

INTERNATIONAL ISLAMIC UNIVERSITY ISLAMABAD
FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE

Date: 28-02-2024

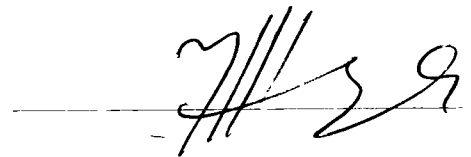
Final Approval

It is certified that we have read this thesis, entitled "Hybrid Landmark Holistic based Facial Recognition System in Video Surveillance" submitted by **Muhammad Asad** Registration No. **115-FBAS/PHDCS/F14**. It is our judgment that this thesis is of sufficient standard to warrant its acceptance by the International Islamic University Islamabad for the award of the degree of PhD in Computer Science.

Committee

External Examiner:

Dr. Muhammad Imran,
Assistant Professor
SZABIST, Islamabad



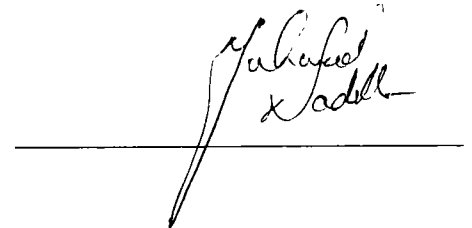
External Examiner:

Dr. Sadaf Abdul Rauf,
Associate Professor
Fatimah Jinah University, Rawalpindi



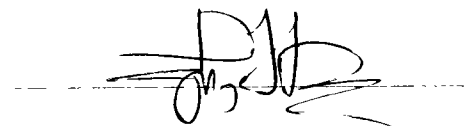
Internal Examiner:

Dr. Muhammad Nadeem,
Assistant Professor,
Department of Computer Science,
International Islamic University, Islamabad



Supervisor:

Dr. Ayyaz Hussain,
Professor,
Department of Computer Science,
Quaide Azam University, Islamabad



TH-27315 1/4

PhD
006.42
MUH

Face recognition system
Facial recognition- Data processing
Pattern recognition system
Computer vision
Video surveillance
Artificial intelligence
Image processing Data analysis

Declaration

“I hereby declare and affirm that this thesis neither as a whole, nor as part thereof has been copied out from any source. It is further declared that I have completed this thesis entirely on the basis of my personal effort, made under the sincere guidance of my supervisor. If any part of this report is proven to be copied or found to be a reproduction of some other, I shall stand by the consequences. No portion of the work presented in this report has been submitted in support of an application for other degree or qualification of this or any other university or institute of learning.”

Muhammad Asad
115-FBAS/PHDCS/F14

Acknowledgement

Praise be to Allah Almighty for His endless blessings upon us. Peace be upon the Prophet Muhammad (S.A.W) who was sent as mercy for this world and hereafter.

I would like to thank my family members for their support and encouragement and express my sincere gratitude to Dr. Ayyaz Hussain for his support and guidance as a teacher during course work and his efforts afterwards towards my research work. His appreciable comments, recommendations and motivation during the PhD have been of great value to me.

I am thankful to all other persons who have directly or indirectly helped me in the completion of my research work.

Dedication

Dedicated to my parents without whose support I would not have been able to complete my thesis work.

My mentor who has been a great guidance and inspiration throughout the thesis work.

I pray that may Allah bless them with a long and blessed life and let them enjoy the best of health throughout.

Muhammad Asad

Abstract

Face recognition has remained one of the main focuses of attention for the computer vision research community for the last couple of decades. A very substantial amount of research has been done in this field and has resulted in improvements that are highly considerable. The reason for the attention that this field has captured is its wide applications and necessity. Not only are face recognition systems an important part of the security and surveillance systems, these systems are needed for everyday tasks like automatic attendance systems and face locks. One main problem with these systems is, however, the computational cost-accuracy tradeoff. The systems that are capable of producing satisfactory results have a very high computational cost and are thus unusable in many scenarios. A robust system is thus needed that attains required results without producing high burden on the computational power of the system.

This research focuses on developing a face recognition system that is capable of attaining results comparable to the state-of-art systems while minimizing the computational requirements and is divided into two sections. Firstly, machine learning approach to design a system that uses landmarks and holistic features are utilized in combination to attain high results. This was accompanied with experimentation to propose voting strategy to attain higher results. The outcome of this section confirms the hypothesis that pose specific classification can indeed benefit in attaining improved results with common machine learning approaches.

Once verified, this research has been extended to deep learning-based approach. A framework was developed with a similar hypothesis in mind as above. In addition, a data augmentation strategy was devised and tested with the help of experimentation. The implemented strategy delivered an accuracy of 98.13, which is comparable to most of the state-of-the-art systems, while requiring only 856 MFLOPS.

Publications:

1. M. Asad, A. Hussain, and U. Mir, "Low complexity hybrid holistic – landmark based approach for face recognition." *Multimedia Tools and Applications*, vol. 80, no. 20, pp. 30199-30212, 2020, doi: 10.1007/s11042-020-08872-z.

Table of Contents

Chapter 1	Introduction	2
1.1	Biometric Systems	2
1.1.1	Advantages of Biometric Systems.....	3
1.2	Face Recognition.....	3
1.2.1	Face Recognition as a Biometric System	4
1.2.2	Importance of FR.....	5
1.2.3	Applications of Face Recognition	5
1.2.4	Categories of FR Systems.....	6
1.2.5	Generic FR System.....	7
1.3	Investigated Research Topics.....	8
1.4	Problem formulation	8
1.5	Motivations and Objectives.....	9
1.6	Problem Statement	9
1.7	Research Questions	10
1.8	Research Contributions	10
1.9	Dissertation Outline	10
Chapter 2	Literature Review	12
2.1	Traditional Face Recognition Techniques.....	12
2.1.1	Local Approaches.....	13
2.1.1.1	Techniques Based on Local Appearance.....	13
2.1.1.2	Techniques Based on Key Points	18
2.1.2	Holistic Approach.....	20
2.1.2.1	Linear Techniques	20

2.1.2.2	Nonlinear Techniques	24
2.1.3	Hybrid Approach	25
2.1.4	Limitations and Challenges in Traditional Techniques	29
2.2	Deep Learning-based Face Recognition Techniques	30
2.2.1	Advantages and challenges in deep learning	32
2.3	Research Gaps	33
Chapter 3	Hybrid Holistic-Landmark Based Low Complexity System	35
3.1	Proposed Methodology	35
3.1.1	Overview	36
3.1.2	Landmark Estimation	36
3.1.3	Illumination Normalization	40
3.1.4	Feature Extraction	42
3.1.5	Classification	44
3.2	Experiments and Results	45
3.2.1	Database	45
3.2.2	<i>Simulation Setup</i>	46
3.2.3	Performance Measures	47
3.2.4	Results and Discussion	48
3.3	Chapter Summary	49
Chapter 4	Low Complexity CNN Based FR System	51
4.1	Background	51
4.1.1	Artificial Neural Networks	51
4.1.2	Convolutional Neural Networks	57
4.2	Work Methodology	59
4.2.1	Preprocessing	60
4.2.2	Classification	64

4.2.3	Simulation Setup	65
4.2.4	Performance Measures	66
4.2.5	Results	67
4.3	Chapter Summary.....	69
Chapter 5	Conclusions and Future Work	71
5.1	Conclusions	71
5.1.1	First Approach	71
5.1.2	Second Approach.....	72
5.2	Future Work	72
Chapter 6	References	75

Table of Figures

Figure 1.1: Types of biometric systems	3
Figure 1.2: Steps involved in finding facial matches	4
Figure 1.3: Categorization of FR assessment protocols	6
Figure 1.4: A generic FR system	7
Figure 2.1: Methods for face recognition	12
Figure 2.2: LBP working	14
Figure 2.3: 4f optical configuration	16
Figure 2.4: Illustration of dimensionality reduction using PCA	21
Figure 2.5: MM-DFR FR flowchart	27
Figure 3.1: Flow of proposed facial recognition system	35
Figure 3.2: Convolutional Experts Network	37
Figure 3.3: Implementation of Landmark extraction on the FR dataset (a) the original image, (b) the extracted landmarks, (c) the pose estimation results, and (d) image after ROI extraction.	39
Figure 3.4: Pitch, yaw, and roll values.	39
Figure 3.5: Steps involved in Tan and Triggs normalization algorithm	41
Figure 3.6: Random images from dataset	45
Figure 4.1: Working of a neuron	52
Figure 4.2: Feed forward propagation example	53
Figure 4.3: Graph of Sigmoid, Tanh, and Relu functions	56
Figure 4.4: Structure of CNN (Courtesy: https://www.mdpi.com)	57
Figure 4.5: Example of convolution operation	58
Figure 4.6: Pooling example	59
Figure 4.7: Fully-connected layer	59
Figure 4.8: Overall system workflow	60
Figure 4.9: Example of pitch, yaw and roll values	62
Figure 4.10: Working of RepMLP	64

List of Tables

Table 2.1:	FR using local appearance based techniques	17
Table 2.2:	FR using key points based techniques.....	19
Table 2.3:	FR using linear techniques	24
Table 2.4:	FR using non-linear techniques	25
Table 2.5:	FR using Hybrid approaches	28
Table 2.6:	FR using deep learning based techniques	32
Table 3.1:	BIWI Kinect dataset summary	46
Table 3.2:	Pose categories for classification	46
Table 3.3:	Bagging and proposed classifiers employed with TT and CLAHE approaches on frontal face images.	48
Table 3.4:	Result obtained for classification with and without pose estimation.	49
Table 4.1:	Popular activation functions	54
Table 4.2:	Datasets used for pose estimation algorithm	62
Table 4.3:	Classification datasets	63
Table 4.4:	Pose categories for classification before flipping images	66
Table 4.5:	Results on FaceScrub dataset	68

Chapter 1

Introduction

Introduction

To cope with demanding needs of the era, human beings depend on technology for vast majority of fields. Technological advancements can be leveraged in various fields such as medical science for detecting breast cancer, self-driving cars, or simulation systems, to transform significant ideas into useful applications. Facial recognition is one of the emerging technologies that has been gaining interest of the researchers for the last decades. It allows automatically identifying a person as he or she appears in front of the camera. This chapter introduces this technology along with background of the field. A more detailed technical background relevant to each module is presented in the corresponding chapters.

1.1 Biometric Systems

Biometric systems enable efficient recognition and authentication of a person with the help of unique biological characteristics. Biometric systems are based on two types of measurements: physiological and behavioral as shown in Fig. 1.1 [1]. Examples of physiological features include patterns captured from fingerprints, iris, facial images, DNA, urine, saliva, etc. Behavioral measurements, on the other hand involve voice patterns, signature dynamics, keystroke undercurrents, etc. Physiological measures are considered more stable over time when compared to behavioral ones.

Earlier, more basic usage of biometrics can be associated to the Babylonian empire at 500BC; however, first systematic record of authentication systems based on biometrics took place in Paris, France in the 1800s. In this system, Bertillon developed a practice where fixed body dimensions were used for the classification and comparison of prisoners. This system depended on distinctive biological features for verification of identity. During 1880s, fingerprints were used to identify prisoners and as signature of a person on agreements. During this era, it was well established that a person's identity can be portrayed by fingerprints and that one using it may be held accountable [1].

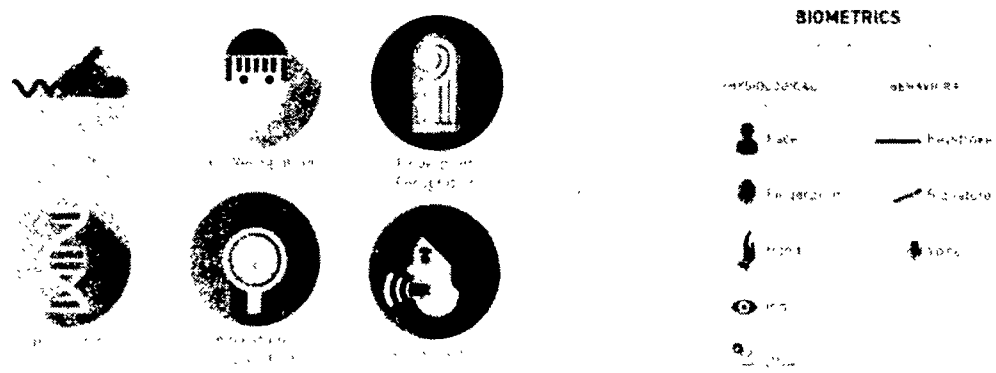


Figure 1.1: Types of biometric systems

1.1.1 Advantages of Biometric Systems

Biometric identification has certain benefits that make this method of recognition much superior than the traditional ones. Perhaps the first and foremost of these is that these techniques are more secure. Passwords may be hard to break, but the fact that these are formed and used outside a person leaves a possibility of copying. Moreover, placement of a password does not require presence of the person himself leaving a door open for misuse. Biometrics, in contrast, use features that are inherent to a person and thus provide a much higher level of security. The importance of security becomes more pronounced when a single password may need to be used numerous times. An average user in the US, for instance, connects up to 130 online servers to a single e-mail account. If needed, biometrics can be combined with manual inputs like passwords or draw patterns to ensure better level of security.

In addition to that the biometrics are easier to use. Depending on the method employed, the method may require minimal or no intervention from the individual. This adds to the fact that one does not have to memorize these as in the case of passwords.

1.2 Face Recognition

Facial Recognition (FR) systems allow people in a facial database to be automatically identified. These involve input from one or more cameras or archived videos/images. Video surveillance (VS) networks, for instance, entail use of many cameras and are an important part of a dependable decision-support system. A real-time apprise can be sent to an operator

when a person of interest approaches a camera or search archived video in video-to-video FR.

1.2.1 Face Recognition as a Biometric System

Face recognition, as the name suggests, uses facial features to identify a person. Like with all biometric systems, FR uses recorded features of known persons to be compared against the features captured during inference to recognize that person. This input may come from live

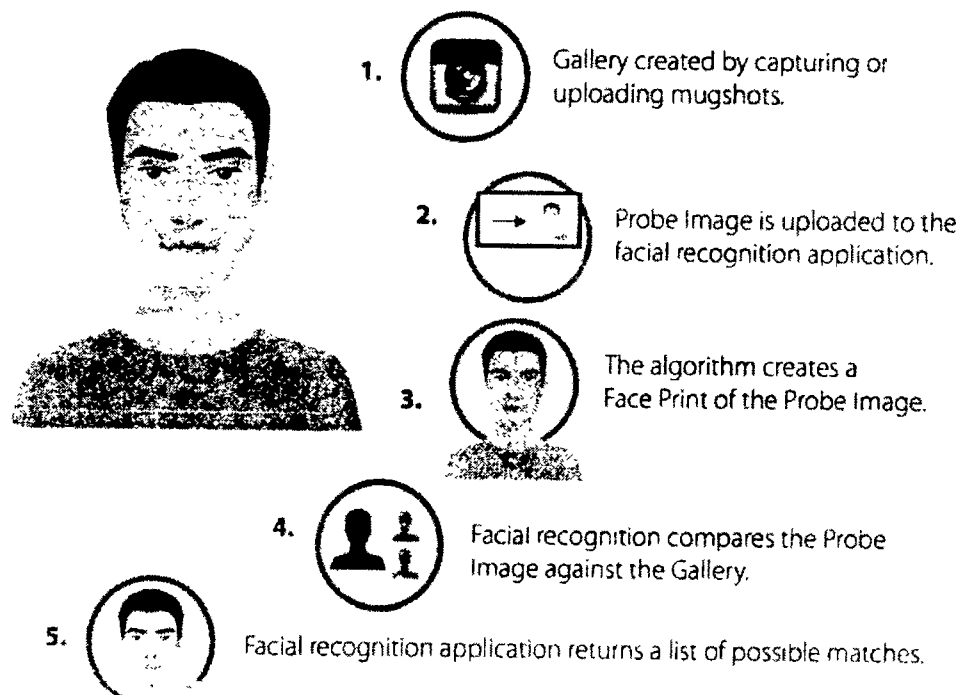


Figure 1.2: Steps involved in finding facial matches

input from a camera or an archived video or images. This method of biometric recognition has become more popular over time. Steps involved in FR are shown in Fig. 1.2 [1]. FR has been the subject of interest for neurologists, psychologists, and computer scientists [2]. It involves detecting, capturing, and comparing facial features using biometric machines or sensors.

1.2.2 Importance of FR

Due to this widespread growing usage of FR systems, the global market that this technology has grasped is worth billions of dollars annually. Some of the main reasons why this technique is gaining such importance are summarized below [2]:

- As face is the most common medium of identification, it is simpler for an operator to recognize a person reliably.
- Unlike other biometric systems like fingerprint identification, it does not require intrusion.
- Common cameras can be used for input rather than requiring specialized apparatus.
- Input can be acquired from a significant distance.

1.2.3 Applications of Face Recognition

Applications of Face Recognition range in vast number of settings [3] including security, automated attendance, crime investigation, banking, etc. One of the main applications of face recognition and verification technologies is access control. This application is becoming more popular over time. Face Recognition systems are useful for controlling access to personal gadgets, homes, vehicles, offices, and other locations. With this method, an advanced level of security can be expected where one can be sure that only he/she can access his device. This approach is also becoming popular among various banks where biometrics can be used with the ATM transactions instead of entering a pin code. Automated attendance of employees or students etc. is also being accomplished by FR systems. This provides us a convenient record of all entry and exit timings of each relevant person. A much stronger security and surveillance system can be ensured with this technology when compared to what humans can provide. A system may be set to trigger an alert if an unwanted person or any person beyond a given closed group is captured. Advantages of such systems include continuous surveillance and convenient updating of new persons' data into the system. Take for instance a new person has just joined a department. It is obviously difficult for every person in all offices to immediately be updated about this person. Centralized approach of many FR systems can solve this problem with great convenience. Likewise, FR systems can be used for forensics. Searching through data fetched from all cameras at a crime scene may be time taking. An FR-based system on the other hand may immediately provide information about which persons were present at the crime scene or whether a particular person was there.

This wide diversity of applications dictates different requirements from these scenarios adding to the challenges FR has to cope with. To elaborate, a crime investigation system may need to cope with faces from all poses while video surveillance requires low computational cost for fast execution to cope with high frame rates of videos. Thus, creating a generalized system has numerous challenges.

1.2.4 Categories of FR Systems

Face Recognition systems can broadly be characterized into two categories of watchlist screening and person reidentification [4]. Watchlist screening focuses on techniques to determine if facial image matches a specific person of interest. In contrast to watchlist screening, person reidentification system searches and identifies any of the persons in the database. Here, a system may assume faces to appear only from within the enrolled persons or have an ability to reject persons from outside its learned sets. The two types of person-reidentification systems are termed as close set and open set respectively [2,5]. Here, the term closed set refers to the applications where source and target sets have the same classes. In contrast, in an open set recognition system the overlap between source and target is either not present or limited to a few classes. The above-described categories have been summarized in the Fig. 1.3.

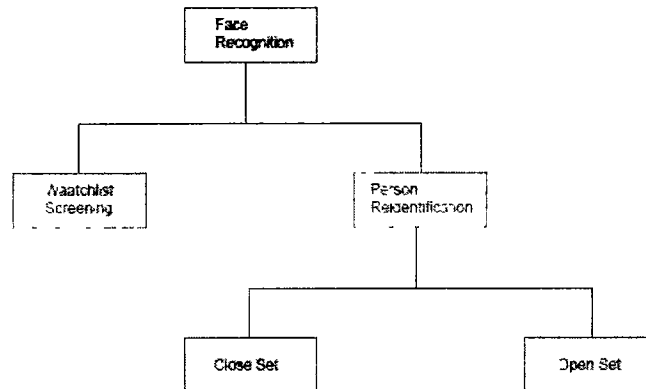


Figure 1.3: Categorization of FR assessment protocols

Though possible to achieve high accuracy for FR in a controlled environment; in a real-life environment, certain challenges like variances in illumination and pose as well as occlusions introduce difficulties that are hard to cope with. Face is a 3D object with different appearance from each pose yet faces of different persons have a significant similarity.

1.2.5 Generic FR System

This section depicts a generic FR system's flow in order to better clarify how FR systems work. As shown in Fig. 1.4, a FR system requires a number of processes. Following image capture, a pre-processing phase is used to ameliorate the quality of image by application of a transformation which results in a faster and/or more accurate categorization. Regions of Interest (ROIs) in a video frame or an image are then searched that correlate to the faces. Several features are extracted after this step, after which non-redundant and discriminant features are chosen. An input pattern $\mathbf{q} = (q_1, q_2, \dots, q_D)$ is then formed with the help of these features.

From the images that correspond to individual i , reference patterns $\mathbf{p}^i = (p^i_1, p^i_2, \dots, p^i_n)$ are then extracted, where n is the number of images used. These patterns are used to create a unique face model that gets maintained in a database. During operational mode, a classifier is trained, which enables mapping of new input patterns to the associated class defined previously. For recognition, a neural or statistical classifier or a template matcher is often utilized.

Patterns for input images of individuals are compared to the facial models of enrolled ones in operational mode. The probability that the pattern q correspond to an individual i are thus of the main concern, where $i = 1, 2, \dots, I$. Output of the system is the identity of the best matching individual.

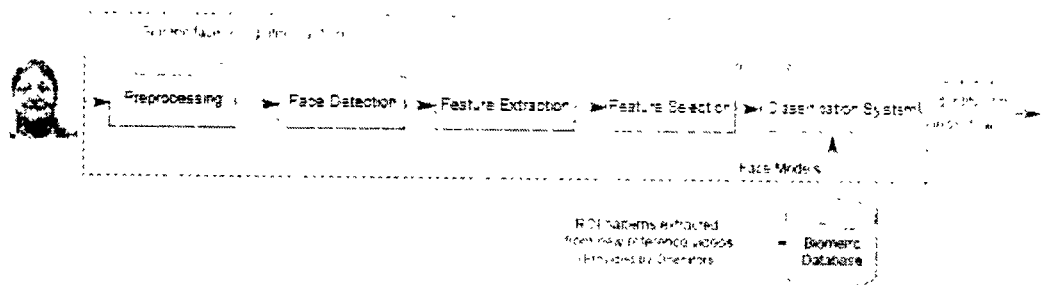


Figure 1.4: A generic FR system

Most Image Processing applications work with the approach used by the FR system mentioned above. One disadvantage is that this method does not automatically account for a

person's vast range of face appearance. Expecting a classifier to classify visibly diverse facial photos of an individual in the same class is unrealistic.

1.3 Investigated Research Topics

At the top level, this research investigates techniques to achieve high classification accuracy for face recognition and verification tasks while utilizing a very low computational power and memory. Among the top reasons that keep face recognition a complex task is illumination, pose and occlusions as these are responsible for intraclass variability. This research focuses the first two with the greater focus on the second.

In the first module, a holistic features-based approach was utilized where a pose specific classification architecture was proposed to assess against the commonly utilized technique without any use of this method. It is this module in which illumination normalization techniques have also been experimented with.

For illumination, a set of techniques have been tested and compared based on the accuracy of the overall results achieved. This needs to be done as the different normalization techniques may perform differently based on the scenario and the features being used. The techniques compared were carefully chosen considering what techniques are giving more promising results in the state of the art.

The main drawback faced in the first module was the requirement of sufficient images from all defined poses. This may prove cumbersome during enrollment. Moreover, with the recent success of CNN based approaches, it was decided to benefit from this more advanced technique. Thus, in the second module, a CNN based low computational cost approach was devised and investigated along with a data augmentation section responsible for generating synthetic images to overcome the problem in the first module.

1.4 Problem formulation

The goal of this thesis is to develop a face recognition system that can accurately identify and verify human faces in real-world scenarios. This work aims at developing a face recognition system that can accurately recognize faces in environments with lighting and pose variations while maintaining an acceptable computational cost – accuracy tradeoff. To achieve this, the FR problem can be formulated as follows:

Let x be the feature vector of the face to be identified and c_i be the class for individual i with ‘ C ’ the set of all the classes pertaining to every individual in the database, then closest facial model m to that of input image needs to be identified. Thus:

$$m = \arg \max_{c_i \in C} \text{Sim}(x, c_i) \quad (1.1)$$

where $\text{Sim}(x, c_i)$ is the similarity score of input facial model and the class of individual i .

1.5 Motivations and Objectives

Face recognition is a complex task with high intraclass variability and high interclass similarity. Face being a 3D object is not only visually different from different sides, other factors like illumination variations also contribute to intraclass variabilities [2]. On the other hand, faces of different persons from similar pose and external conditions have a significant similarity. Thus, face recognition proves to be a complex task to handle. The need of effective and efficient FR systems has become crucial. This is because of the vast applications of FR systems as this has entered numerous fields of life including personal applications like face lock, industrial applications like automated attendance systems and outdoor applications like surveillance [3]. Owing to these factors, much attention is given to this field by researchers. While recent research has shown some significant improvement, the computational cost of these systems is generally in billions of FLOPS. This limits the use of these high accuracy systems to limited applications and scenarios. Due to the number of FLOPS and parameters, these systems are able to achieve low frame rate. Additionally, systems with low computational power and memory cannot benefit from these systems. The primary aim of this work is thus:

- Devise a computationally cost-effective solution for face recognition without compromising accuracy.
- Suggest a solution that allows better recognition on wider pose angles.
- Investigate impact of controlling intraclass variability, particularly caused by pose variations

1.6 Problem Statement

FR techniques are needed to accurately recognize faces in unconstrained environments. Owing to their computational cost, current FR systems are not suited for real-time applications [6]. In addition, for large pose differences, performance of existing methods is

low in terms of computational cost and accuracy [7]. A computationally efficient FR system is required that can provide quick and reliable identification in demanding environments.

1.7 Research Questions

This research focuses on answering the following questions:

1. What will be the impact of controlling intraclass variability due to pose variation on face recognition technique?
2. Can comparable performance be achieved with face recognition technique having low computational cost in an unconstrained environment?

1.8 Research Contributions

The research provides two different implementations of face recognition focused on achieving high accuracy in low computational cost systems. In this work, the first implementation was done with conventional techniques of Machine Learning while in the next step this work proceeds towards a framework for deep-learning based approach. Below are the contributions of this work:

- Face Recognition Methodology is devised to handle intraclass variability caused by large pose variation in face poses.
- Low computational cost solution is proposed for face recognition without compromising accuracy.
- Our experimentations resulted in improvement in landmark estimation technique.
- A Zernike moment combination for FR is given, and experiments are carried out to identify the best illumination normalization method to be utilized with these moments.
- A novel voting set of classifiers has been suggested.
- Data driven AI technique is applied to improve accuracy of low-cost systems.

1.9 Dissertation Outline

Rest of the dissertation has been organized as follows. Chapter 2 describes the modern methods used in FR and lays out some of the prominent work done in each. Methodology of the work has been described in chapter 3 and chapter 4, where each of the two chapters describes a different framework to solve problems faced in FR systems. Overall conclusions of the work done in this thesis have been outlined in chapter 5.

Chapter 2:

Literature Review

Literature Review

Face recognition systems allow us to automatically identify and verify a persons’ identity with the help of their facial features. These systems can be broadly divided into two main subcategories, namely traditional and deep-learning based.

2.1 Traditional Face Recognition Techniques

Traditional face recognition systems use techniques such as feature extraction, normalization, and classification to recognize faces. Depending on the detection and recognition method (Fig. 2.1 [8]), these systems can be classified into three main categories [8]: (1) local approach; (2) holistic approach, and (3) hybrid approach. The first method classifies people based on specific facial traits rather than the entire face. In contrast, the second method uses the complete face as an input before projecting it onto a small subspace or in the correlation plane. Hybrid method improves facial recognition accuracy by combining local and global characteristics.

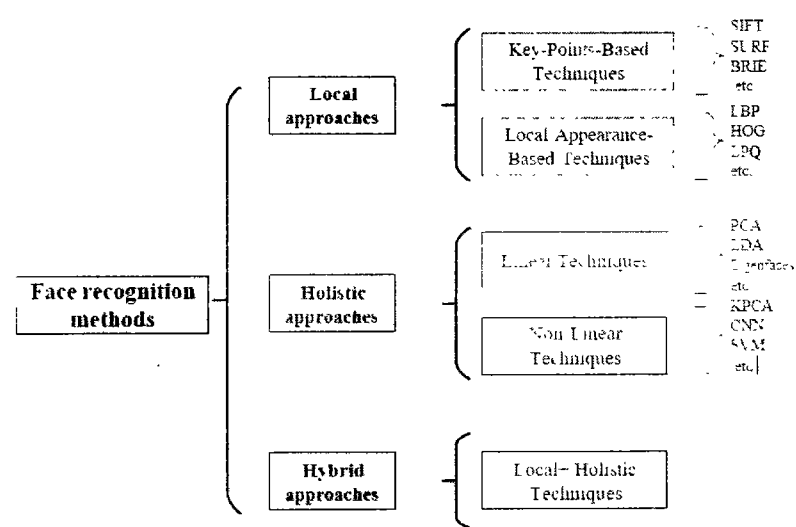


Figure 2.1: Methods for face recognition

2.1.1 Local Approaches

Local techniques to face recognition only address a subset of facial traits. They're more sensitive to face expressions, occlusions, and posture [8]. The basic goal of these approaches is to find distinguishing characteristics. These approaches can be divided into two subgroups: (1) To extract local features, techniques based on local appearance are applied, and the facial image is separated into small parts (patches) [9,10]. (2) With the aim of extracting the features centered on these places, key-points-based algorithms are utilized to detect the relevant points in the facial image.

2.1.1.1 Techniques Based on Local Appearance

These techniques are geometrical and are also referred to as a feature or analytic approach.

In this scenario, a set of different vectors with small areas (patches) or small dimensions represent the face image. To extract more information, techniques based on local appearance focus on crucial parts of the face like eyes, and lips etc. In this approach, pixel orientations, histograms [11, 12], geometric qualities, and correlation planes [9, 13] are used to define local features.

- LBP (local binary pattern) and variants: LBP is a versatile texture extraction technique that may be applied to any object [14]. FR [9], expression identification of faces and segmentation and classification of texture are just a few of the applications where it excels.

The LBP method initially separates the face picture into spatial arrays. Then, within each rectangular region of the array, a kernel of 3×3 pixels ($p_0 \dots p_8$) is convolved over the rectangular region. The center pixel (p_0) of the output is used as the threshold, using the value of the center pixel $i(p_0)$ as a basis for thresholding, to generate the binary code. If center pixel value is greater than neighboring pixels, it is given a zero; otherwise, it is assigned a value of one. The texture information of the current location is represented by the binary code. Finally, a histogram of these codes is constructed for each array square, and feature vector is created by concatenating histograms. As indicated in Equation 1, the LBP is given by a 3×3 matrix.

$$LBP = \sum_{p=0}^8 2^p s(i_0 - i_p), \text{ with } s_x = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (2.1)$$

where i_0 and i_p are the centre and neighbourhood pixel intensity values, respectively. The LBP technique is depicted in Fig. 2.2 [15] as a procedure.

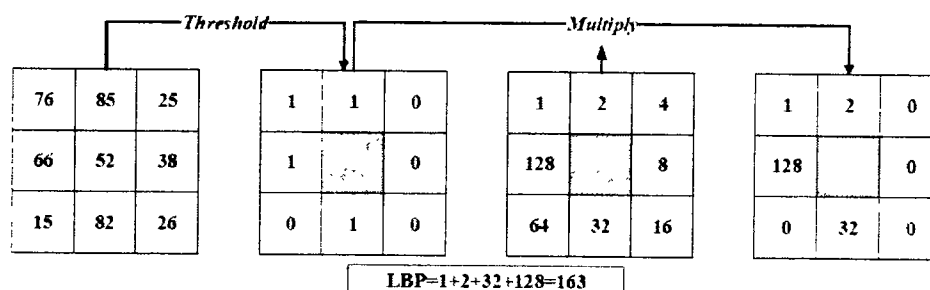


Figure 2.2: LBP working

Khoi et al. [16] offered a quick face identification method based on LBP, PLBP, and RI-LBP. To extract hierarchical representations of data, Xi et. al. [17] demonstrated the LBP network, a new unsupervised deep learning-based approach (LBPNet). The topology of the LBPNet is identical to that of the convolutional neural network (CNN). LBPNet is similar to other unsupervised approaches, according to experimental findings acquired using available benchmarks (e.g., LFW and FERET). Laure et al. [18] developed a system for resolving face recognition problems involving a wide range of characteristics like illumination, expression and various poses. This technique is based upon two different approaches: K-NN and LBP. Because of its invariance to target picture rotation, LBP has become among the most extensively used FR algorithms. Bonnen et al. [19] presented a "multiscale local binary pattern (MLBP)" a variant of the LBP technique for feature extraction. The local ternary pattern (LTP) technique [20] is another LBP extension that is less noise sensitive than the original LBP technique. To compute the differences between nearby pixels and the center pixel, this technique employs three phases. For face representation, Hussain et al. [21] designed a local quantized pattern (LQP) approach. LQP is a modified form of local pattern features that is light-resistant at its core. Using ternary split coding, the LQP features sample pixels from the neighboring ones and generates a binary codes pair. The quantization is based on each code utilizing a distinctly learned codebook.

- HOG (Historical Oriented Gradients) [22]: The HOG approach serves as one of the most useful descriptors to describe edge and shape. Using the distribution of edge direction or the gradient of light intensity, the HOG approach can characterize the shape of face. This method

works by dividing the entire face picture into small regions or areas, generating a histogram of pixel edge direction or gradients of direction for each of the cells. The facial image's feature is then extracted by merging the histograms of the complete cells. The HOG descriptor computes feature-vectors in the following way [11,12,23,24]: Calculate the amplitude of each cell's first-order gradients in both the horizontal and vertical directions after partitioning the local image into cells. Applying a 1D mask, $[-1 \ 0 \ 1]$, is the most typical way.

$$G_x(x, y) = I(x + 1, y) - I(x - 1, y) \quad (2.2)$$

$$G_y(x, y) = I(x, y + 1) - I(x, y - 1) \quad (2.3)$$

where $I(x, y)$ gives the pixel value of the point (x, y) , and $G_x(x, y)$ and $G_y(x, y)$ are the amplitudes of horizontal and vertical gradients, respectively. The gradient's magnitude and each pixel's (x, y) orientation are calculated as follows:

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \quad (2.4)$$

$$\theta(x, y) = \tan^{-1} \left(\frac{G_y(x, y)}{G_x(x, y)} \right) \quad (2.5)$$

With tri-linear interpolation, every pixel in the cell is voted into 9 bins based on the magnitude of the gradient and its orientation. Each cell's histograms are created pixel by pixel using direction gradients and the facial picture feature is created by concatenating the histograms of all cells. Karaaba et al. [22] suggested a robust face recognition method that the author terms as multi-HOG. This method utilizes on a mixture of various histograms of oriented gradients (HOG).

A vector of distances is created by the author between the reference and target facial images to aid with identification. Pyramid histogram of gradient (PHOG) descriptor and Laplacian filters were used by Arigbabu et al. [25] to suggest an innovative face recognition method. In addition, a support vector machine (SVM) with several kernel functions is employed to examine the face recognition problem.

- FR systems that focused correlation filter (CF) are shown to be resilient, accurate, efficient, and discriminative. Since the initial application of an optical correlator [26], correlation techniques have picked attention in the field of facial identification. High discrimination

ability, shift-invariance, desirable noise robustness, and intrinsic parallelism are all advantages of these approaches. Several optoelectronic hybrid correlation filter (CF) methods, like the joint transform correlator (JTC) [27] and VanderLugt correlator (VLC) [26], have been introduced as a result of these advantages. These methods are used to determine the level to which reference and the target images are similar. Correlation peak is used to make the decision. The "4f" optical setup is used in both approaches (VLC and JTC) [28]. Two convergent lenses generate this setup (Fig. 2.3 [31]). The fast Fourier transform (FFT) is used to process the facial image F , which comes from the first lens in the Fourier plane S_F . Using optoelectronic interfaces, a specific filter P (e.g., phase-only filter [29]) is applied in this Fourier plane. Finally, with the 2nd lens on output plane, the inverse fast Fourier transform (IFFT) is performed to generate the filtered face image \hat{F} .

The VLC approach, for example, is implemented using two Fourier transform structures in cascade resulting from two lenses [30], as seen in Fig. 2.3 [31]. The following is a description of the VLC technique: To obtain a target spectrum S , a 2D-FFT is first performed on image of interest. After that, the target spectrum is multiplied by the filter obtained with the 2D-FFT of a reference image, and the result is placed in the Fourier plane. Result of the correlation is then recorded on the correlation plane, with inverse FF affecting this multiplication.

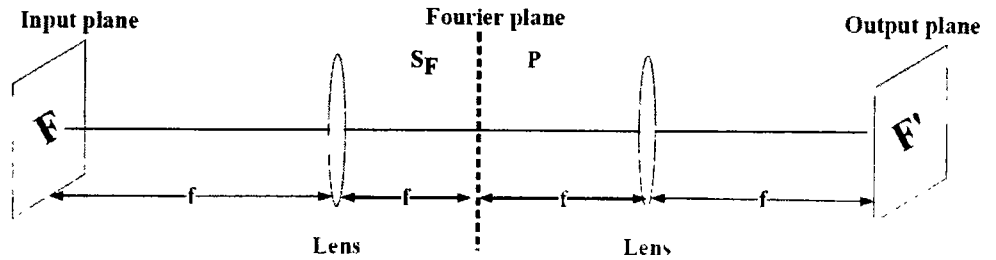


Figure 2.3: 4f optical configuration

The peak intensity of the correlation result is used to obtain the level of similarity among the reference image and the target image.

$$C = FFT^{-1}\{S^* \circ POF\} \quad (2.6)$$

The inverse fast FT (FFT) operation is represented by FFT^{-1} , the conjugate operation is represented by $*$, and element-wise array multiplication is represented by \circ . A phase-only (POF) filter was proposed by Horner and Gianino [32] to improve the matching process.

Correlation peaks with better discriminating abilities can be produced by the POF filter. POF filter can be defined as:

$$H_{POF}(u, v) = \frac{S^*(u, v)}{|S^*(u, v)|'} \quad (2.7)$$

where $S^*(u, v)$ is the complex conjugate of the reference image's 2D-FFT. By definition, the peak to correlation energy (PCE) is the energy in the intensity of the correlation peaks normalized to the total energy of the correlation plane in order to evaluate the decision.

$$PCE = \frac{\sum_{i,j}^N E_{peak}(i, j)}{\sum_{i,j}^M E_{correlation-plane}(i, j)} \quad (2.8)$$

where i and j are the coordinates of the coefficient, M is the correlation plane and N is the peak size of the correlation spot. The correlation peaks energy is E_{peak} while the total correlation plane energy is $E_{correlation-plane}$. Recognition and identification applications frequently use correlation approaches [28, 30, 33-36]. Performances of this approach based on the "4f" identifying configuration utilizing Geforce 8400 GS GPU from Nvidia, for example, were described in [30]. The decision is made using the POF filter. Leonard et al. [33] gives yet another essential piece of study in this field. Several specialized filters such as POF and BPOF etc., are employed to choose the best filter depending on its noise, scale and rotation sensitivity. Napoléon et al. [9] proposed a novel technique for fields of verification and identification that is based on improved 3D modelling under various lighting circumstances and enables for the reconstruction of faces in various positions. To boost efficiency of a correlation technique, the LBP descriptor is employed under diverse illumination situations. The VanderLugt correlator is developed to perform the identification. The tests are carried out on the PHPID database, with pitch value ranging from -30° to $+30^\circ$.

Table 2.1: FR using local appearance based techniques

Author	Date	Method	Matching	Results
Khoi et. al. [16]	2016	LBP	MAP	5% (TDF), 13.03% (CF1999), 90.95%(LFW)
Xi, Meng, et. al. [17]	2016	LBPNet	Cosine similarity	97.80% (FERET), 94.04% (LFW)

C. Guo [18]	2017	LBP and KNN	KNN	85.71% (LFW), 99.26% (CMU-PIE)
K. Bonnen et. al [19]	2013	MRF and MLBP	Cosine similarity	86.10% (AR), 95%(FERET)
J. Ren et. al. [20]	2013	Relaxed LTP	Chisquare distance	95.75% (CMU-PIE), 98.71% (Yale B)
M. Karaaba et. al [22]	2003	HOG and MMD	MMD/MLPD	68.59% (FERET), 23.49%(LFW)
O. A. Arigbabu et. al. [25]	2010	PHOG and SVM	SVM	88.50% (LFW)
Leonard et. al. [33]	2018	VLC correlator	ASPOF	92% (PHPID)

2.1.1.2 Techniques Based on Key Points

Using some geometric features of the surface of face, approaches based on key-points are utilized to recognize certain geometric features. Detection of key-point and extraction of features are two important steps in these approaches [9,37-39]. The first step focuses on how well the face image detectors in the key-point elements operate. In the second stage, the information conveyed by the face image's key-point features is represented. SURF, SIFT and BRIEF techniques are extensively employed to characterize the features of the facial image, despite the fact that these approaches can fix missing portions and occlusions.

- SIFT (scale invariant feature transform) [40, 41]: SIFT is an algorithm for detecting and describing the image's local features. This method is commonly utilized to link two images together using their local descriptors containing information that allows them to be matched. The SIFT descriptor's main goal is to turn an image into a point-of-interest-based representation. The key information of the face is contained in these points. SIFT has rotation and scale invariance. It is widely used nowadays and is quick, which is important in applications that require immediate response; however, it has a drawback of the time it takes to match the vital spots. The algorithm consists of four steps: (1) detection of the space-maximum scales and minimum points, (2) finding of characteristic points, (3) orientation assignment, and (4) a characteristic point descriptor. Furthermore, the SIFT descriptor is better suited to describing faces that are changed by lighting, scale, translation, and rotation

[41]. Using integral images, SURF seeks to determine the highest possible estimate of the Hessian matrix to substantially minimize the computing time for processing

- Binary robust independent elementary features (BRIEF) [37,41]: In terms of assessment, the BRIEF descriptor is based on changes in pixel intensity that are similar to the binary descriptors binary robust invariant scalable (BRISK) and fast retina key point (FREAK). The BRIEF description makes smooth the image patches to reduce noise. After that, the descriptor is represented by the differences in pixel intensity. In pattern recognition, this descriptor is considered to have highest accuracy.

Some of the methods for FR using Surf, SIFT and BRIEF approaches are discussed here. L. Lenc et. al. [40] uses SIFT based Kepenekci automatic face recognition system that is shown to outperform the original Kepenekci approach. It also proposes a new corpus creation algorithm and two novel supervised confidence measure methods to identify incorrectly recognized faces, which are very efficient for the task. Du et al. [42] proposes to use the Speed-Up Robust Features (SURF) detector and descriptor in face recognition, which is scale and in-plane rotation invariant and has comparable or better performance than the Scale Invariant Feature Transform (SIFT). The proposal is based on the fact that SURF is faster due to its 64-dimensional feature size and indexing scheme, which makes it suitable for matching. Vinay et al. [43] proposes two novel detector-descriptor variants (SURF detector with SIFT descriptor and SIFT detector with SURF descriptor) to improve the efficacy of contemporary face recognition systems that use SIFT and SURF algorithms. The proposed techniques are evaluated on standard benchmarks and compared with classical SIFT and SURF algorithms. Calonder et al. [44] proposes a new binary descriptor called BRIEF that can be computed directly based on simple intensity difference tests, resulting in a very fast feature extraction and matching process, with comparable recognition accuracy to SIFT and SURF.

Table 2.2: FR using key points based techniques

Author	Date	Method	Matching	Results
L. Lenc et. al. [40]	2015	SIFT	<i>a posterior</i> probability	97.30% (FERET), 95.80%(AR), 98.04%(LFW)
G. Du et. al. [42]	2015	SURF	FLANN distance	95.60%(LFW)

Vinay et al. [43]	2009	SURF + SIFT	FLANN distance	78.86%(LFW), 96.67% (Face94)
M. Calonder et al. [44]	2011	BRIEF	KNN	48%

2.1.2 Holistic Approach

The goal of holistic or subspace approaches is to process the complete face without removing any facial regions or feature points (eyes, nose, mouth and eyes etc.). These techniques' main goal is to turn the facial image into a pixel matrix, which is subsequently converted into feature vectors to facilitate treatment. These feature vectors are then applied in low-dimensional space. On the other hand, holistic or subspace approaches have become popular because of their resilience to changes (illumination, facial expression, and pose). Furthermore, depending on the technique employed to depict the subspace, these approaches can be classified as linear or non-linear strategies.

2.1.2.1 Linear Techniques

Eigenfaces (principal component analysis; PCA), Fisherfaces (linear discriminative analysis; LDA), and independent component analysis (ICA) are the most prominent linear algorithms for face identification systems.

- Principal component analysis (PCA) [45, 46] and Eigenface [47]: Eigenfaces is a prominent technique that uses holistic methods to extract feature points from a face image. The basic components obtained using the PCA technique are used to construct Eigenfaces or face templates. The PCA approach condenses a vast number of possibly related variables into a reduced number of "principal components." PCA reduces the data space's large dimensionality (observed variables) to the feature space's reduced intrinsic dimensionality (independent variables), which is required to economically characterize the data. This has been demonstrated in Fig. 2.4 [44]. The Eigenvectors of the covariance matrix are calculated by PCA, and large eigenvectors are used to project the data onto a feature space of lower dimensionality. The Eigenvectors produced by PCA are referred to as Eigenfaces in face representation and recognition.

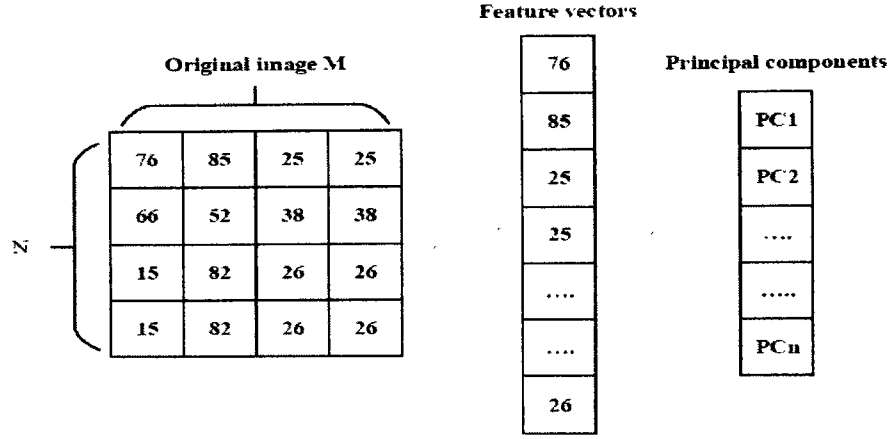


Figure 2.4: Illustration of dimensionality reduction using PCA

A vector of dimension $M \times N$ can also be used to consider an image, thus a typical image of 44 pixels creates a vector of 16 pixels. Let the training set of photos be $\{X_1, X_2, X_3 \dots X_N\}$. The following is the definition of the set's average face:

$$X = \frac{1}{N} \sum_{i=1}^N X_i \quad (2.9)$$

Calculate the estimated covariance matrix to reflect the dispersion degree of all feature vectors linked with the average vector. The following is the definition of the covariance matrix Q :

$$Q = \frac{1}{N} \sum_{i=1}^N (X - X_i)(X - X_i)^T \quad (2.10)$$

The Eigenvectors and their accompanying Eigen-values are calculated with the help of following equation:

$$CV = \lambda V, \quad V \in R_n, V \neq 0 \quad (2.11)$$

where V denotes the collection of the matrix Q of the eigenvectors and its eigenvalue. Project i_{th} person's training photos to the associated Eigen-subspace:

$$y_k^i = w^T x_i, \quad (i = 1, 2, 3, \dots, N) \quad (2.12)$$

The primary components, also referred to as eigenfaces, are the projections of x and are called y_k^i . The facial images are created by combining the "primary components" of these

vectors in a linear fashion. To categorize facial features, neural networks and wavelet fusion are used. For evaluation, the ORL [46] database is employed.

- Linear discriminative analysis (LDA) and Fisherface [49, 51]: The Fisherface approach uses the same similarity premise used by Eigenfaces. By applying LDA approaches rather than PCA techniques, this strategy aims to minimize the high-dimensional picture space. The LDA technique is often utilized to reduce dimensionality and recognize faces [51]. A key difference between PCA and LDA is that PCA works in unsupervised fashion while LDA is a supervised approach that makes use of data. For each sample of each class the within-class and between-class scatter matrices S_W and S_B are defined as:

$$S_B = \sum_{i=1}^c M_i (x_i - \mu)(x_i - \mu)^T \quad (2.13)$$

$$S_W = \sum_{i=1}^c \sum_{x_k \in X_i} M_i (x_k - \mu)(x_k - \mu)^T \quad (2.14)$$

where c is the number of unique classes, M_i is the number of training samples in class i and X_i represents the set of samples belonging to class i with x_k being the number image of that class and μ is the mean vector of the samples in i . In relation to the total average for all classes, S_B describes the dispersion of attributes, while each face class's spread of characteristics around the mean is described by S_W . The goal is to maximize the $\det|S_B|/\det|S_W|$ ratio, i.e., to minimise S_W while increasing S_B .

- Independent component analysis (ICA) [52]: The ICA approach is used to compute a space's basic vectors. This approach enables autonomous component analysis by reducing statistical dependence between the several underlying vectors using a linear transformation. These are observed being not orthogonal to one another. In addition, since ICA acquires images within variables that are statistically independent, it is feasible to attain improved efficiency by acquiring images from several sources in uncorrelated variables.

- PCA, LDA, and ICA improvements: Different sorts of studies are being conducted in order to improve linear subspace algorithms. For facial region extraction and dealing with variance in noise, Z. Cui et al. [53] presented a novel technique known as spatial face region descriptor (SFRD) technique. The following is a description of this method: The reconstruction coefficients are sum-pooled over the patches within each region and each facial image is divided into numerous spatial regions, from which token-frequency (TF) features are extracted. Finally, in the leading eigenvectors, the dimensionality of the features is reduced

and the noise removed. by using a variation of PCA called “whitened principal component analysis (WPCA)” to extract the SFRD for face pictures. Furthermore, the authors in [54] suggested an LDA variant referred to as Probabilistic LDA (PLDA) to find directions in space with the highest distinguishability, which are best for both face recognition for frontal pose and under changing poses.

- Gabor filters: A Gaussian window locates spatial sinusoids in Gabor filters allowing us to determine the frequency, orientation, and scale of features in images. In the work by [55], Gabor filters are adjusted based on the shape as well as position of the object to estimate the feature vectors of an image of face paired with PCA to improve performance in unconstrained contexts for face identification. The Gabor features are subjected to PCA to eliminate redundancies and provide the finest face image description. Lastly, the similarity is assessed using the cosine measure.
- Frequency domain analyses [56,57]: Frequency domain analysis approaches provide a representation of the human faces based on low-frequency constituents with high energy. The DFT, DCT, and DWT algorithms are independent of data and so involve no training.
- Discrete wavelet transform (DWT): DWT is another linear FR algorithm. The authors of [56] employed a new patch approach and a 2D discrete wavelet transform (2D-DWT) methods for recognition of faces. Two highest-frequency sub-bands of 2D-DWT were constructed using an approach for integral projection using the average image from all training samples, a non-uniform patch strategy for the top-low-frequency level's sub-band is provided. This patch approach is better for preserving the integrity of local information and for reflecting the face image's structural feature. The outcome is determined by the neighbor classifier while designing the patching method that employs the training and testing samples.
- Discrete cosine transform (DCT) [59] can be employed in local and the global FR settings. DCT is a modification that signifies a finite block of data by summing a series of cosine operations that oscillate at various frequencies. This methodology is commonly employed in FR algorithms [57].

Subspace or holistic approaches are unable to fully capture the characteristics of geometric alterations of face images because of their constraints in retaining linearity in face identification. In case of linear data structures, linear approaches provide an acceptable

representation of facial images. Many sorts of research, however, employ kernels to generate a huge space to impose linearity on the problem when the facial picture data structures are non-linear.

Table 2.3: FR using linear techniques

Author	Date	Method	Matching	Results
H. J. Seo et. al. [45]	2011	LARK and PCA	L2 distance	85.10%(LFW)
Z. Cui et. al. [53]	2013	BoW	ASM	99.43%(AR), 99.50%(ORL), 82.30%(FERET)
Peng Li et. al. [54]	2012	PLDA	Accuracy	90.10%(LFW)
V. Perlibakas et. al. [55]	2004	PCA and Gabor filter	Cosine metric	87.77%(FERET)
Z.-H. Huang et. al. [56]	2015	2D-DWT	KNN	90.63%(FERET), 97.10%(LFW)
Z. Sufyanu et. al. [57]	2015	DCT	NCC	93.40%(Yale)

2.1.2.2 Nonlinear Techniques

- Kernel PCA (KPCA) [58]: is a nonlinear variant of PCA method. This method adds nonlinearity with the help of nonlinear kernel function like RBF, polynomial. KPCA computes the Eigenvectors of the kernel matrix, while the covariance matrix is computed via PCA. Furthermore, KPCA is a depiction of the PCA approach upon feature space of higher-dimensionality transformed by a kernel function. The KPCA technique estimates the function of the kernel K of a distribution with n data points $x_i \in R^d$ in three important steps, The data points are then translated onto a high-dimensional feature space called F .

The performance of the KPCA approach is highly affected by the selection of kernel matrix. The Gaussian or polynomial kernel is a linear kernel that is commonly utilized. The use of KPCA has proven successful in identifying originality [59] and recognizing speech [46].

- Kernel linear discriminant analysis (KLDA) [62]: KLDA extends the linear LDA technique by applying kernels, similar to how KPCA extends linear PCA. Based on spectral regression

kernel discriminant analysis, Arashloo et al. [60] suggested binary class-specific kernel discriminant analysis classifier (CS-KDA) which was nonlinear in nature.

- SVM is a supervised learning algorithm that uses data to classify and predict outcomes. It has the advantage of working well across a wide area. FR using SVM is possible after facial feature extraction [61], and it has been shown to produce better results when a large number of data sets are chosen directly during training. The Least Square SVM (LS-SVM) [62] is a popular extension of the (SVM) that has been successfully used for FR. This has the benefit of quick computation, speed, and a high recognition rate.

- Hu, S. et. al. [63] proposes illumination processing technique for face recognition termed as AHFSVD-face that uses singular value decomposition (SVD) bases and nonlinear singular values to obtain high-frequency facial features. The technique is shown to perform well under different illumination variations on various face databases.

Table 2.4: FR using non-linear techniques

Author	Date	Method	Matching	Results
S. Azuan et. al. [58]	2009	KPCA	Accuracy	81.5%(Yale), 89.6%(ORL)
S. R. Arashloo et. al [60]	2014	CS-KDA	Accuracy	95.89%(LFW)
Jianhong et. al. [62]	2009	LS-SVM	Correct Recognition	96%(ORL)
Hu, S. et. el. [63]	2017	AHFSVD-Face	Accuracy	95.8% (CMU PIE), 52%(LFW)

2.1.3 Hybrid Approach

To draw from the advantages from both subspace as well as local procedures, hybrid approaches were built on local and subspace properties. These provide great possibilities to ameliorate the performance of face recognition.

- Gabor wavelet and linear discriminant analysis (GW-LDA): Fathima et al. [64] suggested a hybrid strategy known as HGWLDA for FR that combined Gabor wavelet with LDA. This technique involves approximation and shrinking the size of the grayscale image. The authors

used a combination of Gabor filters with variable scales and orientations to convolve the grayscale face image. The inter-class space is then maximized while the intra-class space is reduced using the 2D-LDA subspace approach. The k-nearest neighbor (k-NN) classifier is used to categorize and recognize the test face image. Features for test image of the face are compared to each feature of the training set to perform recognition.

- Over-complete LBP (OCLBP), LDA, and WCCN: Barkan et al. [65] presented a novel face image representation derived using OCLBP, LDA, and WCCN. The LBP approach is employed in a multi-scale modified version in this representation. To minimize the high dimensionality representations, the LDA approach is used. Finally, face recognition uses a metric learning technique that is within class covariance normalization (WCCN).
- Advanced correlation filters and Walsh LBP (WLBP): Juefei et al. [66] used high-dimensional Walsh LBP to develop a single-sample periocular-based alignment-robust face recognition algorithm (WLBP). This technique requires only a single sample for each subject class, thanks to the 3D generic elastic model, that is both precise and efficient to create new face images using a large variety of rotations in three dimensions. The proposed solution outperforms advanced algorithms with a high accuracy of 89.69 percent under four assessment protocols using the LFW database.
- SIFT features, Fisher vectors, and PCA: Using the SIFT descriptor and Fisher vectors, Simonyan et al. [49] created a unique approach for face recognition. Owing the large dimensionality of the Fisher vectors, a class - specific dimensionality reduction is proposed by the authors. Following that, these vectors were linearly projected into a subspace of lower dimensionality.

The goal of this methodology is to characterize a face image using dense Fisher vectors-based encoding and SIFT features in both controlled and unrestricted settings to obtain good performance on the trying LFW dataset.

- CNNs and stacked auto-encoder (SAE): Ding et al. [68] suggested a framework for multimodal deep face representation (MM-DFR) using CNNs from the source face image, produced forward looking image of face by 3D model of the face, and evenly sampled patches of the images. In the suggested MM-DFR approach, a CNN based method is employed to extract the facial features, and an SAE method utilizing three-layer is used to

reduce the deep features with high dimensionality into a condensed facial signature. To do so, the LFW database is employed. MM-DFR's performance is evaluated for identification. Fig. 2.5 [81] depicts the proposed MM-DFR framework's flowchart.

- **PCA and ANFIS:** Sharma et al. [69] proposed a technique in which PCA and ANFIS were used to create a pose-invariant face recognition system. To extract the features of a picture, the PCA technique is used, and the ANFIS classifier is designed to identify people in a variety of poses. For the face recognition problem, the suggested PCA–ANFIS based system performs better than LDA–ANFIS and ICA–ANFIS. For evaluation, the ORL database was employed.

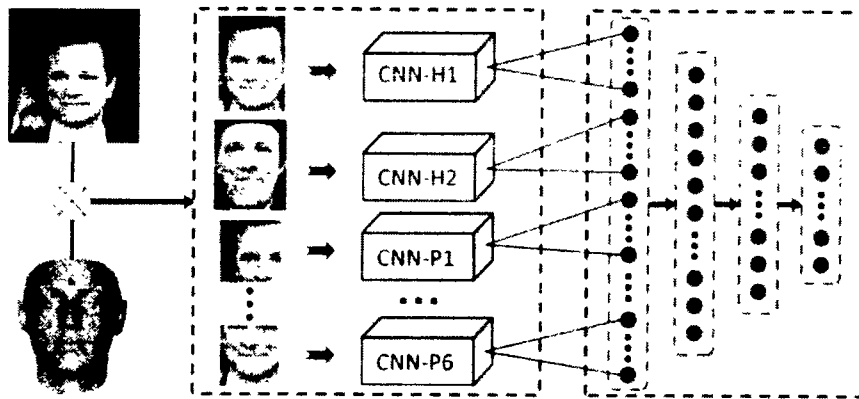


Figure 2.5: MM-DFR FR flowchart

- **DCT and PCA:** Moussa et al. [70] used DCT and PCA techniques to create a rapid face recognition system. To extract facial features, a genetic algorithm (GA) technique was utilized, which allows for the removal of unnecessary features and a reduction in the number of features. In addition, to extract features and reduce dimensionality, the DCT–PCA technique was applied. The decision is made using the minimal Euclidian distance (ED) as a method for measure. The efficiency of this technique is demonstrated using a variety of face databases.
- **PCA, SIFT, and iterative closest point (ICP):** Mian et al. [71] offer a hybrid matching-based Face recognition system that is multimodal (2D and 3D) and is efficient and resilient against face expressions. It utilizes Hotelling transform to automatically fix the posture of a 3D face based on its texture. Afterward, a unique three-dimensional spherical face representation (SFR) in combination with the SIFT descriptors is utilized to create a rejection classifier,

which, by deleting the faces of a vast number of candidates, allows for quick recognition in enormous galleries. A modified iterative closest point (ICP) technique is used to make the decision. While testing on complete FRGC v2 database, this system achieved a verification and identification rates of 98.6 percent and 96.1 percent respectively, indicating its lower sensitivity and resilience to facial expressions.

- **GABOR wavelets, PCA, and local Gabor binary pattern histogram sequence (LGBPHS):** Cho et al. [72] suggested a hybrid FR system with high computational efficiency that takes into account both global and local information. To minimize dimensionality, the PCA approach is utilized. The local Gabor binary pattern histogram sequence (LGBPHS) technique is then used to conclude the recognition step, which aims to lower the computational cost caused by Gabor filters. Under different lighting conditions, the experimental findings demonstrate a higher recognition rate than the PCA and Gabor wavelet approaches. To demonstrate the efficiency of this method, the Extended Yale-B Face Database has been employed.

- **PCA and Fisher linear discriminant (FLD) [73, 74]** A hybrid face recognition as well as representation technique was presented by Sing et al. [74] that takes advantage of both local and subspace features. To obtain the local features from the entire image, it is divided into smaller sections; however, to extract the global characteristics, the entire image is used. The merged feature vector is then subjected to PCA and Fisher linear discriminant (FLD) algorithms to reduce dimensionality. The evaluation makes use of the CMU-PIE [75], FERET [76], and AR face [77] databases.

- **SPCA-KNN [78]:** A SIFT, PCA, and KNN features based novel face recognition method was proposed by Kamencay et al. [78]. The local features are extracted using the Hessian-Laplace detector and the SPCA descriptor. The SPCA is used to recognize human faces. Based on the learned features, the KNN classifier is utilised to discover the nearest faces. The experiment's results show that the unsegmented and segmented ESSEX database has a recognition rate of 92 percent and 96 percent respectively with 700 training images.

Table 2.5: FR using Hybrid approaches

Author	Date	Method	Matching	Results
A. A. Fathima	2015	GW-LDA	KNN	88%(AT&T),

et. al. [64]				94.02%(FACES94), 88.12%(MITINDIA)
O. Barkan [65]	2013	LDA, OCLBP, and WCCN	WCCN	87.85%(LFW)
F. Juefei-Xu [66]	2015	ACF and WLBP	-	89.69%(LFW)
K. Simonyan et. al. [49]	2013	Fisher + SIFT	Mahalanobis matrix	87.47%(LFW)
R. Sharma et. al. [69]	2015	PCA-ANFIS, ICA-ANFIS, LDA-ANFIS	ANFIS	ORL (96.66%, 71.30%, 68%)
MOUSSA et. al. [70]	2018	DCT-PCA	Euclidian distance	92.62%(ORL), 99.40%(UMIST), 95.50%(YALE)
A. Mian et. al. [71]	2007	Hotelling transform, SIFT, and ICP	ICP	99.74%(FRGC)
H. Cho et. al [72]	2014	PCA-GABOR Wavelets	Bhattacharyya distance	95%(Extended Yale Face)
J. K. Sing et. al. [74]	2012	PCA-FLD	SVM	71.98%(CMU), 94.73%(FERET), 68.65%(AR)
P Kamencay et. al. [78]	2012	SPCA-KNN	KNN	96.80%(ESSEX)

2.1.4 Limitations and Challenges in Traditional Techniques

Traditional face recognition techniques have played a crucial role in the development of the field. However, they come with certain limitations and challenges that have hindered their effectiveness in various scenarios.

Sensitivity to variations: Traditional techniques are sensitive to changes in lighting, pose, and facial expressions. These variations can significantly impact the recognition accuracy, making it difficult to achieve reliable results in real-world applications

Feature extraction: Classical algorithms rely on manual feature extraction, which can be time-consuming and prone to errors. This process requires domain expertise and may not always capture the most discriminative features for accurate recognition

Scalability: Traditional methods often struggle with scalability, as they may not perform well when dealing with large datasets or multiple subjects. This limitation can hinder their applicability in large-scale applications, such as surveillance or social media analysis

Occlusion and disguise: Traditional techniques are generally not robust against occlusion or disguise, which can significantly affect recognition performance. Partially occluded faces or those with accessories like glasses, hats, or masks can pose challenges for these methods

Computational cost: Some traditional face recognition algorithms can be computationally expensive, making them unsuitable for real-time applications or resource-constrained environments

Holistic vs. local approaches: Traditional techniques can be broadly categorized into holistic and local approaches. Holistic methods, which consider the entire face as input, may not be robust against local variations or occlusions. On the other hand, local approaches, which focus on specific facial features, may not capture the global structure of the face, leading to reduced recognition performance

Despite these limitations, traditional face recognition techniques have laid the foundation for the development of more advanced methods, such as deep learning-based approaches. These modern techniques have shown significant improvements in addressing the challenges faced by traditional methods, offering better recognition performance, robustness against variations, and scalability for large-scale applications. However, deep learning techniques also come with their own set of challenges, such as the need for large annotated datasets, computational cost, and the interpretability of the learned features.

2.2 Deep Learning-based Face Recognition Techniques

Deep learning-based face recognition systems use deep neural networks to learn representations of facial features directly from the input images. It employs successive hidden-layers to process information, hierarchically organized for feature learning and representation or pattern classification [79]. Overall, deep learning-based face recognition systems have surpassed traditional systems in terms of accuracy and robustness, but they also

require large amounts of labeled training data and computational resources. Some of the recent deep-learning based techniques and their results have been discussed below. These results have been summarized in Table 2.6.

MobileFaceNet [80] was designed specifically to perform with high accuracy in real time on machines with low computational power. It builds on some other model with high discriminative ability to perform face recognition with more prominent distinction between faces of different persons. The activation function used in this model is PReLU.

VGG-Face [81] and FaceNet [82] are among the first to explore benefits of stacking up multiple convolutional layers in depth as well as width of these layers. VGG-Face is a 16-layer network with 3 fully connected layers at the end of the network. It takes as input a 224x224 RGB image and outputs a 4096-dimensional feature vector. It has been shown to have a very high discriminative ability.

FaceNet is a well-known technique in the face recognition community that uses a triplet loss function-based approach to learn Euclidean embedding for each image using deep CNN. The approach has the advantage of not requiring any 2D or 3D alignments other than scale and translation. The datasets LFW [83] and YTF [84] were employed, with the highest accuracies of 99.63 ± 0.09 and 95.12 ± 0.39 , respectively.

VGG-Face 2 [85] incorporates a range of design improvements over its predecessor, including the use of batch normalization, smaller filter sizes, and residual connections. It encompasses 34 layers compared to the original vgg-face network and achieves better results.

ArcFace [86] is yet another intriguing approach. By replacing Softmax Loss with Angular Margin Loss, it provides a similarity learning method that allows distance metric learning to be addressed in the classification problem. The cosine distance is used to compute the distance between two faces. If the two vectors are equal, $\theta = 0$ and $\cos\theta = 1$ will be returned. $\theta = \pi/2$ and $\cos\theta = 0$ if they are orthogonal. As a result, it can be utilized as a similarity indicator.

Like MobileFaceNet, Vargfacenet [87] utilizes the PReLU activation function. The major difference between these two networks is that MobileFaceNet uses downsampling at the beginning of the network, while the Vargfacenet removes downsampling to better preserve information in the images.

Table 2 6: FR using deep learning based techniques

Method	Year	FLOPS(G)	Dataset	Measure	Result (%)
MobileFaceNetV1 [80]	2018	1.1	LFW	Accuracy	99.55
			Megaface	TAR 10^{-6} FAR	92.59
VGG-Face [81]	2015	15G	LFW, YTF	Accuracy	98.95, 97.3
FaceNet [82]	2017	1.6G	LFW, YTF	Accuracy	99.63, 95.12
VGG-Face 2 [85]	2018	4G	IJB-A, IJB-B	Rank-1	98.2, 90.2
ArcFace [86]	2019	24.2G	YTF, LFW, CALFW, CPLFW and Megaface	Verification performance	98.02, 99.82, 95.45, 92.08 and 98.48
VarGFaceNet [87]	2019	1G	LFW, CFP-FP and AgeDB-30	Verification performance	99.85, 98.5 and 98.15

It can be observed in the results above that deep networks can indeed achieve promising results. However, an enormous computational cost is observed, which limits the usage of these systems to high-end machines. Moreover, the results have much space for improvement on more complex datasets as can be seen for MobileFaceNetV1. This point has been made clearer in chapter four of this work.

2.2.1 Advantages and challenges in deep learning

Deep learning methods have shown significant advantages in face recognition tasks, outperforming traditional techniques in various aspects.

1. Robustness to variations: Deep learning models can learn hierarchical representations of faces, making them more robust to changes in lighting, pose, and facial expressions
2. Automatic feature extraction: Deep learning models can automatically learn discriminative features from large datasets, eliminating the need for manual feature extraction and domain expertise
3. Scalability: Deep learning techniques can handle large-scale datasets and multiple subjects, making them suitable for applications like surveillance and social media analysis

4. Improved recognition performance: Deep learning models, such as DeepFace, DeepID, VGGFace, and FaceNet, have achieved near-human or even superhuman performance on standard face recognition benchmarks

However, deep learning methods also come with their own set of challenges:

1. Need for large annotated datasets: Deep learning models require large amounts of labeled data for training, which can be difficult to obtain and maintain
2. Computational cost: Deep learning models can be computationally expensive, requiring powerful hardware like GPUs for training and inference
3. Interpretability: The features learned by deep learning models can be difficult to interpret, making it challenging to understand the underlying decision-making process
4. Overfitting: Deep learning models can be prone to overfitting, especially when trained on small datasets or with complex architectures. This can lead to poor generalization performance on unseen data

Despite these challenges, deep learning techniques have revolutionized the field of face recognition, offering improved performance and robustness compared to traditional methods. Ongoing research aims to address these challenges and further enhance the capabilities of deep learning-based face recognition systems.

2.3 Research Gaps

The key issues/limitations found in the above techniques are listed here:

- While deep-learning-based approaches surpass traditional approaches in accuracy, their computational cost is generally high.
- Existing techniques provide reasonable face recognition accuracy at the cost of high computational cost.
- In order to deal with intraclass variability, existing methods use high computational strategies which increase computation cost.
- Meta information required by existing methods is not available in most real-world scenarios.
- Existing methods are unable to deal with face images having large pose variations.

Chapter 3: Hybrid Holistic-Landmark Based Low Complexity System

Hybrid Holistic-Landmark Based Low Complexity System

In this chapter the construction of a facial recognition system with a reduced level of computational cost has been discussed. This system is resistant to substantial intra-class variability that is generally observed in face images.

3.1 Proposed Methodology

Our proposed method is subdivided into several parts as shown in Fig. 3.1 [8]. The sections where contributions have been made are highlighted in in green in the figure. Video frames are initially sent into the system. Landmarks are extracted from each frame [88] and utilized for ROI extraction. These landmarks are also used to determine facial posture. Then, to reduce the effects of lighting changes, illumination normalization is performed. After normalization Zernike moments are extracted from the images, which are then subject to feature selection. Finally, the features are supplied to the pose-specific classifier. Intra-class variability is significantly decreased because each classifier performs classification within a defined pose. As a result, the suggested system is capable of overcoming incremental-learning induced knowledge corruption without the use of extremely sophisticated classification methods like ensembles or 3D model fitting techniques.

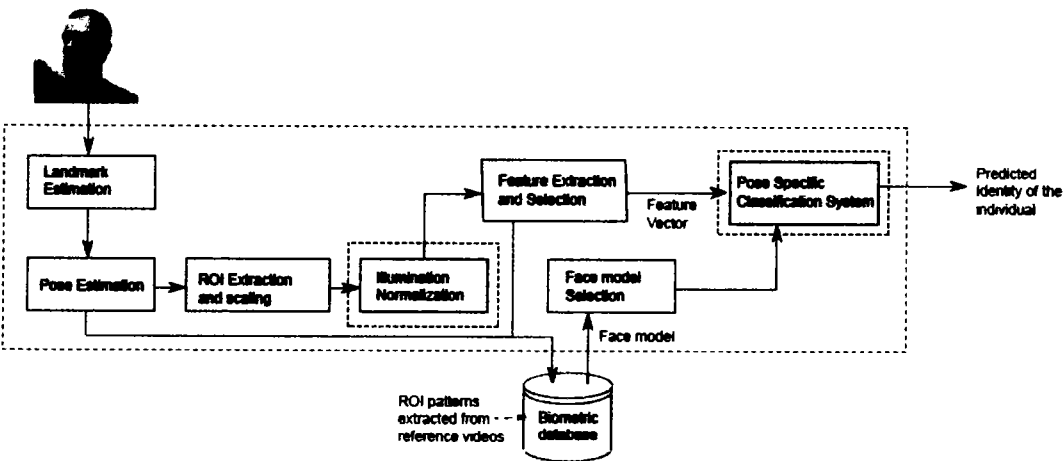


Figure 3.1: Flow of proposed facial recognition system

3.1.1 Overview

Intraclass variability may occur due to number of reasons including variations in facial orientation, face expressions, lighting conditions, etc. Techniques have been proposed to overcome this issue. For instance, technique described in [89] learns faces through ensemble of classifiers in contrast to incremental learning, where a change detection module stimulates creation of new classifier in respective ensemble. While author argues that incremental learning methods fail to produce good results due to high intraclass variability that FR systems need to face, ensembles come with a high computational cost as each subject will need to be compared against each classifier of every ensemble in database during operational mode.

A novel technique is provided in which landmarks are used in conjunction with holistic approaches to improve the system's overall classification accuracy. The cost of computation is also expected to decrease. Here orientation of the faces is found using landmark-based approaches and face images are compared against faces from the most similar pose only. To overcome variations in lighting conditions, utilization of an optimal illumination normalization technique is needed. When compared against commonly applied techniques involving ensembles of classifier an advantage of this technique is the significant decrease in computational cost. Instead of having to compare output from each classifier in the ensemble of each identity, this technique requires to compare only against faces from closest available orientation for each enrolled person.

3.1.2 Landmark Estimation

This work utilizes convolutional experts constrained local models (CE-CLM) [88] for landmark extraction. A modification was made in the existing technique by replacing the originally used Adams optimizer with SGD, resulting in an improvement of root mean squared value (RMSE) from 0.038 to 0.029. This technique benefits from neural architectures and a machine learning technique known as mixture of experts. CLM algorithms independently model the appearance of landmarks using local descriptors. Shape model is then utilized to apply constrained optimization. CE-CLM is a type of CLM models that use convolutional experts network (CEN) as local detector.

Some of the key motivations for using the CE-CLM model for landmark extraction are:

- Robust to occlusions and illumination changes

- Smoothness of tracking enabled by Expectation Maximization
- Ability to perform 3D landmarks extracted without training on 3D data
- No prior information about head pose needed

The network involves two main parts for landmark extraction: Response map computation using CEN and Shape parameter update using a point distribution model (PDM). The following objective needs to be optimized:

$$P^* = \operatorname{argmin}_P [\sum_{i=1}^n -D_i(x_i; I) + R(P)] \quad (3.1)$$

In this equation, P^* is the optimal set of parameters controlling the position of landmarks, D_i is the alignment probability of landmark i in location x_i for input facial image I . R is the enforced regularization. P is the current estimate given as $p = [s, t, w, q]$. Here, s , t and w are rigid parameters where s and t define shape and translation vectors, w is vector of rotation axes defining rotation matrix R and q is vector of non-rigid parameters

The value of x_i is parametrized by P as:

$$x_i = s \cdot R_{2D} \cdot (\bar{x}_i + \phi_i q) + t \quad (3.2)$$

Here, \bar{x}_i is the mean value of i th landmark and ϕ_i is principal component matrix. The CEN network used in this work is shown in the Fig. 3.2.

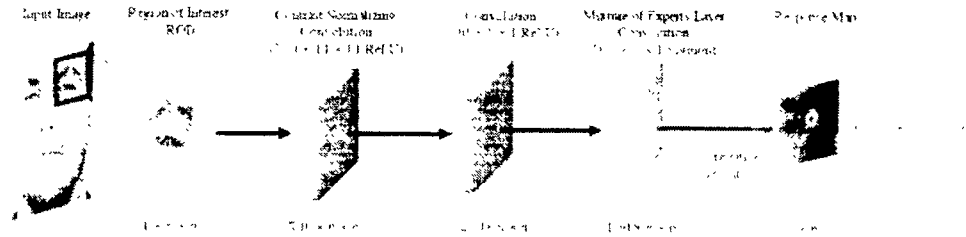


Figure 3.2: Convolutional Experts Network

In this network 19*19 patches for each landmark from the previous frame are fed to the contrast normalizing convolution block. Here, kernel of 11*11 is used with ReLU operation. This outputs a 500 * 9 * 9 feature map. This output is fed to another block with 1*1 kernel operation with the same output width and height as input. This layer also uses ReLU. Finally, the feature map is input to the mixture of experts block with sigmoid activation function. The

reason for using this activation is that its output can be directly considered as probability of landmark i being aligned at position x_i . Mathematically, the response map of the last layer can be given as:

$$\pi_{x_i}^i = p(l_i = 1, \hat{I} = I_{x_i}) \quad (3.3)$$

Where l_i indicates landmark i being aligned. \hat{I} is the image ROI at location x_i .

The landmark estimation technique has been utilized to extract region of interest (ROI), which have the purpose of removing or suppressing extraneous data while simultaneously improving some aspects for future processing. In this work, ROI extraction has been utilized to eliminate all but the face of a person from an image. This approach works by extracting facial landmarks and utilizing them to resize pictures to a preset scale. While each face's landmarks in each image were collected, only the largest image are kept in this research. This is to avoid unlabeled faces in the background. Afterwards, to minimize the impact of lighting changes, illumination normalization is performed. Fig. 3.3 depicts the pose estimation and facial ROI extraction techniques used in this study. Once the image is fed into the system, the landmarks from the face image are retrieved. This has been demonstrated in Fig. 3.3(a). As illustrated in Fig 3.3(b), the output contains 68 landmarks that correspond to the facial shape, lips, nose and eyebrows. These landmarks can be used to infer pose Fig. 3.3(c). The pitch, yaw, and roll values are the outputs of this stage. Fig. 3.4 depicts the pitch, yaw and roll values. The pitch and yaw values have been utilized to decide which posture group the image belongs to. The roll value is discarded in this work. The samples are separated into top, middle, and bottom and front and side orientations of poses.

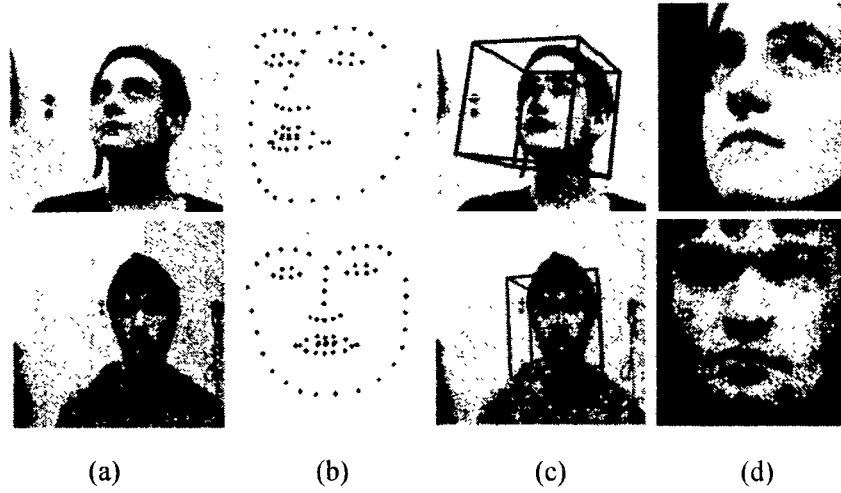


Figure 3.3: Implementation of Landmark extraction on the FR dataset (a) the original image, (b) the extracted landmarks, (c) the pose estimation results, and (d) image after ROI extraction.

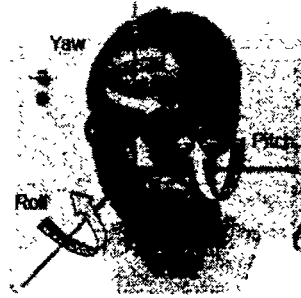


Figure 3.4: Pitch, yaw, and roll values.

To remove background from face images, landmarks depicting the face outline are used. Unlike a frequently used methodology [90], this method of ROI extraction should yield more accurate findings because the image does not contain background pixels, as can be seen in Fig. 3.3. (d). Since the separation between an individual and the camera can affect the apparent size of a face, scaling has been done on these images after extracting ROI. To calculate the scale factor, the maximum of perceived facial height and width are taken and divided by 70. This scale factor is used to resize the segmented image's width and height to ensure that the images are the same dimensions for the next steps.

$$S = \frac{\max(H, W)}{70} \quad (3.4)$$

$$\begin{bmatrix} X' \\ Y' \end{bmatrix} = \begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} \quad (3.5)$$

Here, the image's height and width are H and W , respectively, while the new locations of the X and Y points are X' and Y' . To produce a square picture of $70 * 70$ dimensions, empty columns or rows are added equally on both sides. Pixels are used as a measure of all values.

3.1.3 Illumination Normalization

Invariant features, variation modelling, and canonical form are the three categories of approaches to dealing with illumination-related difficulties [9]. In the first strategy, efforts are made to use features that are less impacted by lighting differences in terms of appearance. Invariant features not only restrict us to a small subset of available features, it has been unable to produce great results in the literature. The goal of variation modelling is to figure out how much variation there is in a given subspace. The closest subspace for the given image is picked for recognition. This method, though reliable, has the disadvantage of necessitating a large training set of data containing sufficient variations in illuminations in images. The canonical form, on the other hand, aims for image-by-image manipulations normalizations or by creating a new image. This category also includes a frequently used method known as Contrast Limited Adaptive Histogram Equalization (CLAHE) [92].

While some normalization strategies are well-known for their superior performance, the performance of any particular preprocessing technique can vary depending on the features utilized. This study tests two image normalization techniques. An empirical investigation conducted by [93] found that the strategy presented by Tan and Triggs (TT) [94] is the most effective across a variety of features. The effectiveness of CLAHE is compared against the TT method with the features utilized in this work. In the following sections, each of these approaches has been described briefly.

Tan and Triggs

Tan and Triggs' [94] preprocessing method has been widely adopted in image processing applications [95-98]. A recent study reported by [93], in which a number of illumination normalization techniques were examined, provided the impetus for testing this technique in features used for this work. When using numerous features, the work demonstrates that this technique has the highest classification accuracy. As shown in Fig. 3.5 [94], it entails a series of actions to combat local highlights and shadowing.

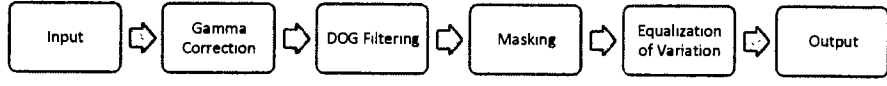


Figure 3.5: Steps involved in Tan and Triggs normalization algorithm

Contrast Limited Adaptive Histogram Equalization

The approach of using the histogram to balance contrast in images is known as histogram equalization [92]. The pixel intensity x is transformed to x' using the transformation function T , which is obtained by multiplying a cumulative histogram with scaling factor. The scale factor makes certain that the intensity value stays within the range of intensity of the image. The intensity range 0 to 255 is a common example. The intensity values of an input image X can be thought of as random variables with values ranging from $[0, L - 1]$, where L depicts the count of distinct grey levels. The intensity at a given spatial position (i, j) is represented by $X(i, j)$. The histogram h is defined as follows:

$$h(X_k) = n_k \quad (3.6)$$

Here n_k is the number of all the pixels at a given level and the k th level of intensity within range $[0, L - 1]$ is given by X_k .

At the j_{th} intensity level, the cumulative distribution function (CDF) describes the possibility that X will have an equal to or less than value to j . This can be determined using the following formula:

$$C(X_k) = \sum_{j=0}^k \frac{h(X_j)}{N}; k = 1, 2, \dots, L-1 \quad (3.7)$$

Here, N is the number of pixels in total in the image. As a result, when mapped to new values, A greater stretch will be produced by a more frequently occurring value. As an outcome, low-contrast areas obtain increased contrast. Because in more frequently occurring values a greater stretch is expected, the visual outcome of stretching is maximized.

Equalization of histograms can thus be expressed as:

$$T(X_k) = (L - 1) \times C(X_k); \text{ for } k = 1, 2, \dots, L - 1 \quad (3.8)$$

This procedure is simple and straightforward; unfortunately, it has certain limits. In circumstances where photos have varying grayscale distributions in different places, the equation above does not perform effectively when applied globally. It's possible that a lot of information is lost, and certain image areas become overly bright. Furthermore, this approach does not preserve image brightness.

After breaking the image into smaller sections, histogram equalization on each region's pixels has been suggested as a solution to this problem [99]. This somewhat solves the problem; however, it does not function well when the image comprises highly localized dispersion. HE has an issue of mapping two extremely similar values to widely differing ones in low-contrast images where the curve for its mapping has an abrupt slope. In the existing literature [100, 101], the technique referred to as Contrast Limited Adaptive Histogram Equalization (CLAHE) [102] is being utilized to deal with this problem by restricting the amount of allowed contrast by HE. At first, the image is separated into almost equal-sized non-intersecting parts. Then, for each location, a clip limit β is determined. The histogram for each region is then reallocated so that the histogram's height does not exceed the clip limit. The following formula is used to get the value for β :

$$\beta = \frac{MN}{L} \left(1 + \frac{\alpha}{100} (S_{\max} - 1) \right) \quad (3.9)$$

Here, α denotes the clip factor and the maximum permitted slope is S_{\max} .

As Tan and Triggs technique is based on a series of steps, the simplicity of CLAHE is clear advantage because it is a single-step procedure. Furthermore, it has the ability to improve certain hidden information. The author of TT, on the other hand, claims that his approach is more resistant to shadows and highlights and is less prone to lose important information.

3.1.4 Feature Extraction

To minimize the dimensionality of images, feature extraction is performed. While relevant features can improve a system's effectiveness, redundant features can degrade efficiency. This study uses Zernike moments, which are produced by applying complex Zernike polynomials onto images. While Zernike moments are known to be complex, this problem has been solved in the work presented in [103]. Not only do these moments have

historical importance, these are still found useful in the recent work [104, 105]. Some of the key attributes of these moments are described below:

- There is minimum redundance or overlap of information in Zernike moments owing to orthogonality of Zernike polynomials.
- In comparison to lower order Zernike moments, Zernike moments of higher orders have a higher sensitivity to details which translates to a far higher sensitivity to noise and computational cost.
- Illumination has minimal impact on these moments.
- Value of Zernike moments is affected by the scale, translation and rotation; however, their magnitude is unaffected by rotation.

32 Zernike moments of low order that meet the criteria in equation 3.1 are considered due to their lower susceptibility to noise:

$$Z_{n,m} \forall \begin{cases} 3 \leq n \leq m \\ |m| \leq n \\ n - |m| = 2k \end{cases} \quad (3.10)$$

Where $k \in N$, The order of the moments is given by n , while the degree of repetition is given by m .

Extracted ROIs, as previously stated, feature solely a person's face and are resized to a standard scale. This should lessen the variation issue brought on by the size of these moments.

3.1.5 Classification

Classification is a type of supervised learning in which a model is trained to predict the class or category of an observation based on its features. The goal of classification is to accurately assign new observations to one of a set of predefined classes. The main novelty of this module lies in the classification approach being utilized. In addition, a voting approach that combines the Bayes Network, Naïve Bayes, and J-48 schemes has been devised to improve classification results at minimal computational cost.

Classification Approach In this research, a classification approach has been devised that utilizes a pose estimation module to classify face images into pose categories. During the training mode, images are stored in their relevant databases for each category, and different classifiers are trained. During operation, the face is recognized using the relevant classifiers after the pose is computed using the same pose-estimation algorithm. Implementation details of this approach can be seen in section 3.2.2.

Feature Selection Correlation-based feature selection has been used to select characteristics that are highly connected with the relevant class. This reduces dimensionality and, consequently, computational cost of the classifier. Best First Search was used as the search strategy.

Voting Approach Voting approach using Naive Bayes, Bayes and Decision Trees was chosen for this research. Average probabilities are utilized as the combination rule, while K2 is used as the Bayes Network search method. Each of these algorithms is known for its ease of use, speed of convergence, and ability to operate with multiple independent variables. This method should provide better accuracy without causing increased computational cost to the system since the used features are independent of one another.

This work employs the Bagging method [106], which is another well-known algorithm that uses the ensemble-based technique, to compare the devised voting scheme. The comparison against Bagging is motivated by work presented in [107].

3.2 Experiments and Results

The suggested pose specific classification system's performance in identifying persons in video frames has been evaluated. These tests are used to compare the efficacy of training inside a specific pose group against that of collective training.

3.2.1 Database

The BIWI Kinect dataset [108] contains images of a total of 20 people with 14 men and 6 women at a resolution of 640 by 480 pixels and a 24-dpi bit depth. Owing to four people returning for another session, the overall number of videos is 24 in which the frames of the videos have been converted into sequences of frames. Over 15,000 photos of people are included in the dataset. Having access to greater degree of facial poses for each individual with yaw ranging in $\pm 75^\circ$ and pitch extending $\pm 60^\circ$ is the key rationale for adopting this database. Despite the fact that ground truth about the head's location and rotation is available, this approach benefits only from the video frames in this database. Due to the vast range of poses, this dataset can be deemed harder for most algorithms, as it considerably increases intra-class variability. Figure 3.6 depicts some of the photographs in the database.



Figure 3.6: Random images from dataset

Table 3.1 summarizes basic information about this dataset.

Table 3.1: BIWI Kinect dataset summary

Dataset	Persons	Frames	Pose
BIWI Kinect [108]	20	15000	Pitch: $\pm 60^\circ$, Yaw: $\pm 75^\circ$

3.2.2 Simulation Setup

As described in section 3.1.2, CE-CLM model was used to extract landmarks. In this work, SGD optimizer was utilized in place of Adams optimizer as used in original work. This yielded better RMSE of 0.029 compared to 0.038 as described in the publication. No other changes were made to the original work.

Poses were divided vertically into top, middle and bottom with each of these having near forward looking and side poses. For this purpose, threshold is established vertically at $\pm 12^\circ$ and $\pm 35^\circ$ to have pitch value for the faces divided into three intervals. Similarly, the yaw thresholds were chosen at $\pm 15^\circ$ and $\pm 45^\circ$. Images were thus separated into multiple datasets as demonstrated in Table 3.2.

Table 3.2: Pose categories for classification

Pose Categories		
Yaw: -45° to -12° Pitch: 15° to 45°	Yaw: -12° to 12° Pitch: 15° to 45°	Yaw: 12° to 45° Pitch: 15° to 45°
Yaw: -45° to -12° Pitch: -15° to 15°	Yaw: -12° to 12° Pitch: -15° to 15°	Yaw: 12° to 45° Pitch: -15° to 15°
Yaw: -45° to -12° Pitch: -15° to -45°	Yaw: -12° to 12° Pitch: -15° to -45°	Yaw: 12° to 45° Pitch: -15° to -45°

Images from first column were then added to the last after flipping them horizontally. The six poses were thus formed as: top (side and middle), middle (side and middle), and bottom (side and middle) poses for each person. Faces with greater pose angles were preserved in a distinct group, but due to pose estimation inaccuracy, they were not considered. Only those people from the database who have at least 20 photographs of each posture been chosen. As a result, data from 8 of the 20 people in the database is available.

Each ROI was scaled to a size of 70×70 after posture estimation. In this study, the magnitudes of Zernike moments were utilized. Because these moments are not invariant to translation and size, ROIs recovered from these images were resized to a standard scale of 70

x 70. Correlation-based selection technique for features is used to reduce the number of features collected from the images, which are then used for classification.

The frontal photographs were the subject of more intensive testing. Once ROIs were extracted, the HE and TT approaches were evaluated to see how well they worked with the suggested feature sets in this work. In addition, two classification schemes were applied. Once the most suitable scenario was determined, the proposed pose-specific classification technique was used to train and test the face photos from each pose. To test the performance of the proposed method, a new dataset was created and classified without using the devised posture estimation module. Except for utilization of pose-specific classification, this dataset contains the same number of photographs and follows the same procedures as proposed technique.

3.2.3 Performance Measures

Given the final output of a classifier for an individual, the true positive rate (TPR) is the proportion of accurately detected positives over the overall positive samples. Similarly, the false positive rate (FPR) is the percentage of negatives incorrectly identified out of the total number of negative samples. Mathematically:

$$TPR = TP / (TP + FN) \quad (3.11)$$

$$FPR = FP / (FP + TN) \quad (3.12)$$

Here $TP = \text{true positive}$, $FN = \text{false negative}$, $FP = \text{false positive}$ and $TN = \text{true negative}$

Area under curve (AUC) is the area under receiver operating characteristic (ROC) curve, where ROC plots TPR against FPR at all classification thresholds. This measure provides threshold independent performance of a classifier.

F1 is a scaler measure that provides a balance between precision and recall. This metric is expressed as:

$$F1 = \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (3.13)$$

This research uses a classification strategy that combines the Bayes Network, Naïve Bayes, and J-48 schemes with the help of voting mechanism. This voting combination was

established after assessing the effectiveness of many algorithms individually and in combination.

3.2.4 Results and Discussion

The results of the frontal face experiments are shown in Table 3.3. The findings are compared to determine which of the normalization procedures among CLAHE and TT is the most effective. Furthermore, the bagging method was compared to the suggested voting-based classification method. In each case, tenfold cross validation is used. In the table below, results only for the frontal poses are reported to compare results with two different normalization techniques.

Table 3.3. Bagging and proposed classifiers employed with TT and CLAHE approaches on frontal face images.

Classifier	Normalization	FP Rate	TP Rate	Precision	AUC	F1-Measure
Bagging	TT	0.040	0.757	0.763	0.950	0.750
	CLAHE	0.010	0.943	0.944	0.992	0.942
Proposed ensemble	TT	0.036	0.786	0.804	0.960	0.787
	CLAHE	0.004	0.979	0.979	0.998	0.979

There is a noticeable improvement when the suggested classifier is combined with CLAHE as a normalization strategy. This is contrary to the initial expectations; yet, as both the preprocessing methodology and the features must be taken into account for better classification, it was deduced that the TT normalization approach is less effective with Zernike moments. The suggested classification ensemble produces significantly better classification results than the Bagging technique. As a result, this work with was proceeded with the voting technique with illumination normalization performed using CLAHE.

This module is focused on proving efficacy of the proposed pose specific approach that lays foundation for the next module. To validate the efficacy of pose specific classification based FR system, the results are compared using a classification method that is identical to the devised method with the exception of pose estimation. Average results after running the system with all six poses in the images have been obtained and are summarized in Table 3.4.

Table 3.4: Result obtained for classification with and without pose estimation.

	FP Rate	TP Rate	AUC	F1-Measure	Precision
W/o pose estimation	0.017	0.899	0.992	0.900	0.903
Proposed system	0.006	0.962	0.997	0.964	0.967

3.3 Chapter Summary

In this chapter, a novel face recognition system is demonstrated on video frames. The system uses a landmark-based approach to extract the region of interest (ROI) and remove backgrounds, allowing it to focus more accurately on the face itself. The scale and orientation of each image are adjusted to eliminate variations, and Zernike moments are extracted for feature extraction. The proposed ensemble and Bagging algorithm are then used for classification, resulting in improved accuracy.

The results show that the pose-specific classification technique not only increased the FR system’s accuracy, but the voting classifiers that were proposed also proved to be effective. These techniques could lead to an increase in the accuracy of face recognition systems, particularly in challenging conditions such as variations in pose. The use of voting classifiers also adds an additional layer of robustness to the system. These improvements could have significant benefits for a wide range of applications, including security, surveillance, and access control.

Chapter 4: Low Complexity CNN Based FR System

Low Complexity CNN Based FR System

This chapter presents a unique paradigm for low-end systems to handle difficult FR problems. Here an attempt has been made to benefit from the power of cellular neural networks to achieve results close to that of the state-of-the-art systems but at a fraction of the computational cost.

4.1 Background

Since this chapter vastly depends on Cellular Neural Networks, which are a field of Deep Learning, it is necessary to provide some background about what these are how they work.

4.1.1 Artificial Neural Networks

To handle large amounts of data, Artificial Neural Networks are generally considered to be the best option among classifiers. This is even more valid for Image Processing and Computer Vision applications as most of the research done in the recent years has utilized this technique. According to the definition [109] “A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates”.

Biological neural networks, which are made up of clusters of chemically coupled neurons, have become the inspiration for artificial neural networks. Any one of these neurons may be connected to numerous other neurons forming an extensive network. Artificial neural networks are one of the paradigms for information processing that have been inspired by processing mechanism of biological neural systems. To properly understand the working of ANNs, it is necessary to understand how a single neuron works.

Artificial neurons can be described as mathematical functions based on biological neurons where each of them takes input, assigns separate weights to them, sums them up and finally passes this sum through an activation function. Fig. 4.1 describes an example. Here, sigmoid has been taken as the activation function. X_1 to X_L are the total inputs and W_1 to W_L are the weights assigned to each.

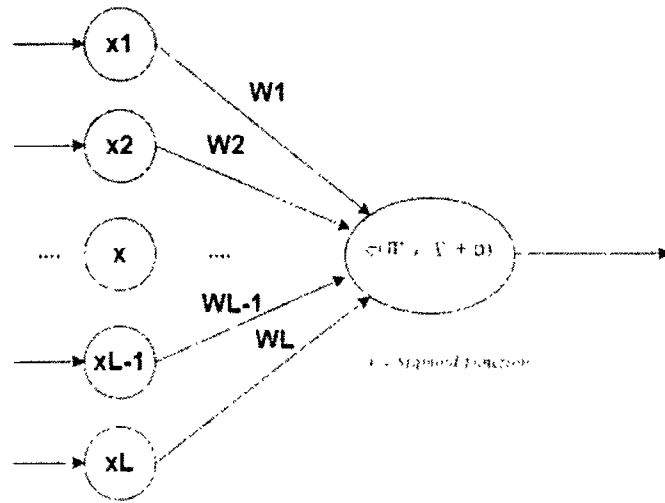


Figure 4.1: Working of a neuron

Feed Forward Neural Networks

Feed-forward is the mathematical calculation of intermediary variables (weights and biases) applied to the input data or outputs of the previous layer. The coefficients' data is stored in the neuron as a result of this operation to be used again when the network performs backward propagation. The terms input layer, output layer and hidden layers have been used to describe function of these layers. Dimensions of the input layers are same as the shape of inputs supplied to the network. This layer serves to provide input to the hidden layer. Depth, width and activation function defined in each of the layers describes the architecture of the neural network. Here, by width signifies the number of neurons in each of the hidden layers. Depth refers to number of hidden layers used. Each of the neurons in the hidden layers has an activation function. This has been described in more detail in a later section of this chapter. Fig. 4.2 shows the overall working example.

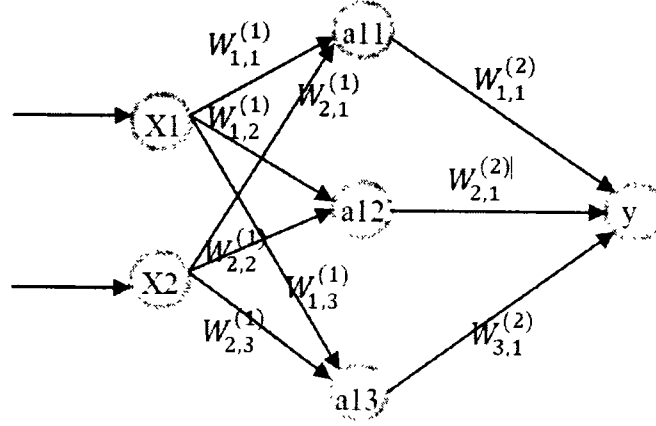


Figure 4.2: Feed forward propagation example

Here, x , a and y in the prefix represent the input, hidden and output neurons. W is the weight with numbers in the subscript representing to and from numbers of neurons as $W_{previous\ layer,next\ layer}^{(layer)}$. The weight matrix for the first layer can be calculated as:

$$W^{(1)} = \begin{bmatrix} w_{1,1}^{(1)} & w_{1,2}^{(1)} & w_{1,3}^{(1)} \\ w_{2,1}^{(1)} & w_{2,2}^{(1)} & w_{2,3}^{(1)} \end{bmatrix} \quad (4.1)$$

No calculations are performed on the input layer. The hidden layer here is a matrix with a single row and three columns – one for each neuron. This is calculated as:

$$a_1 = [a_{11} \ a_{12} \ a_{13}] = XW^{(1)} \quad (4.2)$$

Where X is the vector of input neurons x_1 and x_2 .

For the second layer, $W^{(2)}$ is a matrix with three rows and one column. This can be given as:

$$W^{(2)} = \begin{bmatrix} W_{1,1}^{(2)} \\ W_{2,1}^{(2)} \\ W_{3,1}^{(2)} \end{bmatrix} \quad (4.3)$$

The output can thus be given as:

$$y = a_{11}W_{1,1}^{(2)} + a_{12}W_{2,1}^{(2)} + a_{13}W_{3,1}^{(2)} \quad (4.4)$$

Activation function

An activation function in a neural network defines how in a node or nodes of a layer the weighted sum of the input transforms into an output. It is inserted into an artificial neural

network to assist the network in recognizing difficult data patterns. When comparing to a neuron-based model seen in our brains, the activation function decides what should be triggered onto the next neuron at the end of the process. An activation function in an ANN converts the output signal of the previous cell into a format that may be utilized as input to the following cell. Some of the main benefits of an activation function include keeping the output into limits according to our needs and introducing nonlinearity into the system.

In the past, linear activation functions were experimented on and soon discarded. Here input data was multiplied by intermediate coefficients and generated an output of zero or one. This method was inappropriate for backward propagations. Non-linear activation functions, in contrast, generate an output within a limited range. Activation functions need to be non-linear as classifiers like neural networks generate non-linear patterns which cannot be learned through linear activation functions. Some activation functions that are frequently utilized are described in Table 4.1.

Table 4.1: Popular activation functions

Activation function name	Equation	Output
Sigmoid / logistic	$f(x) = \frac{1}{1 + e^{-x}}$	0,1
Tanh or hyperbolic tangent	$f(x) = \frac{e^x + e^{-x}}{e^x - e^{-x}}$	-1,1
ReLU or Rectified Linear Unit	$f(x) = \max(0, x)$	0 to x

These are also depicted in graphical form in Fig. 4.3. The outcome of each of these is given below:

Sigmoid:

The sigmoid function is an S-shaped curve that maps input values to the range [0, 1]. However, it suffers from the vanishing gradient problem, limiting its use in deep networks.

Some of its key properties are:

- **Range:** The output lies between 0 and 1.
- **Smoothness:** Sigmoid provides smooth transitions.

- **Binary Classification:** Often used in the output layer for binary classification (e.g., predicting probabilities).

Tanh (Hyperbolic Tangent):

The Tanh function is a scaled and shifted version of the sigmoid function. It maps input values to the range $[-1, 1]$. It is commonly used in recurrent neural networks (RNNs) and gradient-based optimization.

Some of its key properties are:

- **Range:** The output lies between -1 and 1.
- **Symmetry:** Tanh is symmetric around the origin (0).
- **Non-linearity:** It introduces non-linearity, making it suitable for hidden layers in neural networks.

ReLU (Rectified Linear Unit):

ReLU is a simple yet powerful activation function. It replaces negative input values with zero and leaves positive values unchanged. It is widely used in deep learning architectures, especially in convolutional neural networks (CNNs).

Some of its key properties are:

- **Range:** The output is non-negative (0 or positive).
- **Sparsity:** ReLU sparsifies activations by zeroing out negative values.
- **Computational Efficiency:** It is computationally efficient due to its simplicity.

In summary:

- **Tanh:** Suitable for hidden layers, provides both positive and negative activations.
- **ReLU:** Efficient, encourages sparsity, and mitigates vanishing gradients.
- **Sigmoid:** Commonly used for binary classification and smooth transitions.

Cost Function

The goal of the cost function is to calculate the neural network error ratio. The sum of the error, which is the difference between the real and expected values, gives the general formula for the cost function. In a single training example, the loss (error) function is:

$$Loss(\hat{y}, y) = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \quad (4.5)$$

This modifies to below mentioned for entire training set:

$$J(w, b) = \frac{1}{m} \sum_{i=1}^m Loss(\hat{y}^i, y^i) \quad (4.6)$$

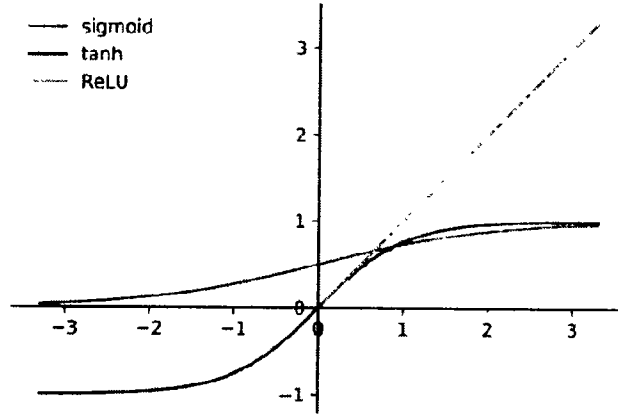


Figure 4.3: Graph of Sigmoid, Tanh, and Relu functions

Backward Propagation

Backward propagation is based on minimizing the cost function error ratio and iteratively adjusting the coefficients by performing the cost function optimization procedure. One of the most commonly used optimizing technique used for training is stochastic gradient descent. The backward propagation involves two steps: computing partial derivatives and updating the weight matrix.

To update weights and biases, the derivatives are calculated in backward propagation as below:

First, the derivative for the output layer is calculated

$$\partial Z^l = a^l - y \quad (4.7)$$

Here, Z represents the output layer. Second, multiply the derivative value of the final layer with the current layer parameters to get the previous layer derivative and update the weight parameters. Biases will use the final layer's derivative value.

$$\partial W^l = \partial Z^l \cdot a^{[l-1].t} \quad (4.8)$$

$$\partial b^l = \partial Z^l \quad (4.9)$$

4.1.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of neural network that focuses on processing of data that has a grid like structure such as images. Though originally designed for images, CNNs have proven their worth over the past few years in several fields not only like Image Processing and Computer Vision but also fields like Voice Recognition etc. According to the definition by Ian Goodfellow et. al., “*Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.*” One of the key advantages of the CNNs is the reduction in number of features.

CNN architecture constitutes of three types of layers (convolutional layer, pooling layer, and fully connected (FC) layer), which are arranged together to form CNN. This has been shown in the image below:

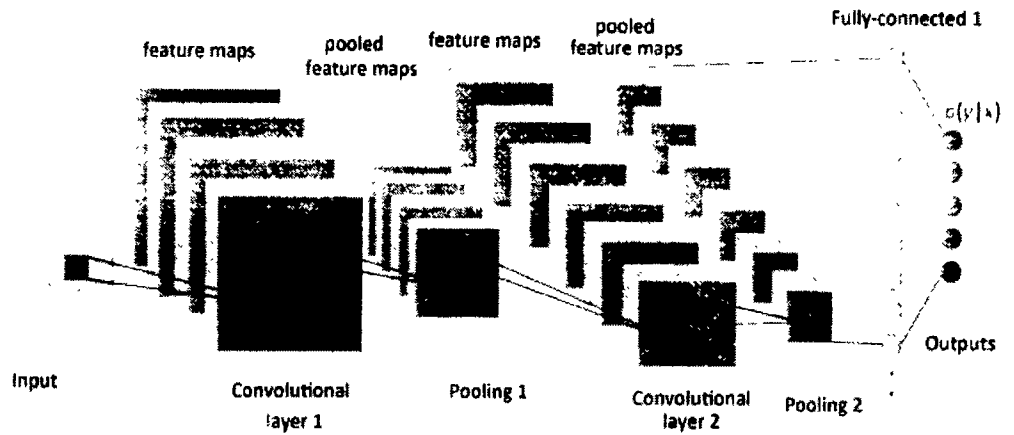


Figure 4 4: Structure of CNN

CNNs typically have an input layer, one or more convolutional and pooling layers, and a fully connected layer in the most common configuration. If the input is an RGB image, the input layer would have to have a depth of 3 to accommodate each of these channels.

Filters and kernels are executed on the original image or other feature maps in the convolutional layers of a deep CNN. The large number of the user-specified parameters in the network are contained in this layer. The number of kernels and the size of the kernels are the most critical parameters. Some common sizes of filters are 3×3 and 5×5 with the depth determined by depth of the input. A filter might be connected with anything; for example, one filter could be associated with seeing specific portion like noses in human images, and the nose filter would indicate how strongly a nose appears in the image, as well as how many times and in what locations it appears.

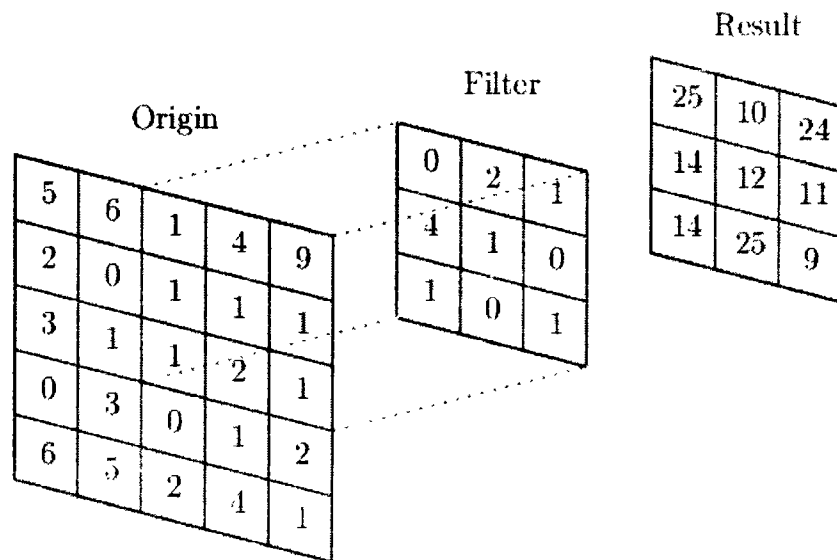


Figure 4.5 Example of convolution operation

As an output of the application of these filters, a feature map is generated for each filter, which is then passed through activation function that performs the function of deciding whether or not a feature is present at that specific location. The output of activation function is called activation map which is often omitted in figures for compactness. Another filtering layer or a pooling layer could come after this operation. A pooling layer selects the desired values like the maximum from the previous layer. This reduces the size of the feature maps for subsequent processing.

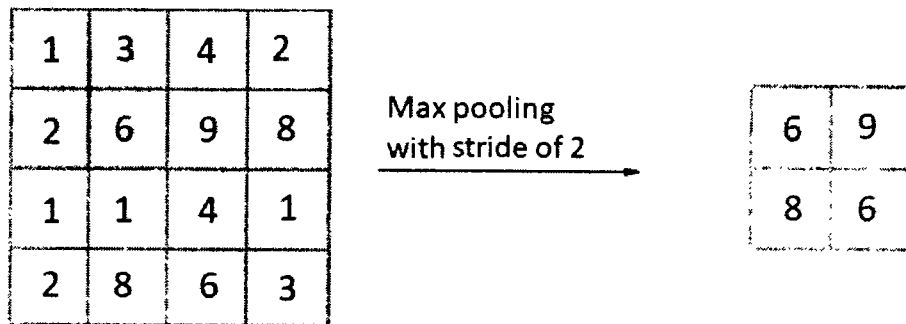


Figure 4.6: Pooling example

Fully connected layers are introduced before the classification output of a CNN to learn non-linear combination of the feature map. This is similar to an MLP's output layer.

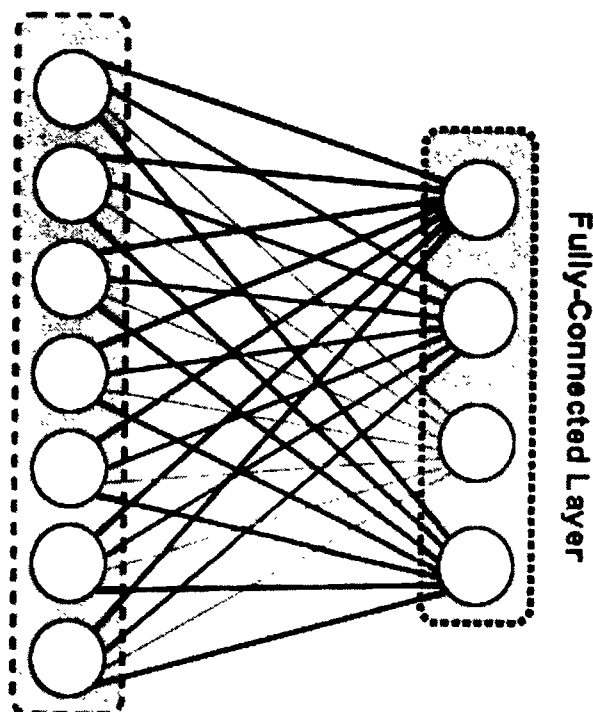


Figure 4.7: Fully-connected layer

4.2 Work Methodology

To achieve the desired results, a network with minimal computational cost was used, drastically decreasing a system's computational requirements and maintaining a flow that

allows these systems to run FR without the need to transmit and receive data to a server. Fig. 4.8 depicts the overall flow of the network, where sections with any improvements highlighted in green. More information on these steps can be found in the sections below.

4.2.1 Preprocessing

Pose Estimation

The pose of a person is obtained in the proposed architecture to determine which pose set a specific image or frame belongs to.

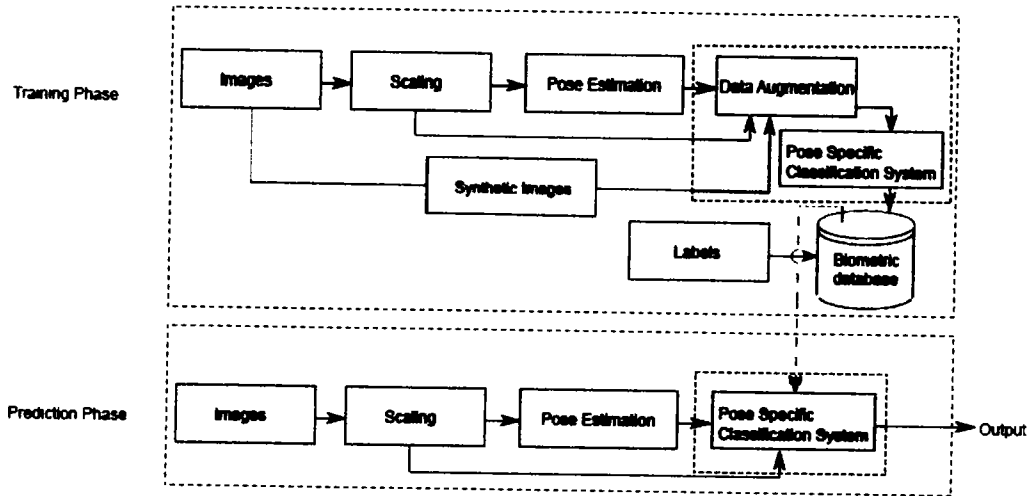


Figure 4.8: Overall system workflow

Pitch, yaw, and roll values are used in Head Pose Estimation (HPE) procedures to indicate the direction of the face as shown in Fig 4.9. To put another way, these values can be described as nodding, shaking (right to left) and tilting of the head respectively. Depending on the range of pitch the system is designed for, these approaches are classified into narrow-angle and wide-angle procedures.

Pose estimation can be done using traditional approaches like template matching and cascade detectors as well as geometric and deformable models that use features like face key-points. In addition to this, regression and classification methods are also used and have gained great popularity. Regression methods utilize mathematical models to directly predict head pose from labelled data, whereas classification methods use a discretized set of poses.

The method utilized in this study [110] is a regression methodology. This methodology, according to the author, is the best performing broad range HPE and is within 1.8 percent of

the best [111] HPE methods at the time. Despite the fact that there are higher performing models in terms of estimation accuracy [111, 112], this method was chosen because of its lesser computational cost, which allows it to run at a high frame rate on low-end devices as well as the ability to reliably detect wider poses. The entire process is described in the paragraphs below for interested readers.

The WHENet approach is a follow-up to the work described in [113]. The backbone network is EfficientNet-B0 in this approach. A binary cross-entropy loss is generated for the classification of each of the pitch, yaw, and roll values into 3° bins with the help of softmax, and a mean squared error (MSE) regression loss is used for the predicted value of softmax output and ground-truth labels. The following equation is used to integrate these classification and regression losses for pitch, yaw, and roll:

$$L = \alpha L_{reg} + \beta L_{cls} \quad (4.10)$$

Here, α and β are constants used to tradeoff one over the other and L_{reg} and L_{cls} are the classification and regression losses.

The classification loss was calculated using the sigmoid function. To avoid unexpected behaviors at wider pitch angles, MSE was used with a novel wrapped loss in the regression. The minimum rotation angle required to align yaw predictions with their associated annotations is penalized by this loss function:

$$\begin{aligned} L_{wrap}(\theta_{pred}, \theta_{act}) \\ = \frac{1}{N_{batch}} \sum_i^{N_{batch}} \min \left[|\theta_{pred}^i - \theta_{act}^i|^2, \right. \\ \left. (360 - |\theta_{pred}^i - \theta_{act}^i|)^2 \right] \end{aligned} \quad (4.11)$$

Here, L_{wrap} denotes the wrapped loss and θ_{act} and θ_{pred} are the target and predicted values of rotation angles.



Figure 4.9: Example of pitch, yaw and roll values

To determine which pose category the image should belong to, only pitch and yaw measurements are used. The pose categories are described in Table 4.4. Self-occlusion is prone to occur when yaw values are high, which can lead to image distortion during reconstruction. To avoid this, all images with negative yaw value of the face were mirrored horizontally thereby eliminating the need for generating synthetic face image towards left from face image directed right and vice versa. This has the added benefit of having fewer sets of poses and, as a result, requiring fewer classifiers. The train and test data used in this model are summarized in Table 4.2.

Table 4.2: Datasets used for pose estimation algorithm

Dataset	Images/frames	Pose	Purpose
300-LP	61,225	large poses	Train
CMU panoptic (processed)	65 Sequences of 5.5 hours	360°	Train
AFLW2000	2000	up to 90°	Test
BIWI Kinect [108]	15,000	Pitch: $\pm 60^\circ$, Yaw: $\pm 75^\circ$	Test

Dataset and Augmentation

As previously stated, the technique relies on a pose-specific classification algorithm. This means that the data must contain a sufficient number of faces from each pose. This is not often expected from most benchmark datasets and is difficult to obtain during enrollment in a real-world scenario. To address this problem, an approach [114] for 3D reconstruction of a face model was utilized. Most frontal images in the dataset for each ID were used to create

the 3D model. Image cannot be rotated to a desirable pose after it has been recreated in 2D. To allow the system to better learn similarity between synthetic and profile images as well as to learn illumination variations, a mix of synthetic and original images from Caspeal [115] and BIWI Kinect [108] datasets were also included in the training set. Since BIWI dataset has missing samples in certain pose categories for some of the persons, no synthetic image was created for those poses considering that a purpose of these datasets is to aid learning similarity between the two types of images. Pose sets for Megaface [116] identities may however contain synthetic only images for some persons as the test image may appear from any of the poses. Background was applied on all synthetic images using Describable Textures Dataset [117]. The 3D reconstruction approach in use was found to be quite robust in a wide range of poses. The technique is used to extract random face images from each pose and save them for training purposes, along with the image's label. Table 4.3 describes these datasets.

Table 4.3: Classification datasets

Dataset	Persons	Frames	Pose	Purpose
BIWI Kinect [108]	20	15000	Pitch: $\pm 60^\circ$, Yaw: $\pm 75^\circ$	Train
Caspeal [115]	1040	99,594	27 poses	Train
Megaface [116]	6,72,057	4.7 M	-	Train
FaceScrub [116]	530	107,818	-	Test

Scaling

The perceived size of a face in a photograph can vary depending on the image size and the distance of a person from the camera. To counteract this, scaling was applied to all images once data was collected. To calculate the scaling factor, the greater number for width and height and divide it by N , where N is the required size for width and height of the final image. The width W and height H of the image are then resized using this factor. Let S represent the scaling factor, and X' and Y' represent the updated locations of points X and Y . Then:

$$S = \frac{\max(H, W)}{N} \quad (4.12)$$

$$\begin{bmatrix} X' \\ Y' \end{bmatrix} = \begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} \quad (4.13)$$

With the use of zero-padding, the smaller of the two dimensions is expanded to N pixels. As a result of this step, all photos being input to the system have a common size, and the system's requirement to manage size variations is reduced

4.2.2 Classification

As a key focus of this research is to keep computational cost of FR low, a recent work termed as RepMLP [118] was used. This network utilizes only 406M FLOPS. The key intuition behind this work is that bringing together the global representational ability and the spatial awareness of fully connected (FC) layers with the local processing of CNNs can advance the performance of a classifier. Some of the convolutional layers are replaced by FC to act as transformation between feature maps. A brief description of this work is given below.

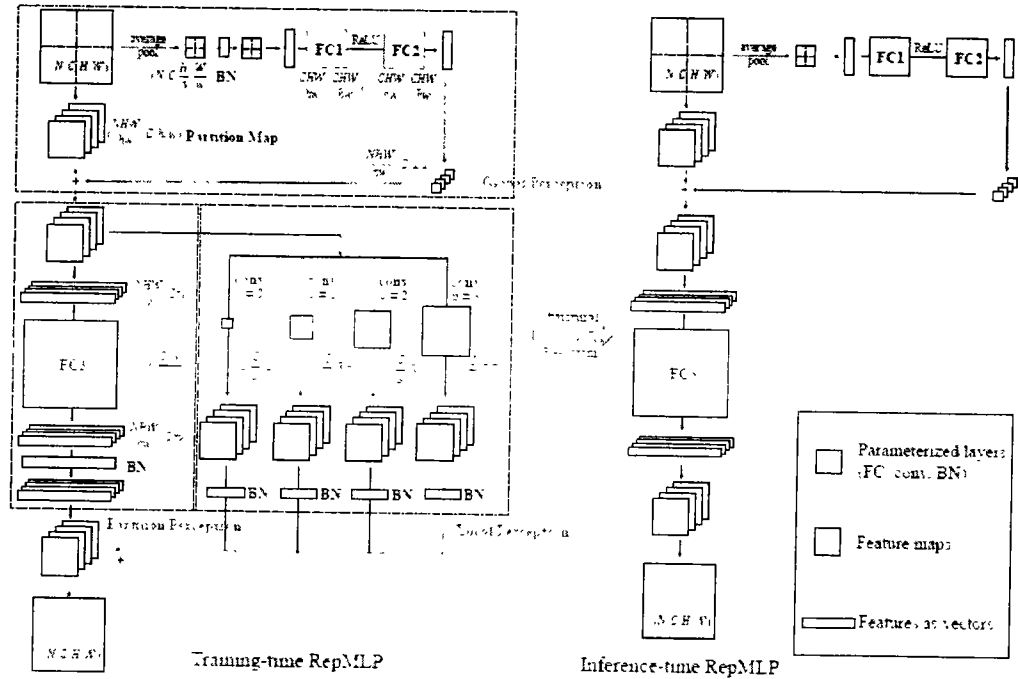


Figure 4.10: Working of RepMLP

RepMLP has three parts during training phase: Global Perceptron, Partition Perceptron and Local Perceptron. Each of these is visible in Fig. 4.10. In the Global Perceptron phase, feature map is broken up into partitions. Parameters can then be shared among these partitions. To introduce correlation into these partitions, a series of operations is performed. For each partition, a pixel is obtained using average pooling. It then goes through batch normalization and afterwards an MLP with two layers. Finally, this output is reshaped and added to the partition map. In Partition Perceptron phase groupwise 1×1 convolutions are applied to the FC. Here, the feature map is first reshaped with a spatial proportion of 1×1 . It is then passed through 1×1 convolution step with a preset number of groups and the output is again reshaped. The Local Perceptron phase applies a number of convolutional layers on the feature map. For inference, the network is squeezed in two ways. The batch normalization is fused in the previous layer and conv layers are merged into an FC layer. For more details about working of RepMLP, reader is directed to the actual publication. This work uses RepMLP implementation on Resnet-50 where ordinary bottleneck blocks are replaced by RepMLP bottleneck blocks. The usual $4 \times$ channel reduction and expansion is enclosed by $r \times$ channel reduction and expansion where r is a parameter having different values for different bottlenecks. Number of groups described above is kept at 8.

4.2.3 Simulation Setup

To reduce the number of images from unseen pose sets, synthetic images are generated as stated in the *Dataset and Augmentation* section above and used to add synthetic images to the original datasets, ensuring that the pose datasets for Caspeal and BIWI datasets have maximum of 10 images per person per pose and a balance of synthetic and profile images. For Megaface [116] training data, it was ensured that each pose contains two synthetic images along with any profile images present for that pose. Face images that were closest to the exact front pose were used to construct the synthetic photographs. In the test case, no synthetic images were included.

Face images are separated into five subsections based on pose angles matching to yaw position during training, and then further divided into three subsections for pitch values. This can be seen in Table 4.4. As photos with yaw values less than -18° were flipped horizontally, a yaw was finally divided into -18° to 18° , 18° to 54° , and above yaw values. Thus, the first and second columns of the table are flipped and added to the 5th and the 4th columns respectively.

Table 4.4: Pose categories for classification before flipping images

Pose categories				
Yaw: below -54° Pitch: above 15°	Yaw: -54° to -18° Pitch: above 15°	Yaw: -18° to 18° Pitch: above 15°	Yaw: 18° to 54° Pitch: above 15°	Yaw: above 54° Pitch: above 15°
Yaw: below -54° Pitch: -15° to 15°	Yaw: -54° to -18° Pitch: -15° to 15°	Yaw: -18° to 18° Pitch: -15° to 15°	Yaw: 18° to 54° Pitch: -15° to 15°	Yaw: above 54° Pitch: -15° to 15°
Yaw: below -54° Pitch: below 15°	Yaw: -54° to -18° Pitch: below 15°	Yaw: -18° to 18° Pitch: below 15°	Yaw: 18° to 54° Pitch: below 15°	Yaw: above 54° Pitch: below 15°

For pitch, thresholds of $+15^\circ$ and -15° were established, with three categories defining any pose that falls within, above, or below this range. There are nine subcategories as a result of this. Random roll values within range of $\pm 15^\circ$ were also introduced while generating the synthetic data. A unique classifier was trained for each of the subsections. The pose of observed faces was determined during inference, and inference was limited to the classifier that corresponded to the pose values.

Pose estimate was performed after the cropped face photos were resized to $224 * 224$ px. This was done as the pose estimation model used here was trained on this size. For training it was kept at $96*96$. The reduced image size necessitates a reduction in the number of FLOPS required. For training, these photos were sent into the network matching to the pose. During inference, the same technique was utilized to choose which classifier to use based on the pose information.

The classification network was trained with batch size of 256, weight decay of 10^{-4} , momentum of 0.9, and cosine learning rate annealing from 0.1 to 0.

4.2.4 Performance Measures

Machine Learning tasks make use of a number of measures to be able to assess the effectiveness of the system and to compare against other systems with similar purpose. Since the base paper along with most of the other recent works of deep learning implementations make comparisons based on accuracy, this work also uses the same performance measure.

Terms like true positive (TP), false positive (FP), true negative (TN), and false negative (FN) describe if the classification was correct (true) or false for positive and negative labels of binary classification. In this case the formula can be given as:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4.14)$$

In summary, correct classifications is being divided with the total. Thus, above equation can be given as below for multiclass classification scenario:

$$accuracy = \frac{\text{correct classifications}}{\text{total classifications}} \quad (4.15)$$

FLOPS, a widely used metric in computational science and high-performance computing, quantifies the computational power of a system. It represents the rate at which a processor can perform floating-point operations per second. FLOPS serves as a crucial performance measure for various applications, including scientific simulations, neural network training, and numerical algorithms.

4.2.5 Results

The system has been tested on FaceScrub [116] test set. A comparison has been made against a number of recent FR models where floating-point values are given in millions in the second column of Table 4.5 and the rank-1 accuracy for each model is given in the third column.

Table 4.5: Results on FaceScrub dataset

Method	Year	MFLOPS	Rank-1 %
MobileFaceNet[80][119]	2018	933	95.8
ArcFace [86]	2019	24211	98.35
AirFace [120]	2019	1000	98.04
VarGFaceNet [87]	2019	1022	94.90
MixFaceNet [121]	2021	626	94.95
RepMLP-FaceRes [118]	2021	406	95.90
ProxylessFaceNAS [119]	2021	900	82.1
PocketNetM[122]	2022	1099	94.13
Proposed method		856	98.13

In the experiments, lower computational cost in terms of FLOPS was achieved than most modern networks, with the exception of MixFaceNet and RepMLP-FaceRes, which have slightly lower MFLOPS (626 and 406 respectively). However, it is important to note that while these two networks have lower computational cost, their accuracy is also slightly lower (around 95%) when compared to the proposed method (accuracy: 98.13%, MFLOPS: 856). This suggests that the proposed method is able to achieve a balance between accuracy and computational cost, providing high accuracy while still maintaining a relatively low computational cost.

The higher accuracy of this proposed method may make it more suitable for applications where accuracy is critical, while the relatively low computational cost may make it more accessible for use in systems with limited computational resources. Additionally, there may be trade-offs between accuracy and computational cost that need to be considered when choosing between these different methods. For example, while ArcFace [86] has higher accuracy than the proposed method, it also requires multiple times the FLOPS (24211 in contrast to 856) during inference, which may make it less suitable for use in systems with limited computational resources.

4.3 Chapter Summary

In this chapter, a novel deep-learning based approach for face verification is presented. The approach aims to address the issue of high time complexity in existing state-of-the-art systems by achieving near accuracy while keeping computational cost to a minimum. The technique focuses on pose-specific classification, as face poses are one of the major causes of complexity in these systems. A framework is devised where pose-specific classification is used, and an approach for data augmentation is suggested and tested using synthetic images.

The results obtained show that the proposed system has the potential to achieve comparable results to existing systems while maintaining low computational cost. This could have significant implications for the development of more efficient and effective face verification systems, particularly in applications where computational resources are limited.

Chapter 5: Conclusions and Future Work

Conclusions and Future Work

5.1 Conclusions

This work addresses the problem of automatic recognition of faces for still images and video frames. The main focus of this research remains on reducing the computational cost of the system. Pose variations being one of the main issues becoming reason for acceptable accuracy being achievable only with the help of highly complex systems in general remained the main focus of attention throughout the research. Different strategies have been devised that involve machine learning approaches for solving the problem, and each one is tested with selected evaluation metrics. The research is driven by a hypothesis that pose being one of the major causes of intraclass variability and thus complexity of the FR systems, mitigating this problem will enable achieving superior results with the help of systems requiring lesser computational power. Two different architectures have been devised that resolve similar problems with the first one based on traditional approaches in FR systems and the second utilizing an approach based on CNN. The latter also resolves some of the issues that were faced during the first; however, the first approach is where the more extensive experimentation was performed.

5.1.1 First Approach

In the research based on traditional ML techniques, advantage was taken from the addition of a pose estimation module. This pose estimation is based on landmark estimation and provides better means for accurate background removal. Face images of each person are divided into predefined pose sets and saved with the information. This allows for the utilization of the proposed pose-specific classification system. The best illumination normalization approach, according to the experiments that gave the most superior performance with the used features, was also utilized. A voting scheme for classification that is lightweight and proved to be better than the famous bagging classification, which is another method that uses voting, was proposed. To assess the validity of the approach devised in this research, several performance metrics were used. The results conclude that the pose-specific classification system indeed improves the results to a great extent while using simple classifiers when compared to a system without the use of this approach.

5.1.2 Second Approach

In the second approach, considering a change of trend in the state-of-art and high effectiveness of the CNN based techniques, an architecture based on CNNs was proposed. However, as the emphasis has always remained on keeping computational cost low, the focus is followed on pose-specific classification and a method for data augmentation is devised to generate and utilize synthetic images to attain sufficient samples for each pose. The mix of original and synthetic images is subject to division into pose categories. A very lightweight two-step architecture is followed where the pose is first estimated and then classification is performed by the classifier responsible for the pose set that the input image belongs to. The experimentation proves that a lightweight model such as this one is indeed able to attain comparable results to the state of the art.

5.2 Future Work

Face recognition and verification tasks have seen tremendous improvements in the past decades. While, traditional machine learning approaches improved the accuracies to a great level, the advent of deep learning techniques, particularly those involving CNNs, have risen it to a remarkable level. Difficult datasets like LFW which have been developed in unconstrained or minimally constrained environments, have been conquered with systems approaching accuracies above 99%. However, certain areas of the field still need attention and have great space for improvement.

One of the great challenges that this field is still facing is occlusions. The recent years have seen a great necessity for facial masks to cover mouth and nose in order to prevent spread of Covid. This has further enhanced the need to FR systems that can detect faces partially covered by these masks. In addition to this, facial occlusions may encompass certain other scenarios like sunglasses and any random object like one's hand partially covering the face. Work is already being done to solve this problem but the systems are far from the accuracy that is needed by demanding applications of face recognition systems.

Yet another important area in the field of face recognition and verification is to detect uncommon variations in the face. Certain plastic surgery caused variations like nose reshape may make a person's face appear somewhat different. This often renders the common face recognition systems incapable of reliably identifying the person. Robustness against such changes still needs to be given a great level of attention.

Lastly, there is ongoing research where attempts are made and methods developed as a result to counter identifiability of a face by FR systems. A good example of such methods is special makeups that prove successful in dodging these systems. A robust system against such approaches may be needed in those applications of FR where a person may not wish to be identified.

Chapter 6: References

References

- [1] Boulos, Mira M., "Facial Recognition and Face Mask Detection Using Machine Learning Techniques" (2021). *Theses, Dissertations and Culminating Projects*. 728. <https://digitalcommons.montclair.edu/etd/728>
- [2] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, Present, and Future of Face Recognition: A Review." *Electronics*, vol. 9, no. 8, p. 1188, 2020, doi: 10.3390/electronics9081188
- [3] W. Ali, W. Tian, S. U. Din, D. Iradukunda, and A. A. Khan, "Classical and modern face recognition approaches: a complete review." *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 4825-4880, 2020, doi: 10.1007/s11042-020-09850-1
- [4] W. N. I. Al-Obaydy and S. A. Suandi, "Open-set single-sample face recognition in video surveillance using fuzzy ARTMAP." *Neural Computing and Applications*, vol. 32, no. 5, pp. 1405-1412, 2018, doi: 10.1007/s00521-018-3649-0.
- [5] R. Roady, T. L. Hayes, R. Kemker, A. Gonzales, and C. Kanan, "Are open set classification methods effective on large-scale datasets?" *PLOS ONE*, vol. 15, no. 9, 2020, doi: 10.1371/journal.pone.0238302.
- [6] Granger, Eric, Madhu Kiran, and Louis-Antoine Blais-Morin. "A comparison of CNN-based face and head detectors for real-time video surveillance applications." *2017 seventh international conference on image processing theory, tools and applications (IPTA)*. IEEE, 2017.
- [7] Taherkhani, Fariborz, et al. "Profile to frontal face recognition in the wild using coupled conditional generative adversarial network." *IET Biometrics* 11.3 (2022): 260-276.
- [8] Kortli, Yassin, et al. "Face recognition systems: A survey." *Sensors* 20.2 (2020): 342.
- [9] T. Napoléon and A. Alfalou, "Pose invariant face recognition: 3D model from single photo." *Optics and Lasers in Engineering*, vol. 89, pp. 150-161, 2017, doi: 10.1016/j.optlaseng.2016.06.019.
- [10] Y. Kortli, M. Jridi, A. A. Falou, and M. Atri, "A novel face detection approach using local binary pattern histogram and support vector machine." *2018 International Conference on Advanced Systems and Electric Technologies (IC_ASET)*, 2018, doi: 10.1109/aset.2018.8379829.

- [11] A. Alfalou, Y. Ouerhani, and C. Brosseau, "Road mark recognition using HOG-SVM and correlation." *Optics and Photonics for Information Processing XI*, 2017, doi: 10.1117/12.2273304.
- [12] J. Rettkowski, A. Boutros, and D. Göhringer, "HW/SW Co-Design of the HOG algorithm on a Xilinx Zynq SoC." *Journal of Parallel and Distributed Computing*, vol. 109, pp. 50-62, 2017, doi: 10.1016/j.jpdc.2017.05.005.
- [13] D. Benarab, T. Napoléon, A. Alfalou, A. Verney, and P. Hellard, "Optimized swimmer tracking system by a dynamic fusion of correlation and color histogram techniques." *Optics Communications*, vol. 356, pp. 256-268, 2015, doi: 10.1016/j.optcom.2015.07.056.
- [14] K. Konstantinidis, A. Gasteratos, and I. Andreadis, "Image retrieval based on fuzzy color histogram processing." *Optics Communications*, vol. 248, no. 4, pp. 375-386, 2005, doi: 10.1016/j.optcom.2004.12.029.
- [15] Pasandi, Mohammad Esmaeel Mousa. "Face, Age and Gender Recognition Using Local Descriptors". *Diss University of Ottawa*, 2014
- [16] P. Khoi, L. Huu, and V. Hoai, "Face Retrieval Based On Local Binary Pattern and Its Variants: A Comprehensive Study." *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 6, 2016, doi: 10.14569/ijacsa.2016.070632.
- [17] Xi, Meng, et al. "Local binary pattern network: A deep learning approach for face recognition." *2016 IEEE international conference on Image processing (ICIP)*. IEEE, 2016.
- [18] C. Guo, "Enhancing Face Identification Using Local Binary Patterns and K-Nearest Neighbors." *Journal of Imaging*, vol. 3, no. 3, p. 37, 2017, doi: 10.3390/jimaging3030037.
- [19] K. Bonnen, B. F. Klare, and A. K. Jain, "Component-Based Representation in Automated Face Recognition." *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 239-253, 2013, doi: 10.1109/tifs.2012.2226580.
- [20] J. Ren, X. Jiang, and J. Yuan, "Relaxed local ternary pattern for face recognition." *2013 IEEE International Conference on Image Processing*, 2013, doi: 10.1109/icip.2013.6738759.
- [21] P. Khoi, L. Huu, and V. Hoai, "Face Retrieval Based On Local Binary Pattern and Its Variants: A Comprehensive Study." *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 6, 2016, doi: 10.14569/ijacsa.2016.070632.
- [22] M. Karaaba, O. Surinta, L. Schomaker, and M. A. Wiering, "Robust Face Recognition by Computing Distances From Multiple Histograms of Oriented Gradients." *2015 IEEE Symposium Series on Computational Intelligence*, 2015, doi: 10.1109/ssci.2015.39.
- [23] A. HajiRassouliha, T. P. B. Gamage, M. D. Parker, M. P. Nash, A. J. Taberner, and P. M. F. Nielsen, "FPGA implementation of 2D cross-correlation for real-time 3D tracking of

- deformable surfaces." *2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013)*, 2013, doi: 10.1109/ivcnz.2013.6727067.
- [24] Huang, Chunde, and Jiaxiang Huang. "A fast HOG descriptor using lookup table and integral image." *arXiv preprint arXiv:1703.06256* (2017).
- [25] O. A. Arigbabu, S. M. Syed Ahmad, W. A. Wan Adnan, and S. Mahmood, "SOFT BIOMETRICS: GENDER RECOGNITION FROM UNCONSTRAINED FACE IMAGES USING LOCAL FEATURE DESCRIPTOR." *Journal of Information and Communication Technology*, 2015, doi: 10.32890/jict2015.14.0.8159.
- [26] A. Alfalou and C. Brosseau, "Understanding Correlation Techniques for Face Recognition: From Basics to Applications." *Face Recognition*, 2010, doi: 10.5772/8935.
- [27] C. S. Weaver and J. W. Goodman, "A Technique for Optically Convolution Two Functions." *Applied Optics*, vol. 5, no. 7, p. 1248, 1966, doi: 10.1364/ao.5.001248.
- [28] A. Lugt, "Signal detection by complex spatial filtering." *IEEE Transactions on Information Theory*, vol. 10, no. 2, pp 139-145, 1964, doi: 10.1109/tit 1964 1053650.
- [29] M. Jridi, T. Napoléon, and A. Alfalou, "One lens optical correlation: application to face recognition." *Applied Optics*, vol. 57, no. 9, p. 2087, 2018, doi: 10.1364/ao.57.002087.
- [30] Y. Ouerhani, M. Jridi, and A. Alfalou, "Fast face recognition approach using a graphical processing unit "GPU"." *2010 IEEE International Conference on Imaging Systems and Techniques*, 2010, doi: 10.1109/ist.2010.5548545.
- [31] M. Zeppelzauer, "Automated detection of elephants in wildlife video." *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, 2013, doi: 10.1186/1687-5281-2013-46.
- [32] J. L. Horner and P. D. Gianino, "Phase-only matched filtering." *Applied Optics*, vol. 23, no. 6, p. 812, 1984, doi: 10.1364/ao.23.000812.
- [33] Leonard, Isabelle, Ayman Alfalou, and Christian Brosseau. "Face recognition based on composite correlation filters: analysis of their performances." (2012): 57-80.
- [34] P. Katz, M. Aron, and A. Alfalou, "A face-tracking system to detect falls in the elderly." *SPIE Newsroom*, 2013, doi: 10.1117/2.1201307.004994.
- [35] A. Alfalou, C. Brosseau, P. Katz, and M. S. Alam, "Decision optimization for face recognition based on an alternate correlation plane quantification metric." *Optics Letters*, vol. 37, no. 9, p. 1562, 2012, doi: 10.1364/ol 37.001562.
- [36] M. ELBOUZ, F. Bouzidi, A. Alfalou, C. Brosseau, I. Leonard, and B.-E. Benkelfat, "Adapted all-numerical correlator for face recognition applications." *SPIE Proceedings*, 2013, doi: 10.1117/12.2014383.

- [37] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a Local Binary Descriptor Very Fast." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281-1298, 2012, doi: 10.1109/tpami.2011.222.
- [38] B. Heflin, W. Scheirer, and T. Boult, "For your eyes only." *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, 2012, doi: 10.1109/wacv.2012.6163013.
- [39] X. Zhu, S. Liao, Z. Lei, R. Liu, and S. Z. Li, "Feature Correlation Filter for Face Recognition." *Advances in Biometrics*, pp. 77-86, doi: 10.1007/978-3-540-74549-5_9.
- [40] L. Lenc and P. Král, "Automatic face recognition system based on the SIFT features." *Computers & Electrical Engineering*, vol. 46, pp. 256-272, 2015, doi: 10.1016/j.compeleceng.2015.01.014.
- [41] Şahin Işık, "A Comparative Evaluation of Well-known Feature Detectors and Descriptors." *International Journal of Applied Mathematics, Electronics and Computers*, vol. 3, no. 1, p. 1, 2014, doi: 10.18100/ijamec.60004.
- [42] Du, G.; Su, F.; Cai, A. Face recognition using SURF features In *MIPPR 2009: Pattern Recognition and Computer Vision*; International Society for Optics and Photonics; SPIE: Bellingham, WA, USA, 2009; Volume 7496, p. 749628.
- [43] Vinay, A.; Hebbar, D.; Shekhar, V.S.; Murthy, K.B.; Natarajan, S. Two novel detector-descriptor based approaches for face recognition using sift and surf. *Procedia Comput. Sci.* **2015**, *70*, 185–197.
- [44] Calonder, M.; Lepetit, V.; Ozuysal, M.; Trzcinski, T.; Strecha, C.; Fua, P. BRIEF: Computing a local binary descriptor very fast. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1281–1298.
- [45] H. J. Seo and P. Milanfar, "Face Verification Using the LARK Representation." *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 4, pp. 1275-1286, 2011, doi: 10.1109/tifs.2011.2159205.
- [46] A. Lima, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "On the use of kernel PCA for feature extraction in speech recognition." *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 2003, doi: 10.21437/eurospeech.2003-704.
- [47] M. Turk and A. Pentland, "Eigenfaces for Recognition." *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991, doi: 10.1162/jocn.1991.3.1.71.
- [48] B. Jyostna Devi, N. Veeranjanyulu, and K. Kishore, "A novel face recognition system based on combining eigenfaces with fisher faces using wavelets." *Procedia Computer Science*, vol. 2, pp. 44-51, 2010, doi: 10.1016/j.procs.2010.11.007.

- [49] K. Simonyan, O. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher Vector Faces in the Wild." *Proceedings of the British Machine Vision Conference 2013*, 2013, doi: 10.5244/c.27.8.
- [50] B. Li and K.-K. Ma, "Fisherface vs. Eigenface in the Dual-tree Complex Wavelet Domain." *2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2009, doi: 10.1109/ih-msp.2009.322.
- [51] R. Agarwal, R. Jain, R. Regunathan, and C. S. Pavan Kumar, "Automatic Attendance System Using Face Recognition Technique." *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology*, pp. 525-533, 2018, doi: 10.1007/978-981-13-1610-4_53.
- [52] M. Annalakshmi, S. M. M. Roomi, and A. S. Naveedh. "A hybrid technique for gender classification with SLBP and HOG features." *Cluster Computing*, vol. 22, pp. 11-20, 2018, doi: 10.1007/s10586-017-1585-x.
- [53] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing Robust Face Region Descriptors via Multiple Metric Learning for Face Recognition in the Wild." *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, doi: 10.1109/cvpr.2013.456.
- [54] Peng Li, Yun Fu, U. Mohammed, J. H. Elder, and S. J. D. Prince, "Probabilistic Models for Inference about Identity." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 144-157, 2012, doi: 10.1109/tpami.2011.104.
- [55] V. Perlibakas, "Face Recognition Using Principal Component Analysis and Wavelet Packet Decomposition." *Informatica*, vol. 15, no. 2, pp. 243-250, 2004, doi: 10.15388/informatica.2004.057.
- [56] Z.-H. Huang, W.-J. Li, J. Shang, J. Wang, and T. Zhang, "Non-uniform patch based face recognition via 2D-DWT." *Image and Vision Computing*, vol. 37, pp. 12-19, 2015, doi: 10.1016/j.imavis.2014.12.005.
- [57] Z. Sufyanu, F. S. Mohamad, and A. A. Yusuf, "A New Discrete Cosine Transform on Face Recognition through Histograms for an Optimized Compression." *Research Journal of Information Technology*, vol. 7, no. 2, pp. 101-111, 2015, doi: 10.3923/rjit.2015.101.111.
- [58] S. Azuan and M. Khali, "PCA-ANN Face Recognition System based on Photometric Normalization Techniques." *State of the Art in Face Recognition*, 2009, doi: 10.5772/6640.
- [59] H. Hoffmann, "Kernel PCA for novelty detection." *Pattern Recognition*, vol. 40, no. 3, pp. 863-874, 2007, doi: 10.1016/j.patcog.2006.07.009.
- [60] S. R. Arashloo and J. Kittler, "Class-Specific Kernel Fusion of Multiple Descriptors for Face Verification Using Multiscale Binarised Statistical Image Features." *IEEE Transactions*

- on *Information Forensics and Security*, vol. 9, no. 12, pp. 2100-2109, 2014, doi: 10.1109/tifs.2014.2359587.
- [61] Ms. Snehal Houshiram GordeI, et al. A Review on Face Recognition Algorithms Volume III, Issue I Issn No.:2350-1146, I.F-2.71
- [62] Xie, Jianhong. "Face recognition based on Curvelet transform and LSSVM." *Proceedings of the 2009 (ISIP'09) Huangshan*,
- [63] Hu, S., Lu, X., Ye, M., et al., 'Singular value decomposition and local near neighbors for face recognition', *Pattern Recognit.*, 2017, 64, (1), pp. 60–83
- [64] A. A. Fathima, S. Ajitha, V. Vaidehi, M. Hemalatha, R. Karthigaiveni, and R. Kumar, "Hybrid approach for face recognition combining Gabor Wavelet and Linear Discriminant Analysis." *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, 2015, doi: 10.1109/cgvis.2015.7449925.
- [65] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz, "Fast High Dimensional Vector Multiplication Face Recognition." *2013 IEEE International Conference on Computer Vision*, 2013, doi: 10.1109/iccv.2013.246.
- [66] F. Juefei-Xu, K. Luu, and M. Savvides, "Spartans: Single-Sample Periocular-Based Alignment-Robust Recognition Technique Applied to Non-Frontal Scenarios." *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4780-4795, 2015, doi: 10.1109/tip.2015.2468173.
- [67] Y. Yan, H. Wang, and D. Suter, "Multi-subregion based correlation filter bank for robust face recognition." *Pattern Recognition*, vol. 47, no. 11, pp. 3487-3501, 2014, doi: 10.1016/j.patcog.2014.05.004.
- [68] C. Ding and D. Tao, "Robust Face Recognition via Multimodal Deep Face Representation." *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2049-2058, 2015, doi: 10.1109/tmm.2015.2477042.
- [69] R. Sharma and M. Patterh, "A new pose invariant face recognition system using PCA and ANFIS." *Optik*, vol. 126, no. 23, pp. 3483-3487, 2015, doi: 10.1016/j.ijleo.2015.08.205.
- [70] M. MOUSSA, M. HMILA, and A. DOUIK, "A Novel Face Recognition Approach Based on Genetic Algorithm Optimization." *Studies in Informatics and Control*, vol. 27, no. 1, 2018, doi: 10.24846/v27i1y201813.
- [71] A. Mian, M. Bennamoun, and R. Owens, "An Efficient Multimodal 2D-3D Hybrid Approach to Automatic Face Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1927-1943, 2007, doi: 10.1109/tpami.2007.1105.

- [72] H. Cho, R. Roberts, B. Jung, O. Choi, and S. Moon, "An Efficient Hybrid Face Recognition Algorithm Using PCA and GABOR Wavelets." *International Journal of Advanced Robotic Systems*, vol. 11, no. 4, p. 59, 2014, doi: 10.5772/58473.
- [73] D. Guru, M. Suraj, and S. Manjunath, "Fusion of covariance matrices of PCA and FLD." *Pattern Recognition Letters*, vol. 32, no. 3, pp. 432-440, 2011, doi: 10.1016/j.patrec.2010.10.006.
- [74] J. K. Sing, S. Chowdhury, D. K. Basu, and M. Nasipuri, "An improved hybrid approach to face recognition by fusing local and global discriminant features." *International Journal of Biometrics*, vol. 4, no. 2, p. 144, 2012, doi: 10.1504/ijbm.2012.046245.
- [75] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE." *Image and Vision Computing*, vol. 28, no. 5, pp. 807-813, 2010, doi: 10.1016/j.imavis.2009.08.002.
- [76] P. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms." *Image and Vision Computing*, vol. 16, no. 5, pp. 295-306, 1998, doi: 10.1016/s0262-8856(97)00070-x.
- [77] J. Mahier, M. El-Abed, B. Hemery, and C. Rosenberger, "Toward a distributed benchmarking tool for biometrics." *2011 International Conference on High Performance Computing & Simulation*, 2011, doi: 10.1109/hpcsim.2011.5999891.
- [78] P. Kamencay, M. Breznan, D. Jelsovka, and M. Zachariasova, "Improved face recognition method based on segmentation algorithm using SIFT-PCA." *2012 35th International Conference on Telecommunications and Signal Processing (TSP)*, 2012, doi: 10.1109/tsp.2012.6256399.
- [79] Licheng Jiao, Liefeng Bo, and Ling Wang, "Fast Sparse Approximation for Least Squares Support Vector Machine." *IEEE Transactions on Neural Networks*, vol. 18, no. 3, pp. 685-697, 2007, doi: 10.1109/tnn.2006.889500.
- [80] Chen, Sheng, et al. "Mobilefacenet: Efficient cnns for accurate real-time face verification on mobile devices." *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*. Springer International Publishing, 2018.
- [81] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." (2015).
- [82] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, doi: 10.1109/cvpr.2015.7298682.

- [83] Huang, Gary B., et al. "Labeled faces in the wild: A database for studying face recognition in unconstrained environments." Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition. 2008
- [84] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity." *CVPR 2011*, 2011, doi: 10.1109/cvpr.2011.5995566.
- [85] Cao, Qiong, et al. "Vggface2: A dataset for recognising faces across pose and age." *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018.
- [86] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition." *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, doi: 10.1109/cvpr.2019.00482.
- [87] Yan, Mengjia, et al. "Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition." Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019.
- [88] A. Zadeh, Y. C. Lim, T. Baltrusaitis, and L.-P. Morency, "Convolutional Experts Constrained Local Model for 3D Facial Landmark Detection." *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, doi: 10.1109/iccvw.2017.296.
- [89] Pagano, E. Granger, R. Sabourin, G. Marcialis, and F. Roli, "Adaptive ensembles for face recognition in changing video surveillance environments." *Information Sciences*, vol. 286, pp. 75-101, 2014, doi: 10.1016/j.ins.2014.07.005.
- [90] Jones, Michael, and Paul Viola. "Fast multi-view face detection." Mitsubishi Electric Research Lab TR-20003-96 3.14 (2003): 2.
- [91] Shiguang Shan, W. Gao, Bo Cao, and Debin Zhao, "Illumination normalization for robust face recognition against varying lighting conditions." *2003 IEEE International SOI Conference. Proceedings (Cat. No.03CH37443)*, 2003, doi: 10.1109/amfg.2003.1240838.
- [92] Ketcham, David J., R. W. Lowe, and J. W. Weber. "Real-time image enhancement techniques." *Seminar on Image Processing*. Vol. 74. Pacific Grove California, Hughes Aircraft Company, 1976.
- [93] C.-K. Tran, D.-T. Pham, C.-D. Tseng, and T.-F. Lee, "Face Recognition under Lighting Variation Conditions Using Tan-Triggs Method and Local Intensity Area Descriptor." *Advances in Intelligent Systems and Computing*, pp. 84-92, 2017, doi: 10.1007/978-981-10-6487-6_11.

- [94] Xiaoyang 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, doi: 10.1109/tip.2010.2042645.
- [95] H. Zhang, S. Wang, X. Xu, T. W. S. Chow, and Q. M. J. Wu, "Tree2Vector: Learning a Vectorial Representation for Tree-Structured Data." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5304-5318, 2018, doi: 10.1109/tnnls.2018.2797060.
- [96] R. M. Anwer, F. S. Khan, J. Van de Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification." *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 138, pp. 74-85, 2018, doi: 10.1016/j.isprsjprs.2018.01.023.
- [97] T. De Freitas Pereira, A. Anjos, and S. Marcel, "Heterogeneous Face Recognition Using Domain Specific Units." *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1803-1816, 2019, doi: 10.1109/tifs.2018.2885284
- [98] X. Liu, L. Lu, Z. Shen, and K. Lu, "A novel face recognition algorithm via weighted kernel sparse representation." *Future Generation Computer Systems*, vol. 80, pp. 653-663, 2018, doi: 10.1016/j.future.2016.07.007.
- [99] A. M. Reza, "Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement." *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, vol. 38, no. 1, pp. 35-44, 2004, doi: 10.1023/b:vlsi.0000028532.53893.82.
- [100] R. Bendjillali, M. Beladgham, K. Merit, and A. Taleb-Ahmed, "Improved Facial Expression Recognition Based on DWT Feature for Deep CNN." *Electronics*, vol. 8, no. 3, p. 324, 2019, doi: 10.3390/electronics8030324.
- [101] O. C. Abikoye, I. F. Shoyemi, and T. O. Aro, "Comparative Analysis of Illumination Normalizations on Principal Component Analysis Based Feature Extraction for Face Recognition." *FUOYE Journal of Engineering and Technology*, vol. 4, no. 1, 2019, doi: 10.46792/fuoyejt.v4i1.309.
- [102] S. M. Pizer, "Adaptive histogram equalization and its variations." *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355-368, 1987, doi: 10.1016/s0734-189x(87)80186-x.
- [103] A.-W. Deng, C.-H. Wei, and C.-Y. Gwo, "Stable, fast computation of high-order Zernike moments using a recursive method." *Pattern Recognition*, vol. 56, pp. 16-25, 2016, doi: 10.1016/j.patcog.2016.02.014.

- [104] Basaran, Emrah, Muhittin Gökmen, and Mustafa E. Kamasak. "An efficient multiscale scheme using local zernike moments for face recognition." *Applied Sciences* 8.5 (2018): 827.
- [105] Ouanan, Hamid, et al. "A novel face recognition system based on Gabor and Zernike features." *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019) Volume 5-Advances Intelligent Systems for Multimedia Processing and Mathematical Modeling*. Springer International Publishing, 2020.
- [106] L. Breiman, "Bagging predictors." *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996, doi: 10.1007/bf00058655.
- [107] M. A. Yaman, F. Rattay, and A. Subasi, "Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Face Recognition." *Procedia Computer Science*, vol. 194, pp. 202-209, 2021, doi: 10.1016/j.procs.2021.10.074.
- [108] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random Forests for Real Time 3D Face Analysis." *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437-458, 2012, doi: 10.1007/s11263-012-0549-0.
- [109] Sisodia, Dilip Singh, Ram Bilas Pachori, and Lalit Garg, eds. "*Handbook of Research on Advancements of Artificial Intelligence in Healthcare Engineering*." *Advances in Healthcare Information Systems and Administration*, 2020, doi: 10.4018/978-1-7998-2120-5.
- [110] Zhou, Yijun, and James Gregson. "WHENet: Real-time Fine-Grained Estimation for Wide Range Head Pose." arXiv preprint arXiv:2005.10353 (2020).
- [111] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation From a Single Image." *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, doi: 10.1109/cvpr.2019.00118
- [112] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation." *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, doi: 10.1109/cvpr46437.2021.00753.
- [113] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-Grained Head Pose Estimation Without Keypoints." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, doi: 10.1109/cvprw.2018.00281.
- [114] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network." *Computer Vision – ECCV 2018*, pp. 557-574, 2018, doi: 10.1007/978-3-030-01264-9_33.

- [115] Wen Gao, "The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 38, no. 1, pp. 149-161, 2008, doi: 10.1109/tsmca.2007.909557.
- [116] 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)*, 2016, doi: 10.1109/cvpr.2016.527.
- [117] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing Textures in the Wild." *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, doi: 10.1109/cvpr.2014.461.
- [118] Ding, Xiaohan, et al. "Repmlp: Re-parameterizing convolutions into fully-connected layers for image recognition." *arXiv preprint arXiv:2105.01883* (2021).
- [119] Martinez-Diaz, Yoanna, et al. "Benchmarking lightweight face architectures on specific face recognition scenarios." *Artificial Intelligence Review* (2021): 1-44.
- [120] Li, Xianyang, et al. "Airface: Lightweight and efficient model for face recognition." *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019.
- [121] Boutros, F., et al. "MixFaceNets: Extremely efficient face recognition networks. In 2021 IEEE IJCB." *IJCB* (2021): 4-7.
- [122] Boutros, Fadi, et al. "Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation." *IEEE Access* 10 (2022): 46823-46833.

