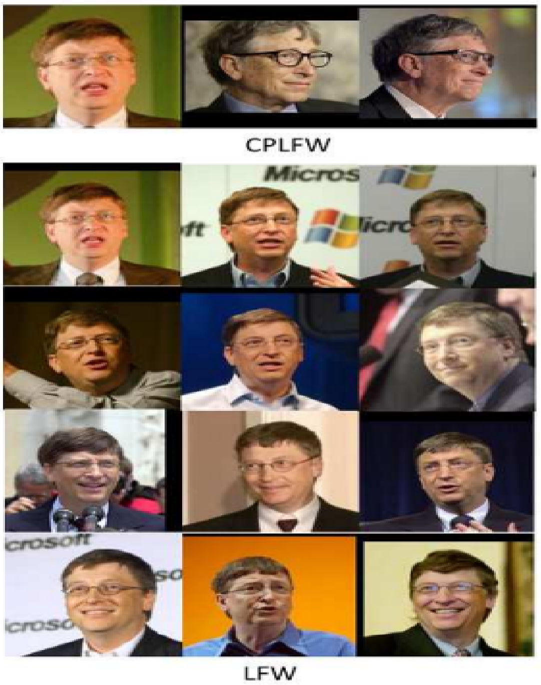


FIGURE 4.7: Image pairs with different poses in LFW [65] and CPLFW [126]

The pose difference represented as yaw angle of facial images can be seen more evenly distributed in CPLFW, according to the Figure 4.8. Furthermore, the difference of facial poses among the majority of positive image pairs in LFW is

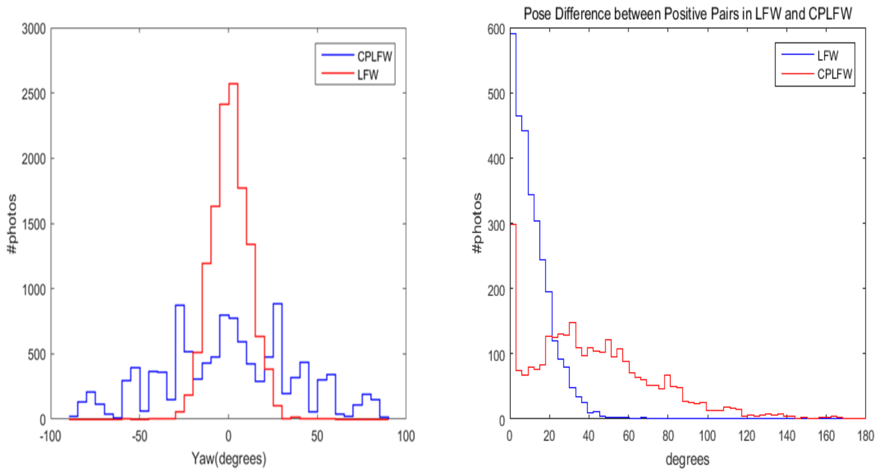


FIGURE 4.8: Pose variation among images present in CPLFW [126]

less than 40 degrees, but the difference is higher in CPLFW's positive image pairs. It confirms that there is intra-class variance in the dataset.

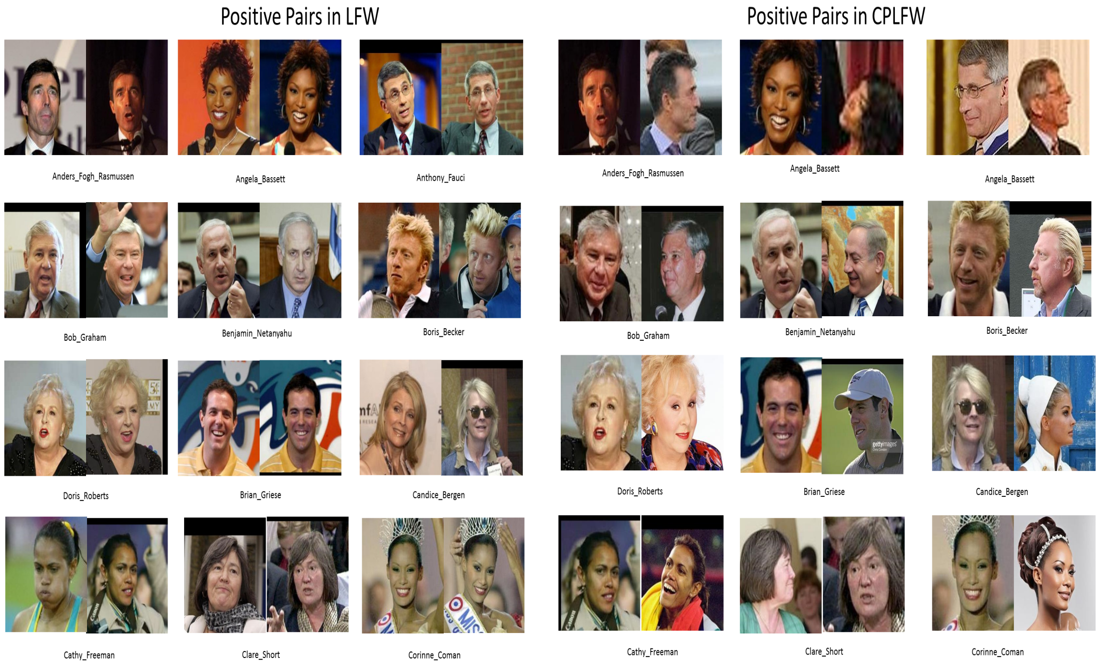


FIGURE 4.9: Comparison of pose of positive pairs of LFW and CPLFW dataset [126]

In comparison to LFW, CPLFW contains positive pairs that represent obvious pose differences, as shown in Figure 4.9. Although there are no changes in gender or race present among negative pairs in CPLFW, this greatly reduces the impact of attribute differences between positive and negative pairs.

#### 4.2.1.1 Evaluation Protocol

Evaluation protocol used to evaluate the face recognition models are same as defined by authors. Using the same identities found in the 10 folds of the LFW, the CPLFW dataset has also been divided into 10 folds. The dataset includes two or three images per person with the following naming convention : $name0001.jpg$, $name0002.jpg$

Authors have selected the positive pairs randomly. They select negative pairs with people who have the same gender and race as each other, so that there are no attribute differences between positive pairs (people who share the same gender and race) and negative pairs (people who are randomly matched by gender and race). The accuracy is computed in the same fashion as explained in Section 4.1.2.

### 4.2.2 CFP-FP

A data set for CFP-FP (Celebrities in Frontal-Profile) [135] in the Wild contains unconstrained images of celebrities in both frontal and profile (side) poses. Based on face verification, the experimental protocol was developed. By selecting a fixed number of frontal and profile images of each subject, the authors were able to obtain a balanced dataset. This data set is made available open-source for research and development. Images of 500 people are included, with 10 frontal and 4 profile shots. They define "frontal" as an image that shows both sides of the face almost equally on the image, and "profile" as an image that shows one eye clearly but less than half the other eye.

Essentially, these definitions mean: There must be less than a 10 degree variation in yaw for the 'frontal' and greater than a 60 degree variation for the 'profile'. Similar to that of LFW, There are 10 splits within the authors' proposed evaluation files, each containing 350 same pairs and 350 different pairs to verify the faces. The suggested protocols are tested on Frontal-Frontal and Frontal-Profile matching.

First row of Figure 4.10 shows the frontal images from the database whereas the second row represents the respective profile images.

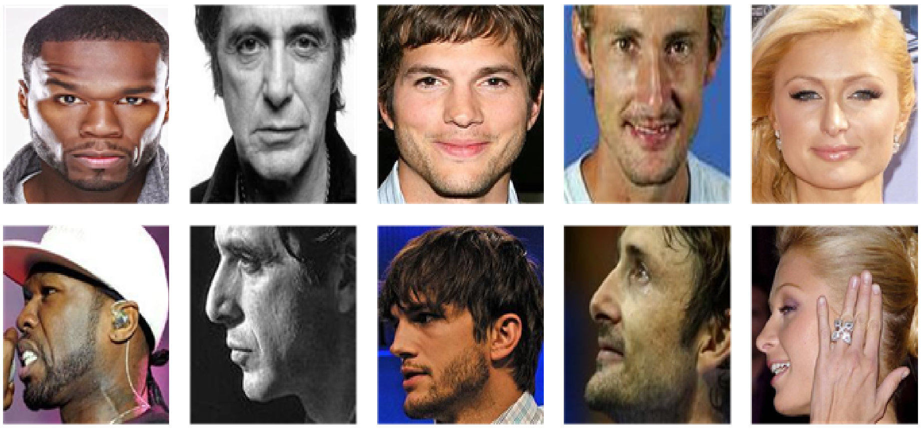


FIGURE 4.10: Frontal and Profile Images from CFP-FP database [135]

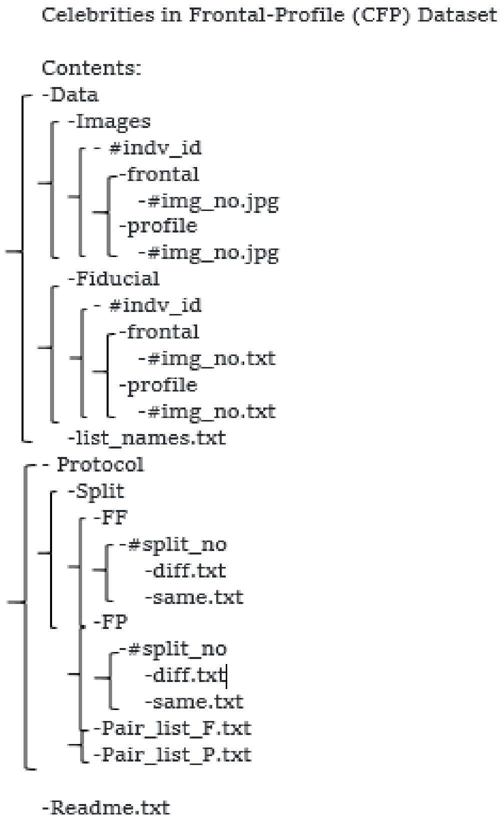
#### 4.2.2.1 Evaluation Protocol

In order to evaluate the models under poses variations, most challenging part of CFP dataset, i.e CFP-FP (Frontal to Profile) matching was performed.

Figure 4.11 shows the database structure of CFP database, Inside there are two main folders as *Data* and *Protocol*, The *Data* folder contains *Images* sub-folder which contains *Individualsidentitynumber* folders and *frontal* and *profile* sub folders with 10 and 4 images of each 500 individuals respectively. *Data* folder also contains folder named as *Fiducial* which contains Frontal and Profile fiducials (30 points) of each individual. *list_names.txt* contains the names of 500 individuals in order.

*Protoco*l folder contains pair information for Frontal-Frontal Verification and Frontal-Profile Verification. It also contains splits folders for all 10 fold verification for Frontal-Frontal. *same.txt* contains 350 same pairs and *diff.txt* retains 350 different pairs, similar is the case with Frontal to Profile matching folder termed as *FP* in Figure 4.11. *Pair_list_F.txt* & *Pair_list_P.txt* contains 5000 Frontal and 2000 Profile images with numbers ranging from 1 - 5000 for Frontal and 1 - 2000 for Profiles.

The associated location of images for this number can be obtained using these *.txt* files. For this research the most challenging scenario CFP-FP (frontal to profile) matching was done using standard 10-folds cross validation data provided by the authors to report the accuracy.



```
Celebrities in Frontal-Profile (CFP) Dataset

Contents:
-Data
  -Images
    - #indv_id
      -frontal
        -#img_no.jpg
      -profile
        -#img_no.jpg
  -Fiducial
    - #indv_id
      -frontal
        -#img_no.txt
      -profile
        -#img_no.txt
  -list_names.txt
- Protocol
  -Split
    -FF
      -#split_no
        -diff.txt
        -same.txt
    -FP
      -#split_no
        -diff.txt
        -same.txt
    -Pair_list_F.txt
    -Pair_list_P.txt

-Readme.txt
```

FIGURE 4.11: Database structure of CFP-FP database

## 4.3 Resolution Effect

As the number of surveillance cameras rises (particularly in metropolitan areas), the footage they acquire will need to be automatically processed. However, such videos are typically shot with significant standoffs, challenging lighting conditions, and a variety of angles of view. Faces in these images are often small, thus the resolution is limited, as seen in Figure 4.12. For a face recognition scenario it is difficult to compare the low-resolution test image to the high-resolution gallery image.

Although there is no widely accepted single criterion for classifying a face in an image as low-resolution, multiple studies have found that face images with a tight bounding box less than 32×32 pixels provide considerable accuracy problems to face recognition systems in both human and computer vision.

In order to evaluate the performance of selected face recognition algorithms, SCface dataset is used.

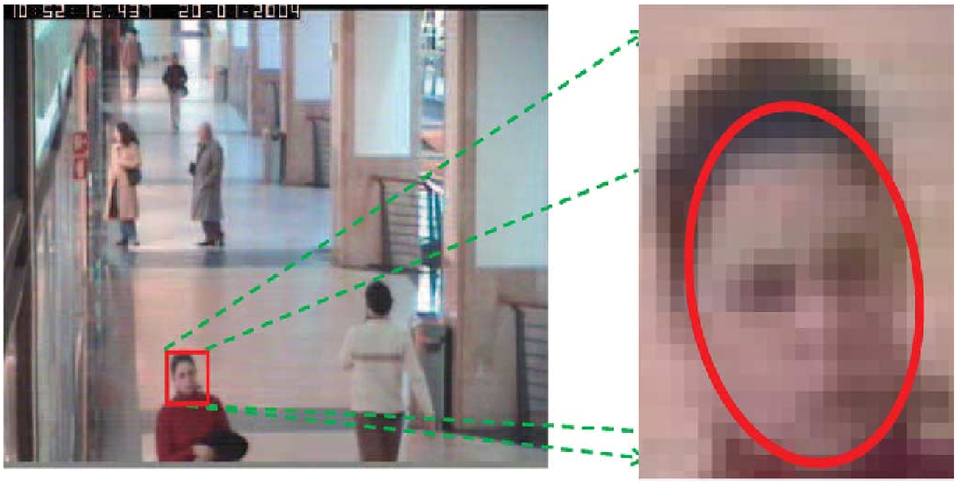


FIGURE 4.12: Low resolution face in a Surveillance video [136]

### 4.3.1 SCface

The SCface [136] database was created primarily to test FR algorithms in realistic scenarios. One might readily conceive a scenario in which a person should be recognised by comparing to a low quality still photo obtained from a video surveillance system with high quality mugshot images in such a configuration. The authors opted to employ commercially available surveillance cameras of varied quality to construct a realistic arrangement. Images in other currently accessible databases are typically captured with the same camera and without the use of proper, commercially available surveillance equipment. IR images were also added in the database since two of the surveillance cameras capture both visible spectrum and IR night vision photos.

The video communications laboratory at the Faculty of Electrical Engineering and Computing at the University of Zagreb, Croatia, was used to capture facial images. Six surveillance cameras, a professional digital video surveillance recorder, a professional high-quality picture camera, and a computer comprised the capture equipment. The authors utilised a high-quality photo camera to capture the mug shot images.

Then they employed five distinct (commercially accessible) surveillance camera types to get surveillance camera images, and a separate surveillance camera was used to capture IR mug photos. The only source of illumination while capturing these images was the natural light that entered via one of the windows. Two (of five) surveillance cameras were also capable of recording in the infrared night vision mode. The sixth camera was placed in a different, darkened room specifically for the purpose of collecting infrared mug pictures. The high-quality camera used to capture visible light mug photos was mounted identical to the infrared camera, but in a separate room with regular indoor lighting and a sufficient flash. Mug shot imaging circumstances are identical to those used in law enforcement or Imagery of passports and other personal identification documents . All 6 cameras (5 surveillance and 1 infrared mug shot) were attached to a professional digital video surveillance recorder that continuously recorded all six video streams on an internal hard drive.

Cam1, cam2, cam3, cam4, and cam5 are the security cameras. cam1 and cam5 are also equipped with infrared night vision capabilities. The authors title their images taken in IR night vision mode as cam6 (basically cam1 when night vision is on) and cam7 (basically cam5 when night vision is on). Cam8 was the name given to the camera used to take infrared mug photographs. All cameras (surveillance and photo) were placed and fixed in identical places and remained stationary during the capture procedure.

The following method was followed by all participants in this experiment. They were required to walk in front of security cameras in the dark first, and then under uncontrolled interior illumination. They were required to halt at three premarked locations along their trek in front of the cameras. This method resulted in the capture of 21 photos per person (cam1-7 at distances of 4.20, 2.60 and 1.00 meters). Following that, subjects were photographed at close range in controlled settings using a digital photographer's camera (HR images with standard lightning conditions). These collections of photos depicts each face in nine distinct angles, going from left to right profile in equal increments of 22.5 degrees. To ensure that each subject's vision was comparable, numbered markers were placed as anchors. As a consequence, each individual has nine photos with perspectives ranging from -90 to +90 degrees, as well as another mug shot at 0 degrees. Finally, individuals entered a dark room equipped with a high-quality infrared night vision security camera capable of obtaining IR mug pictures at close range. This results in a total of 32 photos per subject in the database. Following the capture method, the faces of the subjects were retrieved from the acquired photos. The following names were assigned to the captured images:

- Surveillance cameras (cam 1-7): *subjectID_camNum_distancelabel.jpg.*
- IR frontal mug shot: *subjectID_cams.jpg*
- Visible light mug shot: *subjectID_frontal.jpg*
- Different pose images: *subjectID_angleLabel.jpg*

Thus, each image in the database was assigned a unique name that included information on the subject's unique identifier as well as the distance and imaging conditions at which the image was captured. The distance labels 1, 2, and 3 denote 4.20, 2.60, and 1.00 metres, respectively. The filename 001_*cam*1_1.*jpg* for example, indicates that this image represents subject 001 as acquired by surveillance camera 1 at a distance of 4.20 metres.

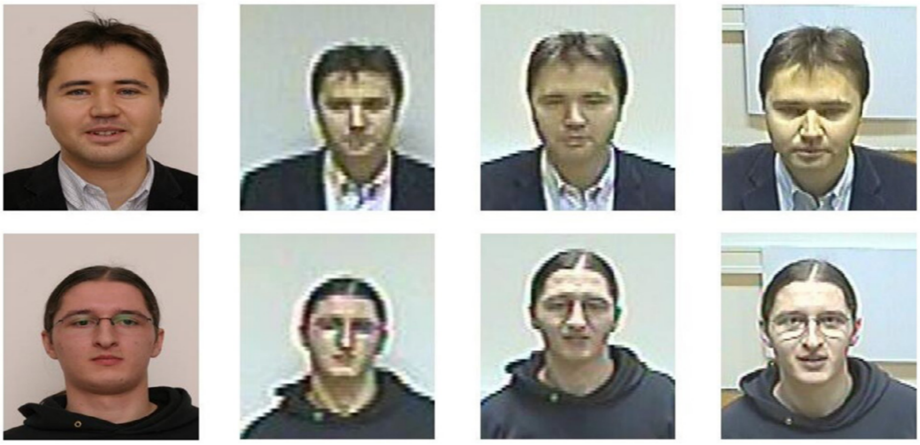


Figure 4.13: Sample Images from SCface Database [136]

Figure 4.13 depicts sample images from the database captured at different distances, here first column contains the high resolution mugshot images, second, third and fourth column represents the images at d1, d2 and d3, which is 4.20, 2.60 and 1.00 meters respectively.

#### 4.3.1.1 Evaluation Protocol

SCface defines face identification with unpaired High and Low resolution faces. It mimics the real-world surveillance watch-list problem, where the gallery contains HR faces and the probe consists of low resolution faces captured from surveillance cameras. Original study of SCface proposed protocol for face recognition was used to test the performance and reporting the accuracy, out of 130 subjects 50 subjects are used for training (fine-tuning) and the remaining 80 for testing. The recognition accuracy at each distance i.e d1, d2, d3 and average accuracy is reported in this study.

## 4.4 Cross Spectral Matching

Matching active infrared (IR) facial probe images to a visible light face gallery images is a novel and difficult challenge. This scenario is prompted by a variety of real-world surveillance tasks, such as facial recognnition at night or in poor atmospheric circumstances. Cross-spectral face recognition (CFR) is used to identify individuals when comparative face images are captured using multiple sensing modalities, such as infrared vs. visible. While CFR is fundamentally more difficult than classical face recognition because to the large diversity in facial appearance caused by a modality gap, it outperforms classical face recognition in scenarios with low or difficult lighting, as well as in the presence of presentation challenges. CFR is more difficult to do than regular FR. This is mainly due to the three factors listed. Firstly, there is significant intra-spectral variance, in which face samples from the same individual might impart greater appearance diversity than face samples from different people within the same modality as seen in Figure 4.14. Second, the modality gap is a source of concern, as here is where appearance variation occurs. This may result in a decrease in the performance of face comparisons. Finally, a constraint has been the scarcity of cross-modality face picture pair training samples. It is noticed that recent breakthroughs in convolutional neural networks (CNNs) and generative adversarial networks (GANs) have enabled significant improvements in CFR [137], [138].

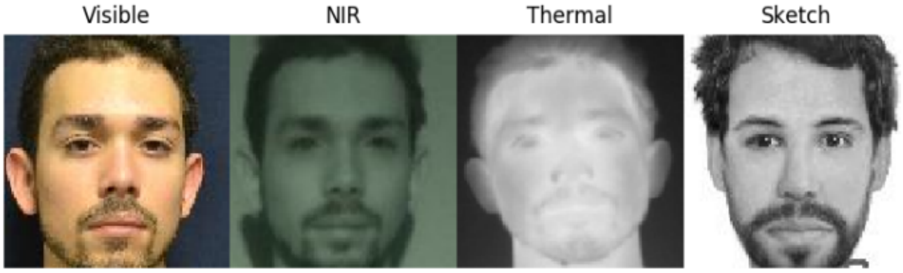


FIGURE 4.14: Heterogeneity among faces of same individual across different modalities [139]

Infrared (IR) spectral bands that have been used in Cross spectral face recognition, can be seen in Figure 4.15. Wolff et al. [140], Buddharaju et al. [141], Kong et al. [142], Bhowmik et al. [143], and Bourlai and Hornak [144] defined infrared light as an invisible, heat-associated energy that may be perceived when radiation or heat is reflected or emitted from an object. Unlike UV rays [145], infrared waves pass through the skin without causing harm. We notice that IR sensors can detect either the infrared light's face-reflection or the heat face-emission from subcutaneous superficial blood vessels. The infrared spectrum has been primarily used in spectroscopy [146], thermography [147], and astronomy [148].

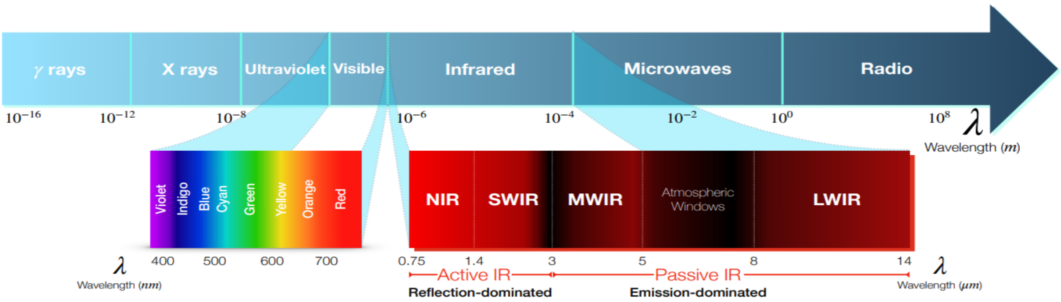


FIGURE 4.15: Heterogeneity among faces of same individual across different modalities [149]

According to ISO-20473:2007, infrared bands are described as near-infrared (NIR) between 0.78$\mu$m–3$\mu$m, mid-infrared (MIR) between 3$\mu$m–50$\mu$m, and far-infrared (FIR) between 50$\mu$m–1000$\mu$m. IR-A (0.7$\mu$m – 1.4$\mu$m), IR-B (1.4$\mu$m – 3$\mu$m), and IR-C (3$\mu$m – 1000$\mu$m).

Several database have been developed to assist the research in cross spectral matching, we are using TUFTS database to evaluate the performance of the selected deep learning models.

### 4.4.1 TUFTS Database

There are approximately 10,000 images in the Tufts [139] Face Database (74 women and 38 men, aged 4 to 70 years old with more than 15 nationalities) containing different image modalities, such as visible, near-infrared, thermal, LYTRO, recorded video, and 3D images.

Due to the widespread use of different sensors in everyday life, cross-modality face recognition is an emerging topic. ML algorithms that are data hungry, face recognition systems rely heavily on existing databases for evaluation and training examples. Unfortunately, there is currently no publicly accessible face database that contains more than two modalities for a given subject. Several images have been acquired for the Tufts Face Database, including photographs, thermal images, near infrared images, recorded video, a computerized facial sketch, and three-dimensional images of the volunteers faces. A protocol was obtained from the Institutional Review Board, and images were collected from students, staff, faculty, and their families at Tufts University.

In order to acquire images participants were seated in close proximity to the camera in front of a blue background. To achieve the optimal image center, each camera was mounted on a tripod and its height adjusted manually. In the acquisition process, the distance between the camera and the participant was controlled carefully. Diffuse lighting was used to ensure a constant lighting condition.

Figure 4.16 shows different image modalities among this database. Four cameras were used to capture 3D images. The camera was moved at 9 equidistant positions in order to form an approximate semicircle around each participant while they were instructed to look at a fixed view-point. Structure-from-motion algorithms were used to reconstruct the 3D models.

Software FACES 4.0 [150] was used for creating computerized facial sketches, one of the most widely used software packages among law enforcement agencies, the FBI, and the US Military. With the software, researchers can select candidates from the database based on their observations or memories.

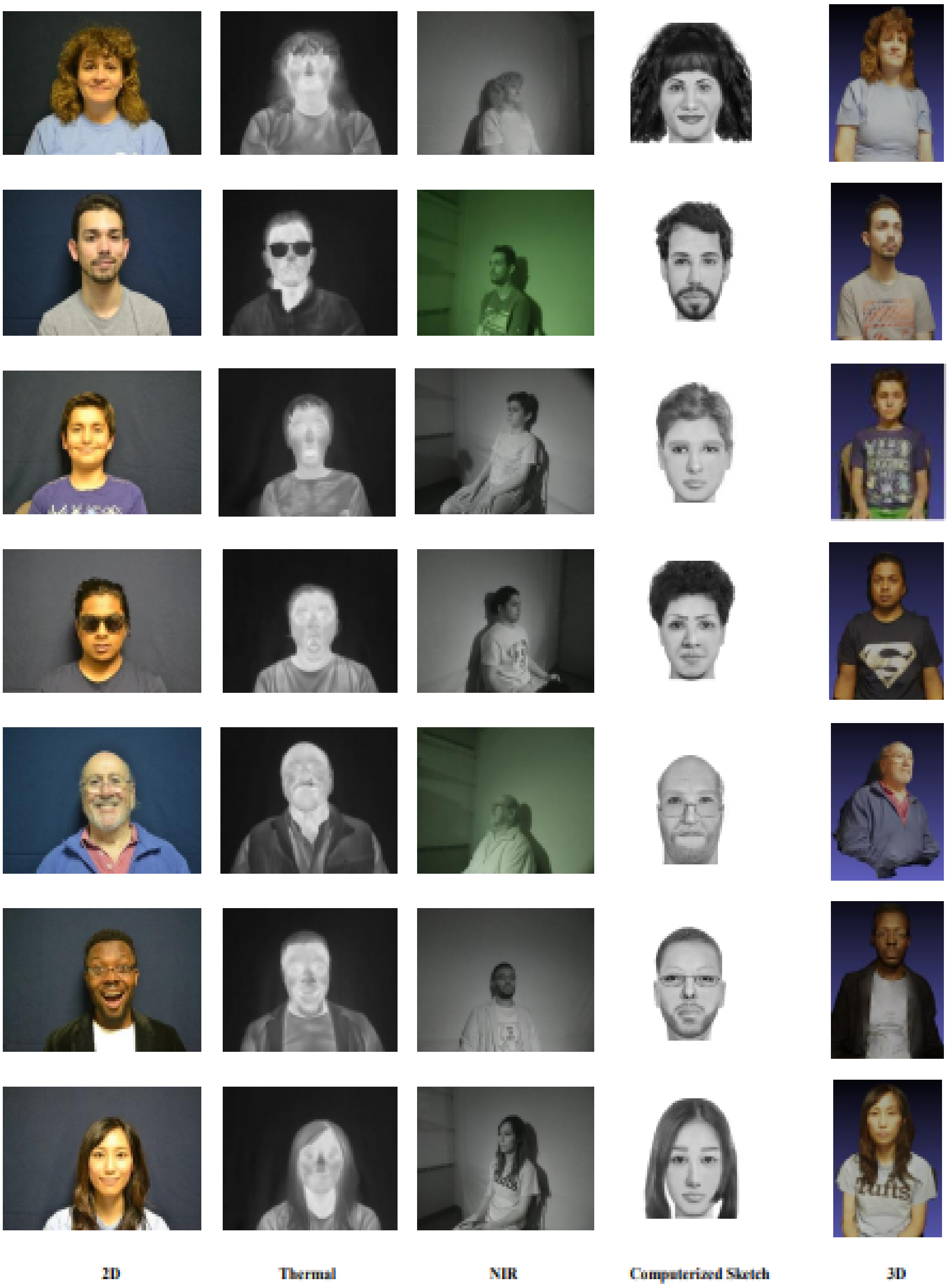


FIGURE 4.16: Images from different modalities present in TUFTS database [139]

FLIR Vue Pro cameras were used to capture the IR images. Nine cameras were placed at nine equidistant positions in a semi-circle around each participant and the participants were instructed to focus on a fixed point. RGB images were captured using a NIKON D3100 camera. Four night vision cameras were used to capture near-infrared images. An 850nm Infrared 96 LED light system was used to maintain the lighting condition for NIR imaging.

#### 4.4.1.1 Evaluation Protocol

In order to evaluate the performance of selected face recognition models, standard watch list identification protocol was used to obtain the results. For both NIR and Thermal images, only frontal images with neutral face expression were selected, for each of the 112 individuals as a probe. RGB frontal images from the database were used as a gallery set to perform the recognition test.

## 4.5 Ethnicity Effect

Several AI systems, including face recognition tools, are built around machine learning algorithms that are trained on labeled data. In a recent study, it was found that algorithms trained with biased data can not discriminate better [151], [152]. The stability of algorithm performance for populations of faces where demographics vary is critical to predicting the accuracy of face recognition when varying venue demographics are present.

In a study by [151] they demonstrated how Word2Vec, a popular embedding space with many applications, encodes societal gender biases. An analogy generator was trained using Word2Vec to fill in missing words in analogies. Using Word2Vec as the embedding, biases are likely to propagate throughout the system. There are no databases based on ethnicity that can be used to generalize research on the effect of ethnicity on gender classification accuracy.

Nearly 117 million Americans have been identified in federal law enforcement face recognition databases. African-Americans are more likely to be subjected to face recognition searches than individuals of other races. [153] Based on an analysis of 100 police departments, it was found that African-Americans have a higher likelihood of being stopped by law enforcement and subjected to face recognition searches than individuals of other races. Civil liberties are threatened by false positives and unwarranted searches. There has been considerable evidence that some face recognition systems misidentify people of color, women, and young people more often than average [154]. To protect citizens' rights and hold vendors and law enforcement accountable, phenotypic and demographic accuracy of these systems as well as their use must be monitored.

Earlier this year NIST, a physical sciences laboratory and non-regulatory agency of the US, also publishes a report [155] analyzing the efficency and accuracy, across racial groups, of 189 various FR algorithms developed and proposed by 99 differentS companies, including Microsoft, Intel, and other big names in in technology and surveillance industry. It has also been found that many of these algorithms misidentify a black or east asian face at the rate between 10 and 100 times greater than a white face. Also, misidentifications of American Indian faces tend to be the most frequent. Black women were less likely than any other demographic to be correctly identified by most algorithms. There are three distinct ways that race can influence the development and performance efficiency of FR technology. Racially disparate results are primarily due to non-diverse training images, human bias, and the availability of high-quality data. In general, lighter skin tones dominate the distribution of faces used to train the algorithm i.e. 83.5% of the faces in the LFW (a benchmark open-source database) dataset are white. Older algorithms used for FR have human selection of facial features, along with poor image quality that predominantly affects darker skin tones. Together, these problems cause the FR algorithms to perform unjust across races. Dark skin tones, in particular, tend to perform badly.

Older face recognition algorithms, which relies on the manual input of humans in the selection of facial features to analyze, may also be impacted by race. In addition to the appearance of human eye and nose to chin distance, colour and length of the eyebrow are possible qualities to consider. A person's choice of features is one of the factors under direct influence of his/her own race, as the research has shown that the race influences recognition of facial features. Among algorithms developed in China, Japan, and South Korea, Asian faces have shown better results than the Caucasian faces, while the opposite was true for algorithms developed in the United States, France, and Germany. This factor, together with the racial composition of the training images, may explain why, in a study by NIST, Asian faces made greater progress than Caucasian faces.

In order to evaluate the performance of our selected face recognition models, we selected the MIVIA Vmer Database with race labels to evaluate the performance.

### 4.5.1 VMER (VGG-Face2 Mivia Ethnicity Recognition)

In the VMER [156] dataset, images have been taken from the original VGGFace2, which includes more than 3.3 million face images, with an average of 362 images per subject (at least 87 images per subject). It also includes gender information, with 62% males and 38% females. The authors asked three individuals representing different ethnicities to label each identity with their ethnicity among the four identified, in order to avoid the other race effect. One African American, one Caucasian Latino, and one Asian Indian, were asked to do this. The authors then applied a majority voting rule to obtain the final annotations, and this allowed us to classify 99% of the face images based on ethnicity; the remaining 1% were sorted based on a tie-break rule by asking a fourth expert.

VMER's final dataset consists of 3,309,742 images of 9129 identities. Training and test sets are not subject overlapping i.e images of a subject provided in the training set is not available in the test set . Face analysis relies heavily on this separation in order to assess a neural network's generalizability. The labels of the database are the following:

1. African American
2. East Asian
3. Caucasian Latin
4. Asian Indian

AA (African American): members of this ethnicity group usually originate from Africa, North America or South America and are oftenly bears darkish skin with the lips and nose region being more prominent.

East Asian (EA): people of this group are either Chinese or have ancestry in East and South East Asia. Their skin tone is lighter, and their nose is relatively small, Besides their almond-shaped eyes, their most distinguishing characteristic is the inclination between their lateral and medial canthi, giving them the appearance of being narrow.

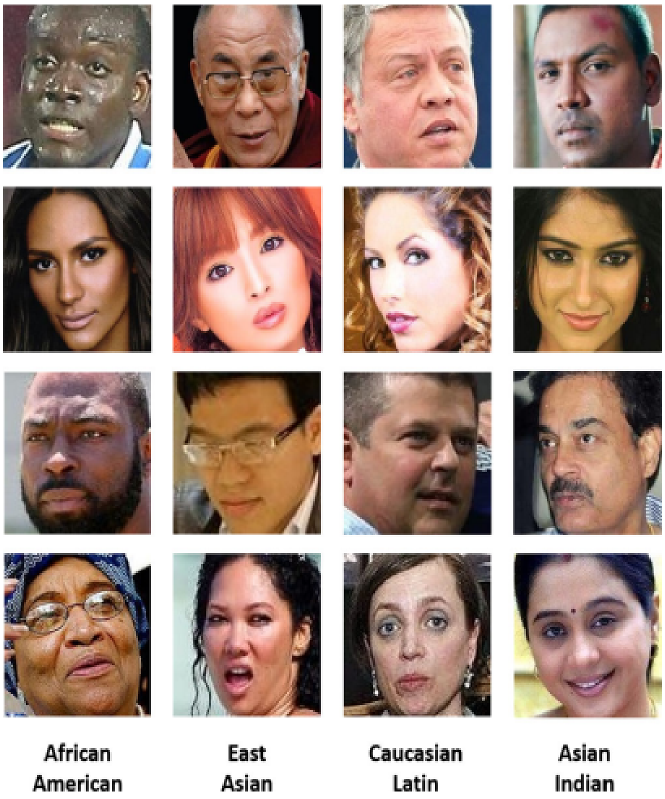


Figure 4.17: 4 different ethnic groups present in VMER Database [156]

Caucasians Latins (CL): As such, people of this ethnicity are derived from Europe, South America, Western Asia, and North Africa. They have a pale or tanned skin tone, a medium nose and lips, and horizontally aligned eyes.

Asian Indian (AI): this ethnicity group consists of people of Indian, South Asian, and Pacific Island descent. Although they share some characteristics with EAs and CLs, there are some slight differences that allow us to differentiate them. They possess a bit darker complexion of skin tones and more prominent facial features compared to East Asians and Caucasian Latins. Figure 4.17 depicts a selection of face images from the four ethnicity groups

#### 4.5.1.1 Evaluation Protocol

As the VMER database is build upon VGGFace 2 database, so in order to evaluate our pre-trained models the test set of the VGGFace 2 is used, which contains images from 500 individuals, All of these folders are manually sorted as by the available four ethnicity labels. Then in order to implement the watchlist (1:N) searching and verification protocol, the gallery images (one per identity) were selected among the available data, the rest of the image were treated as a probe set. The gallery image is manually selected based upon neutral expression, no pose variation (frontal images), even lightning conditions and images captured with no occlusion.

## 4.6 Summary

This chapter include the databases details and the evaluation protocol associated with each database in order to evaluate the model. Two benchmark databases i.e AgeDB and CALFW are used for checking the performance of model under Aging effect, CPLFW and CFP-FP database is used to evaluate the models under pose variation, SCface database is used to mock the scenario of real life surveillance face recognition application in which low resolution probe images are matched with high resolution gallery images. TUFTS database is used to evaluate the cross-spectral performance of face recognition algorithms and atlast VMER database is used to evaluate the models performance under Ethnicity change.

# Chapter 5

# Results and Evaluation

This chapter will discuss the results of face verification and recognition of the models under test and benchmark databases used to evaluate the performance.

## 5.1 Aging

For Deep Face models to be tested under age variations two popular benchmark databases were used for the evaluation. The details and the prorocol for each of these databases was discussed in Chapter 4. Table 5.1 Shows the Verification results of face matching on AgeDB-30 and CALFW, ArcFace with the accuracy of 98.08% and 95.87% leads the table for both the databases. However VGGFace2 performs the worst for AgeDB-30 achieving the accuracy of 85.11%, whereas FaceNET model achieves 83.41%, worst for CALFW database. CosFace and SphereFace shows the average performance for both the databases obtaining 97.30% and 97.91% for AgeDB-30 and 90.30% and 94.97% for CALFW. All of these results reflects the backbone architecture, the training database and the loss function used for the training of models. Here ArcFace published in 2021 performs the best where the database used is more comprehensive and largescale, as well as the angular loss function utilizes for the training provides the large separation margin between classes.

Table 5.1: Verification % of FR Models based on Age datasets

| FR Models | AgeDB-30 | CALFW |
|---|---|---|
| FaceNet | 98.05% | 89.41% |
| VGGFace2 | 85.11% | 90.57% |
| SphereFace | 97.30% | 90.30% |
| CosFace | 97.91% | 94.97% |
| ArcFace | 98.08% | 95.87% |

The bar graph in Figure 5.1 depict the trend among the verification accuracy of all the selected models.

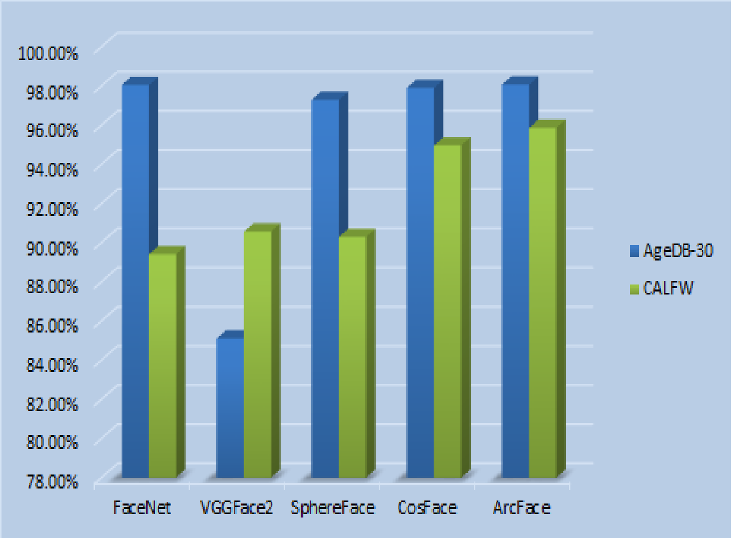


Figure 5.1: Face Verification% of selected FR models on AgeDB-30 and CALFW

## 5.2 Pose

Facial pose variation is one of the most important challenge among different factors that directly effects the performance of FR system, In order to evaluate the performance of selected pre-trained models CPLFW and CFP-FP databases were used, which are considered as benchmark while reporting the accuracy of models, under pose variation. Here Table 5.2 shows the Face Verification accuracy of different models, Here ArcFace achieves the highest accuracy i.e. 92.08% and 94.51% for both CFP-FP and CPLFW, In case of CFP-FP only frontal and profile photos are matched with large pose difference. Whereas SphereFace unexpectedly performs worse among all of the models selected with accuracy of 77.48% on CPLFW and FaceNET achieves 84.55%, the lowest on CFP-FP. The verification protocols and the details of datasets was already discussed in Chapter 4.

Table 5.2: Verification % of FR Models on Pose datasets

| FR Models | CPLFW | CFP-FP |
|---|---|---|
| FaceNet | 81.13% | 84.55% |
| VGGFace2 | 84.01% | 89.48% |
| SphereFace | 77.48% | 93.71% |
| CosFace | 88.88% | 94.40% |
| ArcFace | 92.08% | 94.51% |

ArcFace being trained with the modified loss fuction and ,MS1M comprehensive database with ResNET-100 architecture backbone outclass the rest of the models even when two images to be compared have huge pose difference, The trend among the accuracy of different models can be seen in Figure 5.2. VGGFace 2 model trained with VGGFace 2 comprehensive database shows the below average performance as compared to rest of the models achieving 84.01% and 89.48% accuracy on CPLFW and CFP-FP databases. The inconsistency of the verification accuracy on CPLFW and CFP-FP is seen clearly in the bar graph of Figure 5.2. SphereFace seems to be have a large variance and ARCFace is the one with less variace among the accuracy.

## 5.3 Resolution effect

Like other variations discussed earlier, the effect of resolution of facial image on the performance of face recognition systems is also very crucial specifically in surveillance scenarios, as the discriminative facial features and landmarks are much more evident in high resolution as compared to low resolution image. SCFace, a popular benchmarking database for low resolution face recognition is used to evaluate the performance of the selected FR models, The details about the database and the evaluation protocol was discussed in Chapter 4, here the results are reported in terms of Rank-1 identification accuracy, at the set of images captured at three different distances d1, d2 and d3 which are 4.20, 2.60 and 1.00 meters, respectively.

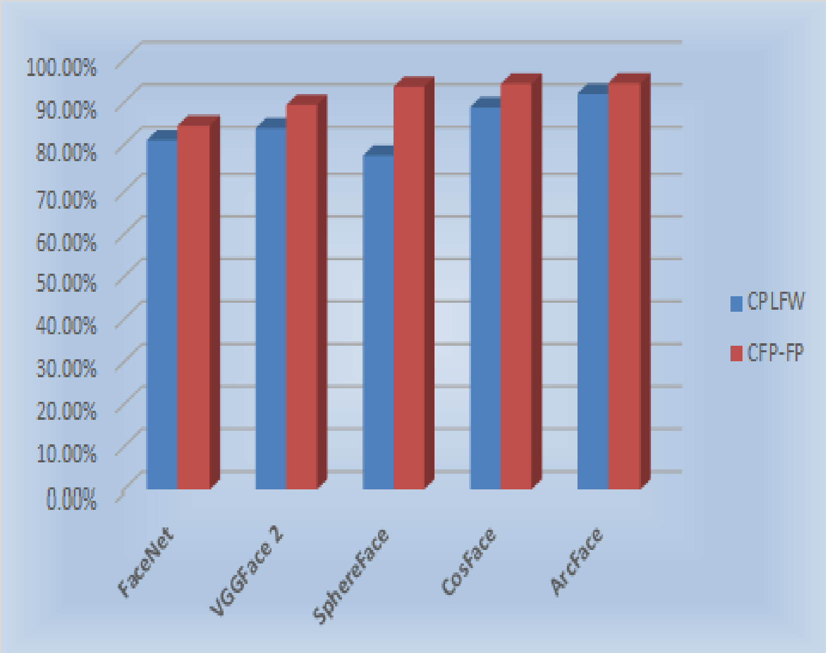


Figure 5.2: Trend of the Verification % on CPLFW and CFP-FP Database

Here again ARCFace leads the table by achieving the accuracy of 67.2% at d1, the most challenging one and 93.2% and 98.0% at d2 and d3 sequentially.

Table 5.3: Rank - 1 Identification % of FR Models on SCFace dataset

| FR Models | d1 | d2 | d3 | Average(%) |
|---|---|---|---|---|
| FaceNet | 25.7% | 24.8% | 31.7% | 27.4% |
| VGGFace2 | 48.0% | 72.4% | 76.3% | 65.6% |
| SphereFace | 61.5 % | 79.0% | 93.8% | 78.1% |
| CosFace | 63.3% | 81.1% | 94.3% | 79.5% |
| ArcFace | 67.2% | 93.2% | 98.0% | 86.1% |

FaceNET shows the worst performance, although it is trained using Triplet loss and large database, but that seems to be unhelpful when the resolution degrades. It achieves only 25.7% at d1 and its best 31.7% at d3. The last column of Table 5.3 shows the Average Identification accuracy at Rank-1, just to have a overall perspective of the performance. VGGFace2 shows 48.0% of accuracy at d1. Figure 5.3 depicts the trend among the accuracy of the selected models. Note that the models used for the evaluation are all trained on High resolution images, evaluating these models on low resolution images, degrade the performance as seen in the bar plot.

Super-resolution techniques are often used to convert low-resolution images into high-resolution ones so that the model trained on high-resolution images can be applied directly. In spite of super-resolution, the details are predicted and discriminative features are sometimes diminished, so classification models may not improve their ability to discriminate between enlarged images. Some recently developed techniques involving knowledge distillation for the training of models to recognize the low resolution images have shown promising results.

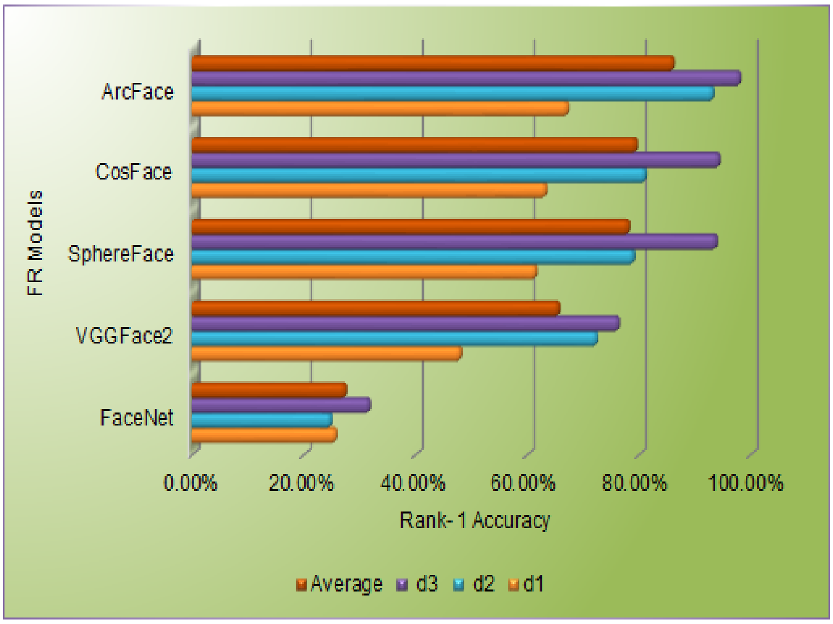


FIGURE 5.3: Rank-1 Identification (%) on SCface Database

## 5.4 Cross Spectral Recognition

The results of Cross Spectral Face Recognition as well as the averaging technique developed for the improvement of recognition will be discussed with detail in Chapter 6.

## 5.5 Ethnicity Effect

Face recognition technology has recently increased in availability, capability, and use, but there have also been statements that possible demographic differences may cause accuracy variations and bias. Research conducted by researchers at both Microsoft Research and the Massachusetts Institute of Technology (MIT) [157], as well as the US National Institute of Standards and Technology (NIST) [158], has uncovered persistent inaccuracies in algorithms for detecting and/or identifying faces of people of color as recently as 2018 and 2019.

A MIT/Microsoft study found that algorithms were less accurate for women than for men, with the biggest errors affecting women with dark skin, up to 35 percent. At least two major challenges have been highlighted. Identifying individuals across racial, ethnic, gender, and age groups requires algorithms with high comparative accuracy and integrating these algorithms into real-world systems such as those in law enforcement or government surveillance is a critical task.

So, in order to evaluate and probe the potential demographic bias in FR algorithms, we have performed the Identification test using VMER Database.

This contains individuals from four different races including African American, East Asian, Caucasian Latin and Asian Indian. The Identification accuracy of these model for the mentioned races is obtained and presented in Table 5.4. Further details about dataset and the evaluation protocol was already explained in Section 4.5.1

Table 5.4: Rank-1 Identification % of FR Models on VMER Database

| FR Models | African American | East Asian | Caucasian Latin | Asian Indian |
|---|---|---|---|---|
| FaceNet | 87.41% | 86.58 % | 88.41% | 85.11% |
| VGGFace2 | 89.29% | 87.56% | 90.52% | 82.65% |
| SphereFace | 88.55% | 88.14% | 91.91% | 89.41% |
| CosFace | 88.31% | 89.15% | 91.00% | 91.74% |
| ArcFace | 92.57% | 91.87 % | 95.14% | 93.75% |

It can be seen that generally all of the FR models have been performing good for the Caucasian Latin type of faces. However for the African American class all of the models are performing under average. ArcFace model can be seen achieving the highest accuracy of 95.14% for the Caucasian Latin class. The performance of VGGFace2 is inconsistent across all the available classes. It also yields the lowest accuracy of 82.65% across all of the models and classes. The trend among the accuracy of different models can be seen in Figure 5.4

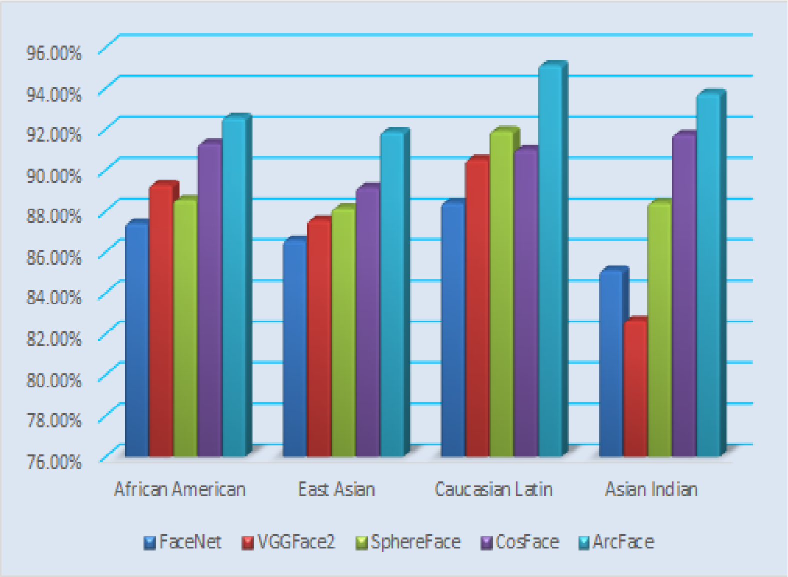


FIGURE 5.4: Rank-1 Identification (%) on VMER Database

## 5.6 Summary

This Chapter discuss the results and findings of the experiments done to evalute the performance under all the five scenarios. The ArcFace model outperforms all of the others. Other models, such as SphereFace and CosFace, also performed well.

# Chapter 6

# An Improved Approach for Cross Spectral Matching

## 6.1 Introduction

Identifying faces in the real world poses a challenge because of illumination variation. It is possible to acquire high-quality images in low-light conditions or complete darkness by using Near Infrared (NIR) or Infrared imaging. The technology has thus been widely adopted in applications like mobile devices, video surveillance and user authentication. Many applications, such as online registration and pre-enrollment using passports or government ID cards, require that face templates be enrolled on the basis of visible (VIS) images. Thus, NIR to VIS face matching has drawn much attention in machine learning and computer vision. Furthermore, it has been the most studied research topic in the field of heterogeneous face recognition (HFR), which is an image matching procedure over multiple spectral (or sensing) domains that contrasts with conventional VIS face recognition for homogeneous conditions.

Currently, for the purpose of deep HFR, convolutional neural networks (CNNs) training on web-scale VIS face datasets and tuning them on NIR-VIS datasets are commonly applied because they are time and cost-effective when obtaining large-scale pair-wised face images from multiple domains [159], [160]. In this work we are about to evaluate the performance of deep face models trained on High resolution visible images and test them for thermal and near infrared (NIR) images. After direct matching, we performed the negative thermal image matching and then also the average score of the predictions of models are used to compute the accuracy, which has shown the increase in identification accuracy.

## 6.2 Evaluation Methodology

For this task TUFTS database is used, as it contains both Thermal and NIR images, detail about the database and the evaluation protocol was discussed briefly in Chapter 4. Firstly, The faces were detected and pre-processed, and all images were cropped to fit the input sizes of the respective models.

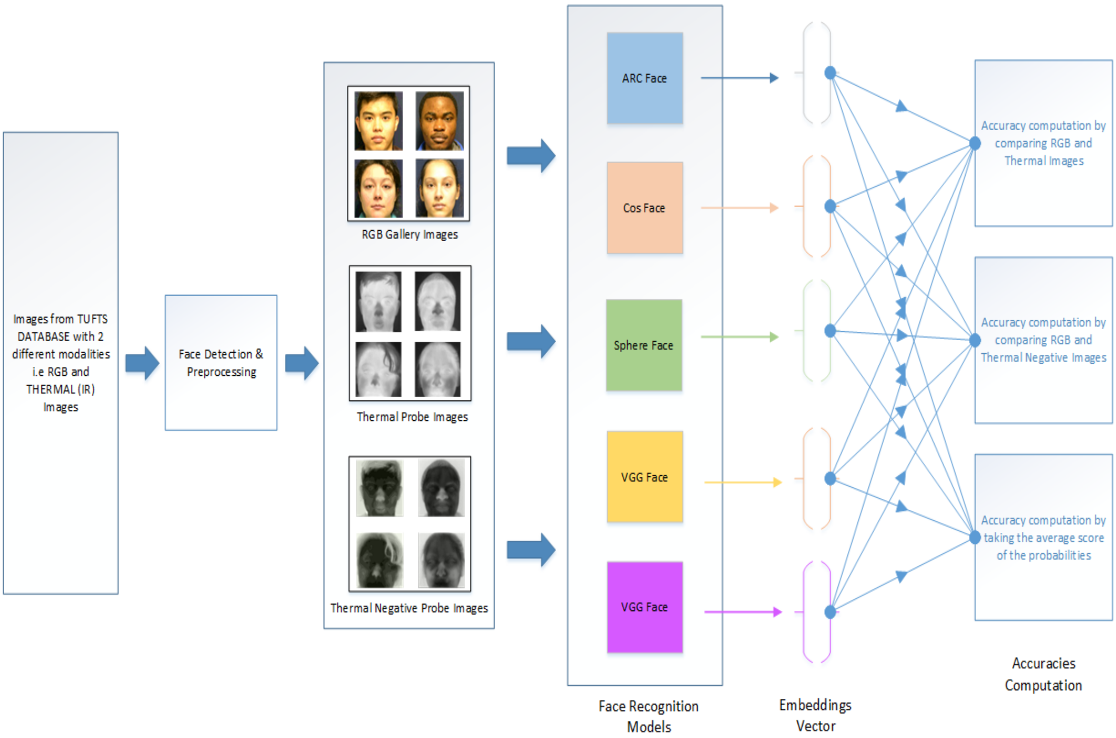

Images from TUFTS DATABASE with 2 different modalities i.e RGB and THERMAL (IR) Images
Face Detection & Preprocessing
RGB Gallery Images
Thermal Probe Images
Thermal Negative Probe Images
ARC Face
Cos Face
Sphere Face
VGG Face
VGG Face
Face Recognition Models
Embeddings Vector
Accuracy computation by comparing RGB and Thermal Images
Accuracy computation by comparing RGB and Thermal Negative Images
Accuracy computation by taking the average score of the probabilities
Accuracies Computation

FIGURE 6.1: Depiction of Proposed Evaluation Methodology

Input image size for ArcFace and CosFace is 112x112, SphereFace is 112x96, VGGFace2 is 224x224, and FaceNet is 160x160. Initially MTCNN was used to detect the faces and perform the alignment, but it only works with Visible Gallery images and Near Infrared Images, for pure thermal images it was unable to detect the faces in most cases, However different thresholds were also tuned of MTCNN, but it didnot helped either. So, RetinaFace algorithm was used to detect faces in case of thermal probe images which performed as per the expectaition and remains successful in detecting all the faces.

FR models extract the features from detected and aligned images and output embedding vectors corresponding to each face, which is essentially an identity code for each face. The embedding vectors are computed both for the RGB images from the gallery as well as for the thermal ones from the probe. Here, we computed accuracies three different ways as depicted in Figure 6.1 and presented the results:

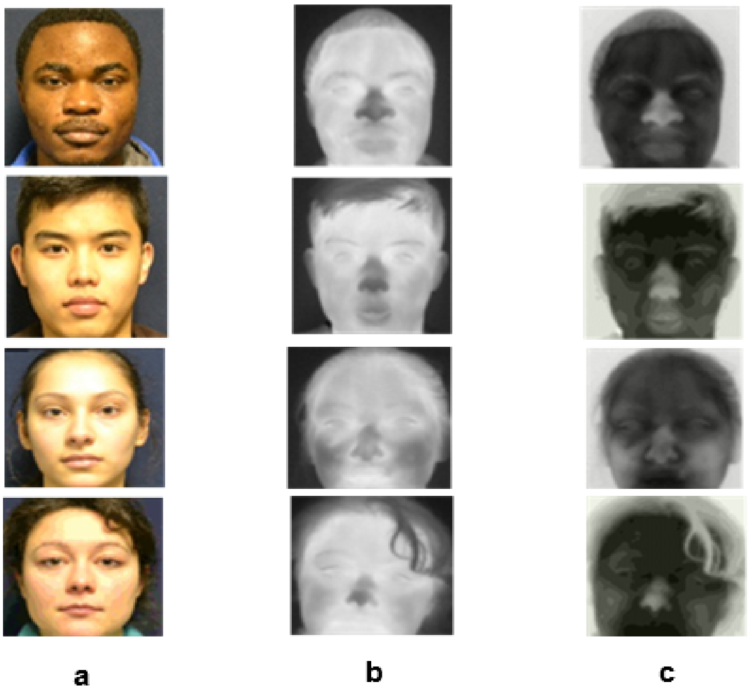


FIGURE 6.2: (a) RGB Gallery Images , (b) Corresponding Thermal Image, (c) Corresponding Thermal Negative Image

1. The Euclidean distance between the RGB gallery embedding and Thermal (IR) gallery embedding is computed.

2. Calculating the Euclidean distance between the RGB gallery embeddings and the Thermal (IR) Negative embeddings of the RGB gallery.

3. Calculation of the Euclidean distance between the RGB gallery and the thermal (average) image. When both the positive thermal  negative thermal

images were input, the average score of embeddings was computed as the average of the probabilities at the output of the models.

We have also reported the results of direct cross spectral matching for Near Infrared (NIR) Images.

Rank-1 accuracy was reported for NIR whereas Rank-1 and Rank-2 accuracy was reported for Thermal images matching using standard watchlist identification protocol.

## 6.3 Results and discussions

Table **??** shows the results of cross spectral face recognition. ArcFace appears to fare better than the others, as the training data used is vast and comprehensive, yet the performance of the other models is not overly poor either, as we are performing the recognition task on the models with input images from two different spectral ranges, i.e. as shown in Figure 6.2. RGB gallery and Thermal probe images.

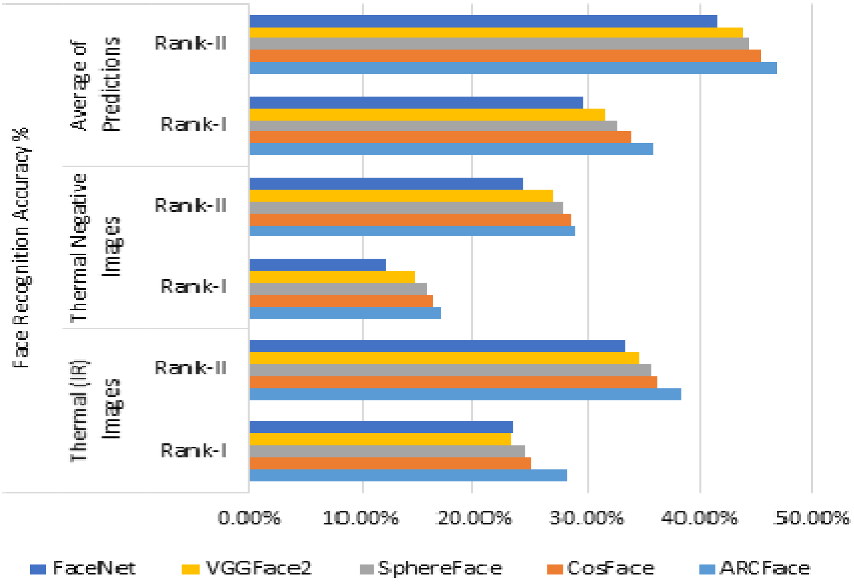


FIGURE 6.3: Trend of FR Accuracy % in Cross Spectral Matching

As seen in the Table 6.1, Direct cross spectral matching results in the low values of accuracy, Here ArcFace achieves the best accuracy of 28.15% among 113 subjects,

Table 6.1: Identification Accuracy % of FR Models (Direct CS Matching)

| | Face Recognition Accuracy % | | | | | |
|---|---|---|---|---|---|---|
| FR Models | Thermal (IR) Images | | Thermal Negative Images | | Average of Predictions | |
| | Rank-I | Rank-II | Rank-I | Rank-II | Rank-I | Rank-II |
| ArcFace | 28.15% | 38.26% | 17.00% | 29.00% | 35.77% | 46.77% |
| CosFace | 25.11% | 36.11% | 16.45% | 28.56% | 33.95% | 45.44% |
| SphereFace | 24.49% | 35.52% | 15.85% | 27.77% | 32.75% | 44.26% |
| VGGFace2 | 23.31% | 34.48% | 14.67% | 26.97% | 31.54% | 43.82% |
| FaceNet | 23.44% | 33.58% | 12.23% | 24.45% | 29.56% | 41.56% |

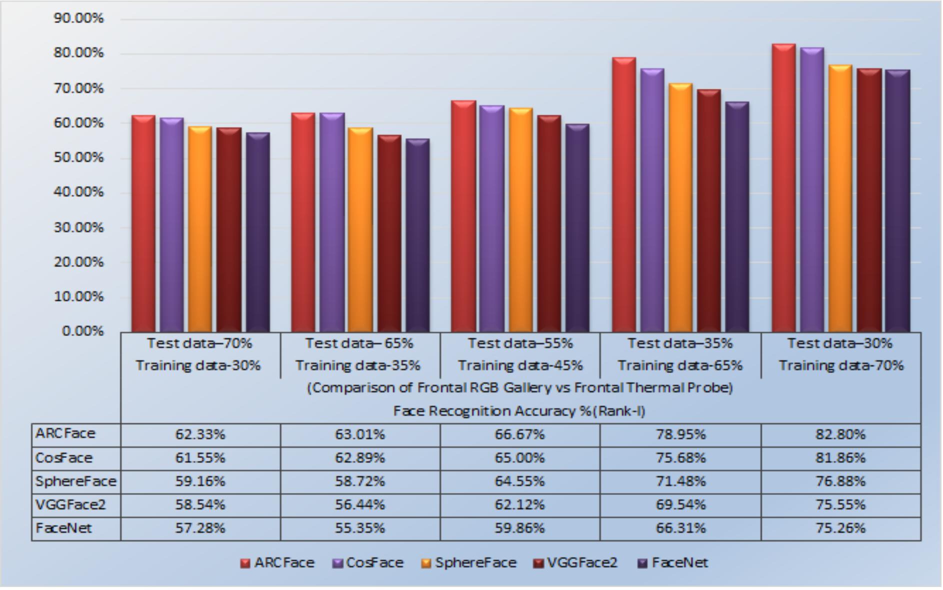
as the model is trained on RGB high resolution images, whereas when the Thermal negative images are used as a probe images and are compared with the RGB mugshot image, the results are more disturbing as the accuracy drops to 17.00% at Rank-1 for ArcFace compared with 28.15% with Thermal positive images.

The last two column of the Table 6.1 presents the Average score results, the average of the score of both the thermal positive and thermal negative images are used for the comparison with RGB gallery images. The averaging scheme worked very well boosting the overall accuracy upto 7% at Rank-1 and 9% at Rank-2.

### 6.3.1 Results after Fine-Tuning

As demonstrated in Table 6.4, the results are promising. However, fine tuning and transfer learning may be required to improve the results for a specific task, such as recognising thermal faces. The advantages of transfer learning include reduced training time and better performance compared to training from scratch. In order to fine-tune the model, the weights of each model were frozen and an extra layer was added at the end, which was then trained at a very low learning rate. As part of our testing, we change the proportion of training and testing data after fine-tuning and compare the RGB frontal images to the images from different angles, as well as the frontal thermal images only.

A total of 1008 images of 112 individuals with 9 faces per subject were used for fine tuning.

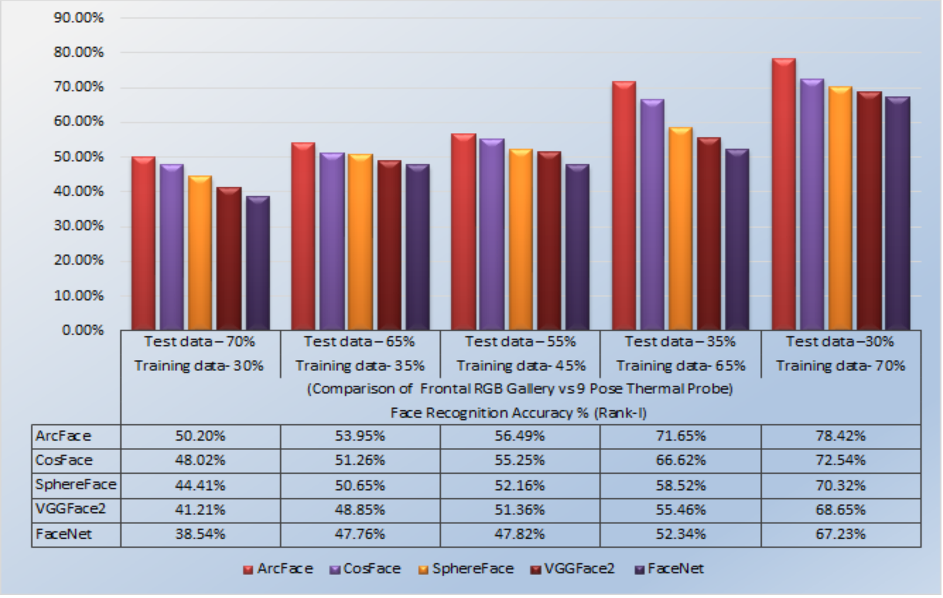


| | (Comparison of Frontal RGB Gallery vs Frontal Thermal Probe)<br>Face Recognition Accuracy %(Rank-I) | | | | |
|---|---|---|---|---|---|
| | Test data–70%<br>Training data-30% | Test data– 65%<br>Training data-35% | Test data–55%<br>Training data-45% | Test data–35%<br>Training data-65% | Test data–30%<br>Training data-70% |
| ARCFace | 62.33% | 63.01% | 66.67% | 78.95% | 82.80% |
| CosFace | 61.55% | 62.89% | 65.00% | 75.68% | 81.86% |
| SphereFace | 59.16% | 58.72% | 64.55% | 71.48% | 76.88% |
| VGGFace2 | 58.54% | 56.44% | 62.12% | 69.54% | 75.55% |
| FaceNet | 57.28% | 55.35% | 59.86% | 66.31% | 75.26% |

FIGURE 6.4: Trend of Face Identification % (Rank-1 by matching both frontal gallery and probe images)

The results of the fine tuning are shown in Tables with Figure 6.4 and 6.5. In Figure and Table 6.4, we are showing results of an experiment in which different proportions of training and testing datasets were used to compute accuracy.

According to this, ArcFace has the best recognition accuracy as at train:test::30:70, its recognition accuracy was 62.33%, whereas it increased to 82.80% when training and testing were split respectively 70% and 30%.



| | (Comparison of Frontal RGB Gallery vs 9 Pose Thermal Probe)<br>Face Recognition Accuracy % (Rank-I) | | | | |
|---|---|---|---|---|---|
| | Test data – 70%<br>Training data- 30% | Test data – 65%<br>Training data- 35% | Test data – 55%<br>Training data- 45% | Test data – 35%<br>Training data- 65% | Test data –30%<br>Training data- 70% |
| ArcFace | 50.20% | 53.95% | 56.49% | 71.65% | 78.42% |
| CosFace | 48.02% | 51.26% | 55.25% | 66.62% | 72.54% |
| SphereFace | 44.41% | 50.65% | 52.16% | 58.52% | 70.32% |
| VGGFace2 | 41.21% | 48.85% | 51.36% | 55.46% | 68.65% |
| FaceNet | 38.54% | 47.76% | 47.82% | 52.34% | 67.23% |

FIGURE 6.5: Trend of Face Identification % (Rank-1 by matching both frontal gallery vs 9 probe images)

We compare only faces from frontal thermal (IR) images with those from frontal RGB images in Table 6.4. However, in Table and Figure 6.5, we compared the frontal RGB maps with all nine different thermal (IR) probe maps of a single subject as available in the dataset, and here the accuracy of ArcFace was just 50.20% at train:test::30:70, which was lesser than that of Table and Figure 6.4 with the same proportion of train/test split. Because of this, the model is able to decide from a significant amount of data, i.e. the thermal images of the subject from 9 different sides. As of train:test::70:30, the accuracy was 78.24%. Accuracy of other models shows the same trend.

## 6.4 Summary

This Chapter presents the comparison of selected models under cross spectral face recognition i.e in our case models are pre-trained on RGB images and they are deployed for thermal face recognition scenario (RGB gallery and Thermal Probe images). A unique approach has also been proposed in this section which helped in boosting the accuracy of face recognition algorithms.

# Chapter 7

# Conclusion and Future Work

## 7.1 Conclusion

The field of FR has been explored with several algorithms in the past. The objective of this research was to evaluate the performance of selected five state of the art face recognition algorithms. To evaluate the performance under five different scenarios i.e Aging, Pose, Low Resolution, Cross Spectral Face Recognition and Ethnicity and a total of seven databases are used to benchmark the performance. ArcFace model published in 2021 have so far performed the best among other models trained on large scale databases. This research has also proposed a new technique for matching the cross spectral images using the deep learning models trained on RGB images. The proposed technique for CFR has boosted the accuracy of all the models, without fine tuning on thermal database. The same trend can be seen after fine tuning the models.

## 7.2 Future Work

In future this research can be extended in a way that more facial modalities like, occlusion which is common in the surveillance applications, facial expressions and illumination effect can be examined for all of these models. This will present a more comprehensive analysis about the performances of all the models. An ensemble classifier can also be implemented based on the output of all of these models so that a reliable face recognition system can be implemented.

# Bibliography

[1] M. R. D. Rodavia, O. Bernaldez, and M. Ballita, “Web and mobile based facial recognition security system using eigenfaces algorithm,” in *2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. IEEE, 2016, pp. 86–92.

[2] B. DeCann and A. Ross, “Relating roc and cmc curves via the biometric menagerie,” in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2013, pp. 1–8.

[3] M. Carandini, “What simple and complex cells compute,” *The Journal of physiology*, vol. 577, no. Pt 2, p. 463, 2006.

[4] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152.

[5] Y. Freund, R. Schapire, and N. Abe, “A short introduction to boosting,” *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[6] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1. Ieee, 2001, pp. I–I.

[7] J. Callaham, “Lumua 730 used in world record attempt,” Nov 2014. [Online]. Available: https://www.windowscentral.com/lumia-730-used-attempt-create-worlds-largest-selfie-bangladesh

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *European conference on computer vision.* Springer, 2006, pp. 404–417.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[13] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.

[14] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with mtcnn," in *2017 4th international conference on information science and control engineering (ICISCE).* IEEE, 2017, pp. 424–427.

[15] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4507–4515.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[17] P. Hu and D. Ramanan, "Finding tiny faces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 951–959.

[18] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "Ssh: Single stage headless face detector," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4875–4884.

[19] X. Shi, S. Shan, M. Kan, S. Wu, and X. Chen, "Real-time rotation-invariant face detection with progressive calibration networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2295–2303.

[20] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[23] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[24] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 761–769.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[26] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[28] G. B. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images," in *2007 IEEE 11th international conference on computer vision.* IEEE, 2007, pp. 1–8.

[29] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE transactions on image processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

[30] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE transactions on pattern analysis and machine intelligence*, vol. 15, no. 10, pp. 1042–1052, 1993.

[31] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern recognition*, vol. 25, no. 1, pp. 65–77, 1992.

[32] K.-M. Lam and H. Yan, "An analytic-to-holistic approach for face recognition based on a single frontal view," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 7, pp. 673–686, 1998.

[33] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A hybrid neural network approach," Tech. Rep., 1998.

[34] A. K. Jain and S. Z. Li, *Handbook of face recognition.* Springer, 2011, vol. 1.

[35] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[36] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[37] P. Comon, “Independent component analysis, a new concept?” *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[38] M.-H. Yang, “Face recognition using kernel methods,” *Advances in neural information processing systems*, vol. 14, 2001.

[39] M. Ringnér, “What is principal component analysis?” *Nature biotechnology*, vol. 26, no. 3, pp. 303–304, 2008.

[40] S. Balakrishnama and A. Ganapathiraju, “Linear discriminant analysis-a brief tutorial,” *Institute for Signal and information Processing*, vol. 18, no. 1998, pp. 1–8, 1998.

[41] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, “A feature selection method based on improved fisher’s discriminant ratio for text sentiment classification,” *Expert Systems with Applications*, vol. 38, no. 7, pp. 8696–8702, 2011.

[42] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, “Face recognition by independent component analysis,” *IEEE Transactions on neural networks*, vol. 13, no. 6, pp. 1450–1464, 2002.

[43] M.-H. Yang, “Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods,” in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, 2002, pp. 215–220.

[44] T. Abhishree, J. Latha, K. Manikantan, and S. Ramachandran, “Face recognition using gabor filter based feature extraction with anisotropic diffusion as a pre-processing technique,” *Procedia Computer Science*, vol. 45, pp. 312–321, 2015.

[45] F. Bellifemine, A. Capellino, A. Chimienti, R. Picco, and R. Ponti, “Statistical analysis of the 2d-dct coefficients of the differential signal for images,” *Signal Processing: Image Communication*, vol. 4, no. 6, pp. 477–488, 1992.

[46] T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[47] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[48] P. Cunningham and S. J. Delany, “k-nearest neighbour classifiers-a tutorial,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–25, 2021.

[49] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction.* Springer, 2009, vol. 2.

[50] M. A. Turk and A. P. Pentland, “Face recognition using eigenfaces,” in *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition.* IEEE Computer Society, 1991, pp. 586–587.

[51] K. Etemad and R. Chellappa, “Discriminant analysis for recognition of human face images,” *Josa a*, vol. 14, no. 8, pp. 1724–1733, 1997.

[52] G. Guo, S. Z. Li, and K. Chan, “Face recognition by support vector machines,” in *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. no. PR00580).* IEEE, 2000, pp. 196–201.

[53] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, “Attribute and simile classifiers for face verification,” in *2009 IEEE 12th international conference on computer vision.* IEEE, 2009, pp. 365–372.

[54] T. Berg and P. N. Belhumeur, “Tom-vs-pete classifiers and identity-preserving alignment for face verification.” in *Bmvc*, vol. 2. Citeseer, 2012, p. 7.

[55] M. Guillaumin, J. Verbeek, and C. Schmid, “Is that you? metric learning approaches for face identification,” in *2009 IEEE 12th international conference on computer vision.* IEEE, 2009, pp. 498–505.

[56] D. Gabor, “Theory of communication. part 1: The analysis of information,” *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[57] T. Barbu, "Gabor filter-based face recognition technique," *Proceedings of the Romanian Academy*, vol. 11, no. 3, pp. 277–283, 2010.

[58] F. Bellakhdhar, K. Loukil, and M. Abid, "Face recognition approach using gabor wavelets, pca and svm," *International Journal of Computer Science Issues (IJCSI)*, vol. 10, no. 2, p. 201, 2013.

[59] S.-H. Lin, S.-Y. Kung, and L.-J. Lin, "Face recognition/detection by probabilistic decision-based neural network," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 114–132, 1997.

[60] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.

[61] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[62] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.

[63] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[64] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[65] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[66] K. Gregor and Y. LeCun, "Emergence of complex-like cells in a temporal product network with local receptive fields," *arXiv preprint arXiv:1006.0448*, 2010.

[67] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *2012 IEEE conference on computer vision and pattern recognition.* IEEE, 2012, pp. 2518–2525.

[68] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1891–1898.

[69] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *European conference on computer vision.* Springer, 2012, pp. 566–579.

[70] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

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

[72] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[73] 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.

[74] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[75] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507*, 2017.

[76] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[77] Y. Wu, H. Liu, J. Y. Li, and Y. R. Fu, “Deep face recognition with center invariant loss,” *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017.

[78] M. A. Hasnat, J. Bohné, J. Milgram, S. Gentric, and L. Chen, “Deepvisage: Making face recognition simple yet with powerful generalization skills,” *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1682–1691, 2017.

[79] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

[80] F. Wang, J. Cheng, W. Liu, and H. Liu, “Additive margin softmax for face verification,” *IEEE Signal Processing Letters*, vol. 25, pp. 926–930, 2018.

[81] J. Deng, J. Guo, J. Yang, N. Xue, I. Cotsia, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition.” *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2021.

[82] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[83] Y. Yamada, M. Iwamura, and K. Kise, “Deep pyramidal residual networks with separated stochastic depth,” *ArXiv*, vol. abs/1612.01230, 2016.

[84] Z. Liu, P. Luo, X. Wang, and X. Tang, “Large-scale celebfaces attributes (celeba) dataset,” *Retrieved August*, vol. 15, no. 2018, p. 11, 2018.

[85] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, “Umdfaces: An annotated face dataset for training deep networks,” *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 464–473, 2017.

[86] D. Yi, Z. Lei, S. Liao, and S. Li, "Learning face representation from scratch," *ArXiv*, vol. abs/1411.7923, 2014.

[87] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74, 2018.

[88] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4873–4882, 2016.

[89] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016.

[90] Z. Mahmood, T. Ali, S. Khattak, and S. U. Khan, "A comparative study of baseline algorithms of face recognition," in *2014 12th International Conference on Frontiers of Information Technology*. IEEE, 2014, pp. 263–268.

[91] K. Subramanian, "Comparative analysis of advanced face recognition technique."

[92] L. Liying and H. Yue, "Comparative study of some face recognition algorithms," in *2008 International Conference on Wavelet Analysis and Pattern Recognition*, vol. 1. IEEE, 2008, pp. 343–346.

[93] W. H. Zhan, "Comparative study of face recognition classifier algorithm," in *Applied Mechanics and Materials*, vol. 696. Trans Tech Publ, 2015, pp. 110–113.

[94] U. I. Bajwa, I. A. Taj, M. W. Anwar, and X. Wang, "A multifaceted independent performance analysis of facial subspace recognition algorithms," *PloS one*, vol. 8, no. 2, p. e56510, 2013.

[95] S. Paul and S. K. Acharya, "A comparative study on facial recognition algorithms," in *e-journal-First Pan IIT International Management Conference–2018*, 2020.

[96] T. Schenkel, O. Ringhage, and N. Branding, “A comparative study of facial recognition techniques: With focus on low computational power,” 2019.

[97] B. Karimi, “Comparative analysis of face recognition algorithms and investigation on the significance of color,” Ph.D. dissertation, Concordia University, 2006.

[98] G. Vladimir, B. Dmitriy, T. N. Win, and N. W. Htet, “A comparative analysis of face recognition algorithms in solving the problem of visual identification,” in *2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus).* IEEE, 2017, pp. 666–668.

[99] A. R. Shinwari, A. J. Balooch, A. A. Alariki, and S. A. Abdulhak, “A comparative study of face recognition algorithms under facial expression and illumination,” in *2019 21st International Conference on Advanced Communication Technology (ICACT).* IEEE, 2019, pp. 390–394.

[100] G. Kumar *et al.*, “A comparative analysis of face recognition algorithms,” 2016.

[101] M. U. Rahman, “A comparative study on face recognition techniques and neural network,” *arXiv preprint arXiv:1210.1916*, 2012.

[102] G. Pala, “A comparative study of deep learning based face recognition algorithms for video under adverse conditions,” Ph.D. dissertation, Marmara Universitesi (Turkey), 2019.

[103] N. Ahmed, F. A. Khan, Z. Ullah, H. Ahmed, T. Shahzad, and N. Ali, “Face recognition comparative analysis using different machine learning approaches,” *Advances in Science and Technology Research Journal*, vol. 15, no. 1, pp. 265–272, 2021.

[104] R. Patel, N. Rathod, and A. Shah, “Comparative analysis of face recognition approaches: a survey,” *International Journal of Computer Applications*, vol. 57, no. 17, 2012.

[105] L. Wolf, T. Hassner, and I. Maoz, “Face recognition in unconstrained videos with matched background similarity,” *CVPR 2011*, pp. 529–534, 2011.

[106] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[107] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[108] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *ECCV*, 2018.

[109] D. Sandberg, "Davidsandberg/facenet: Face recognition using tensorflow," Apr 2018. [Online]. Available: https://github.com/davidsandberg/facenet

[110] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.

[111] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 2011–2023, 2020.

[112] A. F. Agarap, "Deep learning using rectified linear units (relu)," *ArXiv*, vol. abs/1803.08375, 2018.

[113] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.

[114] J. S. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in *NIPS*, 1989.

[115] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *ArXiv*, vol. abs/1704.04861, 2017.

[116] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017.

[117] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[118] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *NIPS*, 2017.

[119] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[120] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *ArXiv*, vol. abs/1512.01274, 2015.

[121] B. Esme and B. Sankur, "Effects of aging over facial feature analysis and face recognition," 2010.

[122] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs, "A study of face recognition as people age," *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, 2007.

[123] A. Lamont, S. Stewart-Williams, and J. Podd, "Face recognition and aging: Effects of target age and memory load," *Memory & Cognition*, vol. 33, pp. 1017–1024, 2005.

[124] C. Gohringer and A. Limited, "Advances in face recognition technology and its application in airports," 2012.

[125] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: The first manually collected, in-the-wild age database," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1997–2005, 2017.

[126] T. Zheng and W. Deng, “Cross-pose lfw : A database for studying cross-pose face recognition in unconstrained environments,” 2018.

[127] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth, “Whos in the picture,” in *NIPS*, 2004.

[128] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. G. Learned-Miller, and D. A. Forsyth, “Names and faces in the news,” in *CVPR 2004*, 2004.

[129] R. Rothe, R. Timofte, and L. V. Gool, “Dex: Deep expectation of apparent age from a single image,” *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 252–257, 2015.

[130] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzàlez, H. J. Escalante, D. Misevic, U. K. Steiner, and I. Guyon, “Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results,” *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 243–251, 2015.

[131] J. C. Kavitha and T. T. Mirnalinee, “Automatic frontal face reconstruction approach for pose invariant face recognition,” *Procedia Computer Science*, vol. 87, pp. 300–305, 2016.

[132] S. Banerjee, J. Brogan, J. Krizaj, A. Bharati, B. RichardWebster, V. truc, P. J. Flynn, and W. J. Scheirer, “To frontalize or not to frontalize: Do we really need elaborate pre-processing to improve face recognition?” *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 20–29, 2018.

[133] C. Ding, C. Xu, and D. Tao, “Multi-task pose-invariant face recognition,” *IEEE Transactions on Image Processing*, vol. 24, pp. 980–993, 2015.

[134] Y. Gao and H. J. Lee, “Cross-pose face recognition based on multiple virtual views and alignment error,” *Pattern Recognit. Lett.*, vol. 65, pp. 170–176, 2015.

[135] C. C. V. P. R. C. D. J. S. Sengupta, J.C. Cheng, "Frontal to profile face verification in the wild," in *IEEE Conference on Applications of Computer Vision*, February 2016.

[136] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *Multimedia tools and applications*, vol. 51, no. 3, pp. 863–879, 2011.

[137] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2439–2448.

[138] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision.* Springer, 2016, pp. 694–711.

[139] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani *et al.*, "A comprehensive database for benchmarking imaging systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 509–520, 2018.

[140] L. B. Wolff, D. A. Socolinsky, and C. K. Eveland, "Face recognition in the thermal infrared," in *Computer Vision Beyond the Visible Spectrum.* Springer, 2005, pp. 167–191.

[141] P. Buddharaju, I. T. Pavlidis, P. Tsiamyrtzis, and M. Bazakos, "Physiology-based face recognition in the thermal infrared spectrum," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 4, pp. 613–626, 2007.

[142] S. G. Kong, J. Heo, B. R. Abidi, J. Paik, and M. A. Abidi, "Recent advances in visual and infrared face recognition—a review," *Computer vision and image understanding*, vol. 97, no. 1, pp. 103–135, 2005.

[143] M. K. Bhowmik, K. Saha, S. Majumder, G. Majumder, A. Saha, A. N. Sarma, D. Bhattacharjee, D. K. Basu, and M. Nasipuri, "Thermal infrared

face recognition—a biometric identification technique for robust security system," *Reviews, refinements and new ideas in face recognition*, vol. 7, pp. 113–138, 2011.

[144] T. Bourlai and L. A. Hornak, "Face recognition outside the visible spectrum," *Image and Vision Computing*, vol. 55, pp. 14–17, 2016.

[145] N. Narang, T. Bourlai, and L. A. Hornak, "Can we match ultraviolet face images against their visible counterparts?" in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXI*, vol. 9472. SPIE, 2015, pp. 597–608.

[146] B. H. Stuart, *Infrared spectroscopy: fundamentals and applications.* John Wiley & Sons, 2004.

[147] K.-P. Möllmann and M. Vollmer, *Infrared thermal imaging: fundamentals, research and applications.* John Wiley & Sons, 2017.

[148] I. S. Glass and I. Glass, *Handbook of infrared astronomy.* Cambridge University Press, 1999, no. 1.

[149] D. Anghelone, C. Chen, A. Ross, and A. Dantcheva, "Beyond the visible: A survey on cross-spectral face recognition," *arXiv e-prints*, pp. arXiv–2201, 2022.

[150] H. K. Galoogahi and T. Sim, "Face sketch recognition by local radon binary pattern: Lrbp," in *2012 19th IEEE International Conference on Image Processing.* IEEE, 2012, pp. 1837–1840.

[151] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[152] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[153] C. Garvie, *The perpetual line-up: Unregulated police face recognition in America.* Georgetown Law, Center on Privacy & Technology, 2016.

[154] P. N. Schuetz, "Fly in the face of bias: Algorithmic bias in law enforcement's facial recognition technology and the need for an adaptive legal framework," *Law & Ineq.*, vol. 39, p. 221, 2021.

[155] P. Grother, P. Grother, M. Ngan, and K. Hanaoka, "Face recognition vendor test (frvt) part 2: Identification," 2019.

[156] A. Greco, G. Percannella, M. Vento, and V. Vigilante, "Benchmarking deep network architectures for ethnicity recognition using a new large face dataset," *Mach. Vis. Appl.*, vol. 31, p. 67, 2020.

[157] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *FAT*, 2018.

[158] P. Grother, M. Ngan, and K. K. Hanaoka, "Face recognition vendor test part 3:," 2019.

[159] M. A. Akhloufi and A. Bendada, "Infrared face recognition using texture descriptors," in *Thermosense XXXII*, vol. 7661. SPIE, 2010, pp. 49–58.

[160] D. A. Socolinsky and A. Selinger, "A comparative analysis of face recognition performance with visible and thermal infrared imagery," in *Object recognition supported by user interaction for service robots*, vol. 4. IEEE, 2002, pp. 217–222.