Biometric Recognition Using 3D Face

Ph.D. Thesis

By

Akhilesh Mohan Srivastava

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

JUNE 2022

Biometric Recognition Using 3D Face

A Thesis

Submitted in partial fulfillment of the
requirements for the award of the degrees
of
Doctor of Philosophy

by
Akhilesh Mohan Srivastava

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
JUNE 2022

INDIAN INSTITUTE OF TECHNOLOGY INDORE

I hereby certify that the work which is being presented in the thesis entitled **Biometric Recognition Using 3D Face** in the partial fulfillment of the requirements for the award of the degree of DOCTOR OF PHILOSOPHY and submitted in the DEPARTMENT/SCHOOL OF **Computer Science and Engineering**, Indian Institute of Technology Indore, is an authentic record of my own work carried out during the time period from **July 2017** to **June 2022** under the supervision of **Dr. Surya Prakash, Associate Professor, Department of Computer Science and Engineering, IIT Indore**.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

27 June 2022

Signature of the student with date

(Akhilesh Mohan Srivastava)

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.

Surya Prakash 27-June-2022

Signature of Thesis Supervisor with date

(Dr. Surya Prakash)

Akhilesh Mohan Srivastava has successfully given his/her Ph.D. Oral Examination held on 02 February 2023.

Surya Prakash 02-Feb-2023

Signature of Thesis Supervisor with date

(Dr. Surya Prakash)

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my heartfelt gratitude to a number of persons who in one or the other way contributed by making this time as learnable, enjoyable, and bearable. At first, I would like to thank my supervisor **Dr. Surya Prakash**, who has been a constant source of inspiration during my work. Without his constant guidance and research directions, this research work would have not been possible. His continuous support and encouragement has always motivated me to remain streamlined in my research work. I am also grateful to him for extending all necessary support to me as Head of the Department of Computer Science & Engineering during his tenure from 2017 to 2020.

I am thankful to **Dr. Somnath Dey** and **Dr. Vivek Kanhangad**, my PG Student Progress Committee (PSPC) members, for taking time out of their busy schedule to evaluate my progress all these years. Their valuable comments and suggestions have helped me to improve my work at various stages. I am also thankful to **Dr. Somnath Dey** for providing all necessary support as current Head of the Department of Computer Science and Engineering.

I would like to appreciate the fine company of my dearest colleagues and friends especially, **Vivek Baghel**, **Sadaf** and **Vikas Chauhan** for their cooperation during the course of my research.

I would like to express my heartfelt respect to my wife **Dr. Garima Srivastava** and my mother **Shakuntala Srivastava** for their love, care and support they have provided to me throughout my life. Special thanks to my M.Tech guide **Professor Maheshwari Tripathi** for their constant support and encouragement.

Finally, I am thankful to God, the most gracious and the most merciful; and all his creatures who directly or indirectly contributed, helped, and supported me. To sign off, I state my quote, which I always tell everyone to motivate them to make this world a better place.

"Having a sharp brain is a gift, having a big heart is a choice."

Akhilesh Mohan Srivastava

Dedicated

to

My family, friends, teachers and IIT Indore

ABSTRACT

Biometrics deals with recognizing a user based on his/her physiological or behavioural characteristics. By applying sensors, both physiological (such as fingerprint, iris, face, ear etc.) as well as behavioural (such as signature, handwriting, voice, etc.) characteristics can be captured from the user. These characteristics are relatively unique, permanent, and difficult to forge and share. Also, a user does not need to worry about forgetting them or making any special effort to carry them along. Among various biometric traits (features), face is the most widely used trait for user identification. The use of human face is becoming popular day by day and many countries are currently using it as a prime tool for the identification/recognition of their citizens.

Face recognition has recently gained considerable attention as one of the most successful applications of image analysis and recognition, especially in the last few years. This can be attributed to at least two factors: the wide spectrum of its real-world applications such as authentication, access control and surveillance, and the availability of viable technologies after several years of continuous studies. However, the existing 2D Human face recognition systems have attained a certain level of maturity, but the limitations imposed by many real-world applications limit their success. The Challenges faced by 2D Human face recognition with changes in illumination and/or position and especially spoofing attacks, for example, are still unsolved. Here 3D face recognition technologies provide a solution to the challenges proposed as it could accurately recognize human faces even in low-light settings, with a variety of facial angles and expressions and can deal with spoofing. But the collection and processing of 3D data is itself a big challenge owing to its high computational power and resource requirements.

Primarily, the work presented in the thesis addresses the issue of scarcity of 3D data, which is causing overfitting in Deep Neural Networks and making training more difficult. To circumvent this issue, the variability of 3D data must be raised by using data augmentation to expand the quantity of available data. Regarding this, the proposed technique uses three different types of sampling approaches for augmenting the 3D data which will make the training of deep neural network model computationally and spatially efficient.

We also prove, using Iterative Closest Point(ICP) and Central Limit Theorem(CLT), that no information is lost while sub-sampling the point clouds. In this work, we are proposing a technique that uniquely combines an efficient 3D object recognition architecture with a one-shot learning network using 3D data augmentation and deep learning. The technique successfully classifies and recognize when applied for the recognition of object classes, which are very similar to each other like faces and ears. Furthermore, to solve high computational power and resource requirements for 3D data, we propose a technique which makes it relatively easy to work with in terms of computational power and time requirements. The proposed method is built on ResNet-34 with Siamese Network model and using transfer learning approach, attains high accuracy in less training time.

LIST OF PUBLICATIONS

(A) From PhD thesis work:

A1. Journal Articles:

Published/Accepted:

J1. **Akhilesh M Srivastava**, Arushi Jain, Priyanka Rotte, Surya Prakash and Umarani Jayaraman, A Technique to Match Highly Similar 3D Objects with an Application to Biomedical Security, Multimedia Tools and Applications, Volume 81, Issue 10, pages: 13159–13178 (2022). DOI :10.1007/s11042-020-10161-8

J2. **Akhilesh M Srivastava** , Priyanka Ajay Rotte, Arushi Jain, Surya Prakash, Handling Data Scarcity through Data Augmentation in Training of Deep Neural Networks for 3D Data Processing, International Journal on Semantic Web and Information Systems (IJSWIS), Volume 18, Issue 1, Article 14, pages: 1-16 (2022). DOI: 10.4018/IJSWIS.297038

J3. **Akhilesh Mohan Srivastava**, Sai Dinesh Chintaginjala1, Samhit Chowdary Bhogavalli1, Surya Prakash, Robust Face Recognition using Multimodal Data and Transfer Learning, Journal of Electronic Imaging, Volume 32, Issue 4, SN. 042105 (2022). DOI: 10.1117/1.JEI.32.4.042105

(B) Other publications during PhD:

B1. Journal Articles:

Published/Accepted:

J1. Iyyakutti Iyappan G, Surya Prakash, Syed Sadaf Ali, Piyush Joshi, Ishan R Dave and **Akhilesh Mohan Srivastava**, ”Ear Recognition in 3D using 2D Curvilinear Features,” *IET Biometrics*, Volume 7, Issue 6, pages: 519-529 (2018). DOI :10.1049/ietbmt.2018.5064.

J2. Anagha R Bhople, **Akhilesh Mohan Shrivastava**, Surya Prakash, ”Point cloud based deep convolutional neural network for 3D face recognition,” *Multimedia Tools and*

Applications, Volume 80, Issue 20, pages: 1573-7721 (2020). DOI 10.1007/s11042-020-09008-z.

B2. Conference Articles:

C1. Aditi Agrawal, Mahak Garg, Surya Prakash, Piyush Joshi, **Akhilesh M. Srivastava**, Hand Down, Face Up: Innovative Mobile Attendance System using Face Recognition and Deep Learning, In Proc. of International Conference on Computer Vision & Image Processing **(CVIP 2018)**, Volume 1024, pages: 363—375, September 29-October 01, 2018, Jabalpur, India.

C2. Mudit Maheshwari, Sanchita Arora, **Akhilesh M. Srivastava**, Aditi Agrawal, Mahak Garg, and Surya Prakash, Earprint based Mobile User Authentication using Convolutional Neural Network and SIFT, In Proc. of Fourteenth International Conference on Intelligent Computing **(ICIC 2018)**, LNCS 10954, pages: 874-880, 15-18 August, 2018, Wuhan, China.

C3. Ishan Dave, Iyyakutti Iyappan Ganapathi, Surya Prakash, Syed Sadaf Ali and **Akhilesh Mohan Srivastava**, ”3D Ear Biometrics: Acquisition and Recognition ,” *15th IEEE India Council Int'l Conference (**INDICON 2018**)*, December 16-18, 2018, Coimbatore, India.

C4. Vivek Singh Baghel, **Akhilesh M. Srivastava**, Surya Prakash, Siddharath Singh, Minutia Points Extractions Using Faster R-CNN, In Proc. 7th International Conference on Advance Computing, Networking and Informatics **(ICACNI 2019)**, Volume 1276, pages:3-10, 20-21 December 2019, Indian Institute of Information Technology, Kalayani, India.

Contents

List of Figures

List of Tables

List of Abbreviations

2D	Two Dimensional
3D	Three Dimensional
2.5D	Depth or range Images
ROC curve	Receiver operating characteristic curve
EUC curve	Error under ROC curve
CMC curve	Cumulative Match Characteristic curve
CNNs	Convolution Neural Networks
FRR	False Rejection Rate
FAR	False Acceptance Rate
EER	Equal Error Rate
GAR	Genuine Acceptance Rate
FNMR	False Non Match Rate
FMR	False Match Rate
UND	University of Notre Dame
IITI Phase-3 Database	Indian Institute of Technology Indore Phase-3 Database
STL	STereoLithography
PCA	Principal Component Analysis
FER	Facial Expression Recognition
GAN	Generative Adversarial Network
ICP	Iterative Closest Point
CLT	Central Limit Theorem
3DMM	3D deformation model
LSVM	Linear Support Vector Machine
LDA	Linear Discriminant Analysis
PDS	Priority-driven Search
MS-LBP)	Multi-scale Local Binary Model
SIFT	Scale Invariant Feature Transform
FRGC v2.0	Face Recognition Grand Challenge Version 2.0
KMTS	Keypoint-based Multiple Triangle Statistics
TPWCRC	Two-Phase Weighted Collaborative Representation Classification

LBP	Local Binary Pattern
SVM	Support Vector Machine
3DFRD	Three-Dimensional Face Depth database
ANN	Artificial Neural Network
AI	Artificial Intelligence
NLP	Natural Language Processing
EEG	Electroencephalogram
NIPS	Neural Information Processing System
VGG	Visual Geometry Group
GPU	Graphics Processing Unit
UND-J2 Dataset	Collection J2, acquired at University of Notre Dame
GTFD Database	Georgia Tech Face-Database
RMS error	Root Mean Squared error
MSE	Mean Squared error
ResNet	Residual Neural Network
ANN	Artificial Neural Network

Chapter 1

Introduction

In the early days of civilization, people lived in small communities and were known to each other individually. They could easily recognize one another without any identity management tool. However, due to the proliferation in population with movement from one place to another, the modern community has prompted a need for well organized and secure identity management system which can efficiently recognize individuals. The system is required to be capable of effectively storing, updating, and deleting individual records (43). The classical recognition methods rely on tokens (such as ID card) and secret knowledge (such as password and PIN). However, these methods have some constraints, for example, ID cards may be lost, stolen, shared whereas password can be forgotten. Additionally, systems based on traditional methods are incapable of distinguishing between a legitimate user and an impostor. Thus, there is a need to overcome these limitations of the conventional recognition methods, which do not require the physical presence of a user at the time of performing the recognition process.

Biometrics can be used to overcome the shortcomings of traditional authentication systems. It refers to a way of automated recognition of a human being based on his/her behavioral or physiological characteristics/features where the physiological (such as, fingerprint, face, ear, etc.) and behavioral (such as, typing rhythm, voice, gait, etc.) features are captured using appropriate sensors. Human biometric traits/features are permanent, unique, and cannot be easily transferred or falsified. Also, biometric authentication is non-repudiable and has been found to be reliable due to which it has become a popular tool for human recognition in current time. The biometric technologies such as face recognition, fingerprint recognition, and iris recognition have been employed in various applications such as identification of staff members in an organization, managing identity information of individuals in public distribution systems, and maintaining patient records in a hospital. In addition to other key applications of biometrics, healthcare fraud prevention, patient privacy enhancement, and healthcare safety are among other several applications where biometrics has played a vital role. A lot of developments are happening in the domain of biometrics and nearly every new technology currently incorporates biometric techniques for a variety of purposes. For example, biometric traits such as fingerprint and face are very routinely employed in mobile phones and laptops to ensure that only the authorized users get access of the devise.

Among various physiological biometric traits such as fingerprint, iris, face, ear, etc., face has been one of the prominent biometrics for human recognition and is found to be having an edge over other biometric modalities due to its non-intrusive nature. The basic idea of face recognition, like in other traits, relies on extracting features from an individual's face and employing them for the recognition. A computer-based program is used for automatic face recognition that uses collected images or video frames to recognize a person. Most of the automated face recognition systems make use of 2D images acquired through cameras for performing the recognition task.

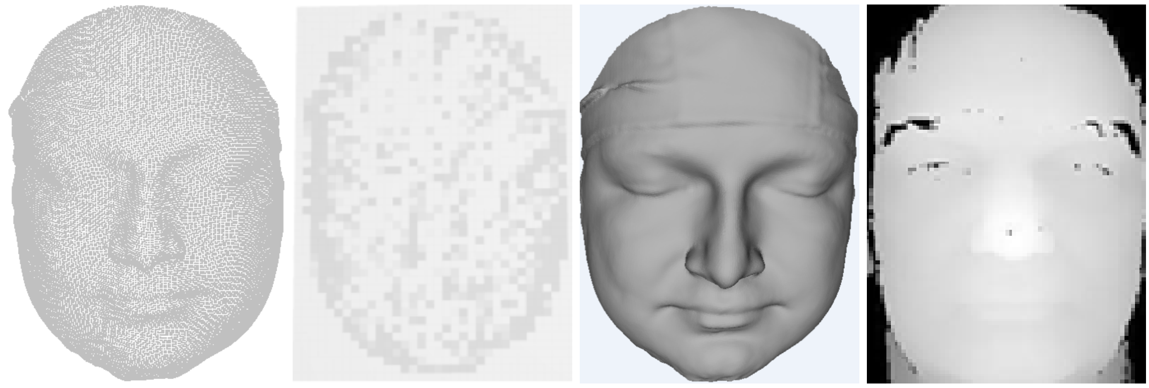
Face recognition is a general phrase used in the biometrics field to refer to two distinct processes: (1) face verification (or authentication) and (2) face identification (or recognition).

- **Face Verification (or authentication)**: Face verification (also known as ”Am I who? I say I am.”) is a one-to-one match procedure in which a probe face image is compared to a template face image whose identity is being asserted. Generally, the receiver operating characteristic (ROC) curve is used to assess the verification performance of a face recognition system. The receiver operating characteristic (ROC) curve plots the genuine acceptance rate (the rate at which legitimate users are allowed access) against the false acceptance rate (the rate at which access has been granted to imposters).

- **Face identification (or recognition)**: In identification (also known as “Who am I?”), the probe image is not labeled with any identity. Therefore, in face identification, a probe face image is matched against all the previously stored face templates in the gallery (reference database) to know the identity of the probe face image. It involves one to many (1:*m*, where *m* is the size of the gallery) matching process where a probe image is located in the gallery by computing similarity score with all face templates of the gallery. An identification procedure returns a sorted list of identities, arranged in the order of best to worst match score. The identification procedure can be either closed-set or open-set. In the closed-set identification, the identity of the input test probe image is known to be present in the gallery. In open-set identification, the identity corresponding to the probe test image may or may not be present in the gallery. Probe images are ranked in order of their scores, and the top k ($k \leq m$) rankings are utilised to estimate the probability that the true matching identity for a probe is detected. The Cumulative Match Characteristic (CMC) curve (refer Figure 1.4) is a common visual representation of these probabilities. The CMC curve illustrates the performance of a biometric system in a closed-set identification problem where X-axis represents the number of top ranks considered. In contrast, the Y-axis represents the probability of the correct identity (the identity of the probe subject) being among the top ranks considered.

Previously, it has been shown that the 2D images are good enough to build object recognition systems. However, recently, 3D data based object recognition systems have seen a rise in popularity over 2D images based systems. Though, the usage of 2D images has produced excellent results in standard settings; the performance of these systems suffer when there is low lighting, pose variations, or occlusions present in the images. These challenges are easily overcome by 3D based face representation, which is capable of capturing all of the geometric information about the face objects. In general, dissimilar unknown objects can be effectively classified and recognised by most of the suggested 3D object recognition approaches (3; 19; 75; 101) in the literature; however, their performance decreases when these techniques are used in the recognition of object classes that are very similar to each another. In this work, a solution has been proposed by developing a general deep network model for 3D object recognition which is capable of matching extremely similar 3D objects, such as 3D faces or ears. Its use has been demonstrated in the application of 3D face recognition in this work. In terms of time and space, the proposed network is quite competent, making it suitable for developing solutions for real-time security applications in a variety of domains.

A 3D face image is an abstract representation of face and can be represented as depth image, point cloud, polygon mesh or voxel. The Figure 1.1 has shown examples of these face representations. These representations have been used in the literature in order to extract the features, and then to perform 3D face recognition. In the depth image based representation, the "depth" or "z" information of an object in the real world is provided. The pixel intensities present in the image show how far the object points are from the camera, that is, from the viewpoint. In surface modelling methods, like mesh, the topological information (connectivity between the points) of the points can be obtained. In point clouds based



(a) point cloud (b) voxel (c) polygon mesh (d) depth image

Figure 1.1: Different representations of 3D face images used in recognition: (a) point cloud, (b) voxel, (c) polygon mesh, (d) depth image.

representation, a 3D object is represented by performing digitization of its surface in the form of an unordered set of data points. In this representation, the data is unstructured and the topological information is absent. The voxel image is a volumetric representation of each point where the change in volume size affects the resolution of the 3D image. Also, this representation does not possess topological information similar to point clouds. Among different representations, point clouds are the most raw form of representation of 3D data and are the directly produced by the 3D scanners.

Volumetric CNNs (60; 71; 95) are the most often used method for processing of 3D data, and their inputs are regularised data in the form of three-dimensional voxels or image grids. This format simplifies weight sharing and other kernel improvements. However, because 3D voxels and grids are so large, they complicate the input rendering process both computationally and physically. As a result of this shortcoming, we have developed a method that can directly consumes the point cloud as input rather than translating it into a standard 3D representation like volumetric representation. Classification and recognition of objects, navigation of mobile robots, etc. are all rapidly emerging areas of research in computer vision that all depend on 3D point cloud data as well as its processing (4; 10; 74) and our developed method has application in all such cases.

The PointNet architecture is a popular framework of convolutional neural networks used in deep learning that it is capable of directly accepting the point clouds as input. The architecture of the PointNet is special in its ability to preserve translational and rotational invariant properties of the point cloud. Though the PointNet is capable of directly consuming point cloud data and in performing classification and recognition tasks, it produces good classification accuracy only when the classes being used in the analysis are very distinct. The performance of the PointNet deteriorates if it is used for comparing the classes which are highly similar like faces of human subjects. The thesis provides a deep neural model which utilizes Siamese Network (31; 94) along with PointNet and is capable of comparing highly similar objects. In the proposed model, PointNet architecture is used for extracting the features of the 3D objects whereas Siamese Network computes a similarity score by comparing a test sample to the reference sample and predicts whether they belong to the same or distinct subject classes. The second-last dense layer of the PointNet architecture is used to extract features from it once the model has been trained on all of the training data from all of the subjects. Subsequently, the collected features are used to train the Siamese Network, which calculates the similarity score between pairs of feature vectors and, lastly, predicts whether two objects supplied into it are members of the same or distinct classes.

Due to the difficulty of gathering a sufficient amount of 3D data, training of the model successfully and without overfitting is difficult. To address the issue of scarcity of 3D data, in this thesis a novel data augmentation strategy for increasing the size of the data is proposed where the augmentation of the data is carried out by generating fixed-size subsets of the input point clouds of the available data samples. The proposed data augmentation preserves all of the characteristics of the original samples. The basic idea behind the proposed data augmentation technique is that the point cloud data generated by the 3D scanners during the scanning process is very bulky and consists of hundreds and thousands of points. Further, in the object recognition process, these many points are not required. So to make the recognition process as efficient as possible in terms of time and space, numerous random subsets of each original point cloud sample are created and utilized. Using these subsets, the size of the database is increased, which is otherwise normally limited due to the fact that usually it contains only a few 3D samples per subject, thus preventing overfitting.

1.1 Motivation and Objectives

The work presented in this thesis belongs to 3D face recognition and derives its motivation from following. First, collecting, modelling, and synthesising realistic 3D human faces and their dynamics has emerged as a hot research topic at the intersection of computer vision and computer graphics. Second, it is one of the most non-intrusive biometric traits available presently and is considered as one of the most attractive biometrics. Face is the most widely used identifier or tokens for representing people. We can see the use of face images in almost all our personal documents; and from this clearly we can understand the relevance of the face as an important biometrics for person recognition. We can obtain a lot of information like age, gender, emotions, expressions, etc. from a human face. Third, while the usage of 2D face images has demonstrated exceptional performance in normal settings, performance degrades significantly in the presence of low illumination, pose changes, and occlusions. These difficulties are overcome by 3D face biometrics, which utilizes 3D face data (which contains geometric information of the face) in lieu of the 2D face images. Fourth, the richness of 3D facial geometrical features of the face is superior to that of 2D face images. This thesis work examines the 3D face data and propose the following effective approaches towards performing the 3D face recognition.

- Motivated by the power of data augmentation three data augmentation techniques have been proposed in this work. These techniques are based on random, systematic and stratified samplings. They help in improving model prediction accuracy as well as reduce the cost of 3D data collection and labeling of the data. By using these techniques, the availability of limited data is increased which in turn gives more variation in 3D point cloud data. To check the efficacy of the generated samples, we check whether the 3D sub-samples created by data augmentation possesses same information as in original samples or not. By applying the Central Limit Theorem over large population, we compute original sample mean and standard deviation values, and compare them with those of the average of the newly created samples through augmentation to check this.

- It is seen that when objects belong to classes that are very similar to each other, the performance of 3D object recognition systems declines. In order to address this issue, an extended model for 3D object recognition has been developed and described in this work. The model combines an efficient object recognition architecture and a one-shot learning network in a novel way. The proposed model is being applied for matching of very similar 3D objects which are 3D faces to demonstrate its effectiveness .

- The challenges faced by 2D face recognition are easily overcome by the 3D face recognition. However, 3D data comes with its own challenges. The computational requirements to process and use 3D data are often cumbersome. To handle these challenges, in this work we suggest the use of 2.5D representation of 3D face data in place of directly using 3D data, along with registered 2D face images. This proposal makes it comparatively much easier to work with in terms of computational power and time requirements. Further, the training time has been further cut-down by the use of transfer learning where a pre-trained deep neural network on 2D face images is used as an initial point while training the deep neural network on 2.5D face data. The use of multi-modal data (2.5D face images along with 2D face images) and transfer learning provides an effective mechanism for face recognition and produces encouraging results. We conduct three experiments to perform the analysis. In Experiments 1 and 2, recognition task is performed using only 2D and 2.5D data, respectively. In Experiment 3, it is performed using multi-modal data and transfer learning.

1.2 Performance Analysis

To evaluate and analyze the strength of the proposed techniques under various scenarios; the following measures, parameters, protocols, and databases are used.

1.2.1 Evaluation Measures

Verification and identification are the essential requirements for the analysis of the recognition performance of any biometric-based authentication technique. A verification system is usually described as a 1-to-1 matching process, as the system tries to match the biometrics of an individual with a similar biometric already in the database. An identification system checks the biometrics of an individual against all others present in the database. It implements 1-to-*n* matching scheme, with *n* being the total number of enrolled subjects in the database. The proposed techniques have been evaluated in terms of following parameters.

- **Verification Accuracy** Verification accuracy and equal error rate (EER) are two common measures used to find out how well a biometric recognition technique is performing the verification task. The verification accuracy is defined as follows.

$$\text{Accuracy} = (100 - \frac{(\text{FAR} + \text{FRR})\%}{2}) \tag{1.1}$$

where, False Acceptance and False Rejection rates (FAR/FRR) are the measures of the probability of a system wrongly accepting an unauthorised person and incorrectly rejecting a legitimate person, respectively. A threshold value is used to determine the best FRR and FAR values. FRR and FAR are directly impacted by changes in the threshold value. FRR and FAR are combined to determine the verification accuracy. EER is the rate at which FAR and FRR are equal (refer Figure 1.2). Details of the parameters used are given below:

False Rejection Rate (FRR): It is the percentage of genuine candidates that are incorrectly rejected by a biometric system. In other words, it is the rate at which a genuine individual is mistakenly identified as an impostor. The FRR is often referred to as the False Non-Match Rate (FNMR). A low FRR score indicates that the system is capable of efficiently capturing intra-class variations via its feature representation approach and matching. Thus, FRR is given by

$$\mathrm{FRR} = \frac{\text{Number of Genuine Persons Rejected}}{\text{Total Number of Genuine Comparisons}} \times\ 100\,\% \tag{1.2}$$

where, Genuine Acceptance Rate (GAR) measures the fraction of the acceptance of genuine candidates and is defined as below.

$$\mathrm{GAR} = (1\ -\ \mathrm{FRR}) \times\ 100\,\% \tag{1.3}$$

False Acceptance Rate (FAR): The rate at which impostors are identified as the legitimate user by a biometric system are referred to as the false acceptance rate (FAR). It is defined as the percentage of candidates that have their biometric information incorrectly accepted by a biometric system. To put it another way, it is the rate at which an impostor is mistakenly regarded as a genuine individual. The low value of FAR

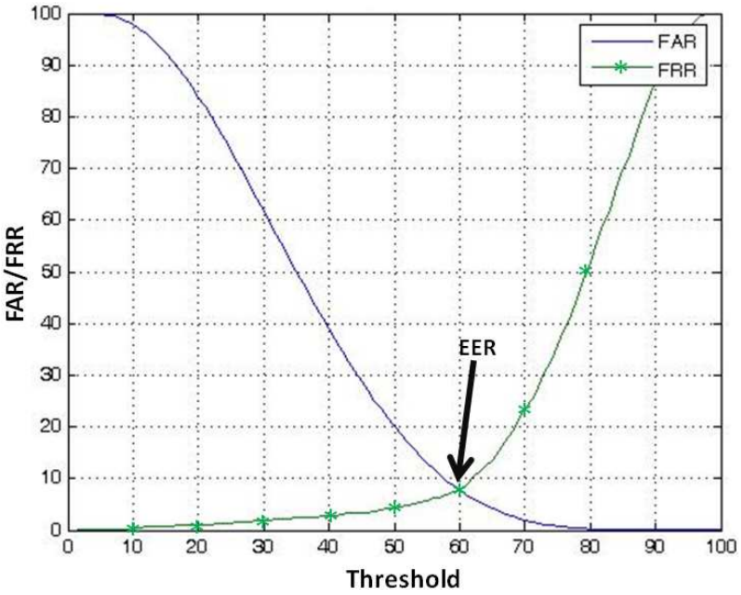


Figure 1.2: An example of Threshold Vs. FAR and FRR curves.

indicates that the biometric system is capable of efficiently capturing inter-class variability through its feature representation and matching techniques. FAR which is also sometime referred as False Match Rate (FMR), is given by a following equation.

$$\text{FAR} = \frac{\text{Number of Imposters Accepted}}{\text{Total Number of Imposter Comparisons}} \times 100\% \qquad (1.4)$$

- **Equal Error Rate (EER):** It is the value of FRR/FAR, when FRR and FAR have the same value, *i.e.*,

$$\text{EER} = \text{FAR} \;\; for \;\; which \;\; \text{FAR} = \text{FRR} \qquad (1.5)$$

- **Receiver Operating Characteristics (ROC) Curve:** ROC curve is a probability curve which shows how well a biometric system is capable of in distinguishing the genuine individuals from the imposters. It is used to evaluate the performance of a verification system. It visually depicts the changes in GAR (Genuine Acceptance

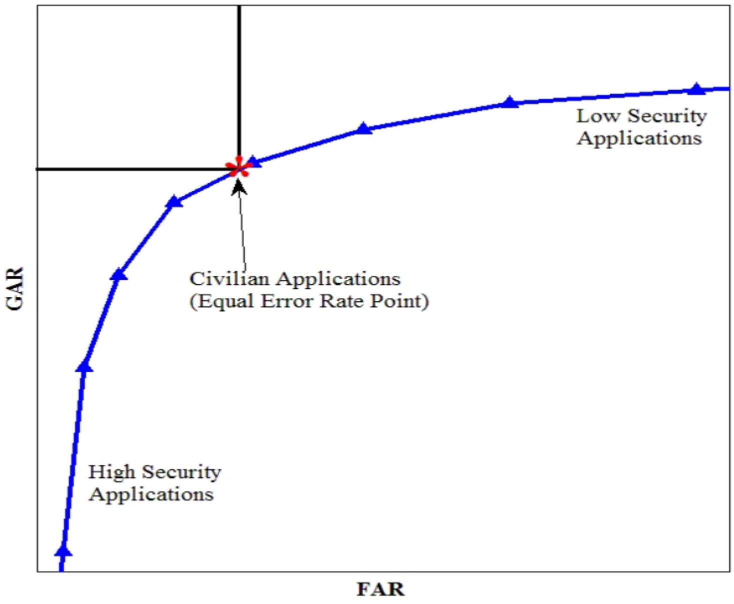


Figure 1.3: Receiver Operating Characteristic (ROC) Curve.

Rate) in relation to the changes in FAR. As the ROC curve depicts the relationship between the FAR and GAR values, it obviates the need of a threshold parameter in the graph. A perfect receiver operating characteristic curve would contain a point at GAR = 100 and FAR = 0. Figure 1.3 shows an example of ROC curve. The ROC curve is an excellent tool for comparing the performance of two biometric systems.

- **Error under ROC Curve (EUC):** The area under the ROC curve (AUC) is a scalar metric that indicates the likelihood that a recognition technique would assign a better score to a randomly chosen genuine match than to a randomly chosen imposter match. The error under the ROC Curve (EUC) is a statistic that is frequently used to provide a better understanding. It is defined as follows.

$$\text{EUC} = (100\ -\ \text{AUC})\% \tag{1.6}$$

- **Rank-*k*:** It is used to analyse the identification performance of a biometric system.

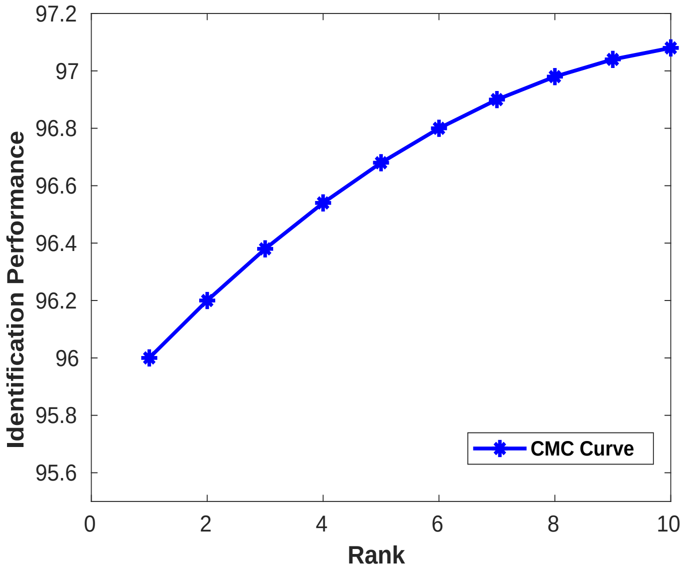


Figure 1.4: Cumulative Matching Characteristic (CMC) Curve.

It shows the proportion of times in which the correct sample occurs with in the top-*k* matches in an identification process. In order to judge the ranking capabilities of an identification system, the cumulative matching characteristic curve (CMC) is used. The accuracy of algorithms that generate an ordered list of probable matches is measured using the CMC curve. For example in facial recognition, the output of the algorithm would be a list of faces from the training-set, ordered from most to least likely to be the test person. So in a CMC curve where the rank-10 accuracy is 50%, it means that the correct match will occur somewhere in the top-10, 50% of the time. Figure 1.4 shows an example of a CMC curve where rank-10 accuracy is shown to be 97.1%. In general, better the algorithm, higher the rank-*k* CMC-percentage assuming that the algorithms have been tested on the same dataset.

1.2.2 Technical Specifications

All experiments presented in the thesis have been carried out on a machine having Intel(R) Xeon(R) Gold 6132 CPUe @ 2.60 GHz, 192 GB RAM and Tesla V100 32 GB graphic processing unit(GPU). Further, Anaconda3.0 environment is being used to perform numerical computations.

1.3 Databases used for Analysis

Rigorous assessments of the proposed techniques have been carried out on three 3D face databases. Details of the databases are provided in this section. The largest publicly accessible 3D face database in terms of number of subjects, the University of Notre Dame (UND) 3D face database (ND-collection D), is discussed in Section 1.3.1 whereas the other two databases are discussed in Subsections 1.3.2 and 1.3.3 respectively. The summary of the 3D face databases used for evaluation purpose has been given in Table 1.1.

1.3.1 UND-collection D Database

The University of Notre Dame (UND)-collection D 3D face database consists of 277 subjects with a total of 953 aligned 3D facial images. All these images present in the database have been captured using Minolta Vivid 900 3D range scanner. The images of the database contain noise in the form of spikes. Due to this, the standard spike removal technique has been used to denoise the 3D images.

1.3.2 BOSPHORUS Database

The Bosphorus database consists of 4666 3D facial scans of 105 individuals. The 3D samples present in the database contains diverse range of expressions, systematic pose

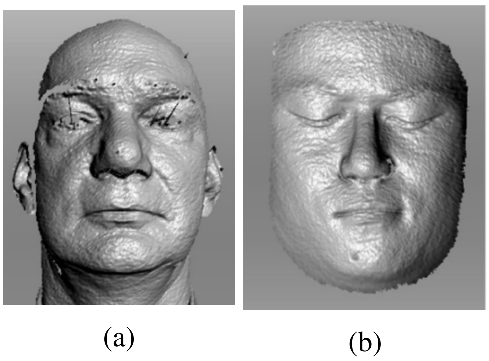


Figure 1.5: A few examples of 3D face samples from UND Database

changes, and occlusions. A total of 299 neutral 3D faces from this database have been used in our experimentation for the analysis. The samples of this database have been acquired using an Inspeck Mega Capturor-II 3D scanner. These scans are aligned and have low noise since noise reduction is performed during data collection by experimental optimization of the acquisition equipment.

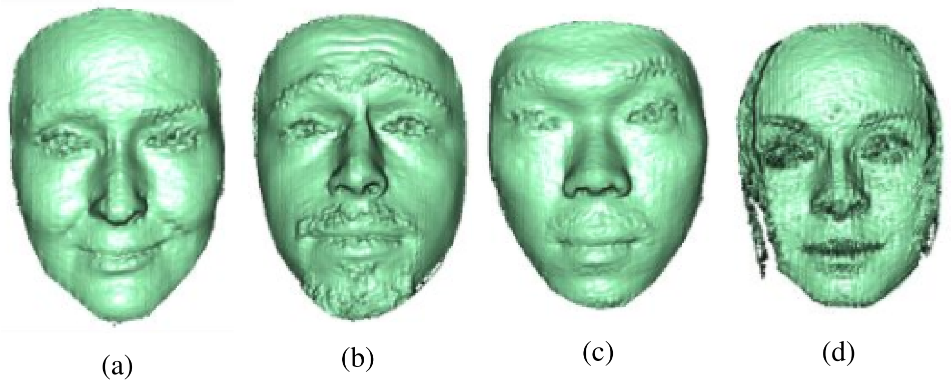


Figure 1.6: A few examples of 3D face samples from Bosphorus Database without texture.

1.3.3 IIT Indore Phase-3 Database

The Indian Institute of Technology Indore (IITI) Phase-3 database contains 3D face images, which have been collected from staff members, teachers and students of the Indian Institute of Technology Indore. All the images in the database are captured in indoor setting. The IITI Phase-3 database contains a total of 445 samples collected from 170 subjects and is our in-house database. In order to scan faces of human subjects, Artec-Eva® 3D scanner is used. The scanner produces high resolution images, is quick, and does not require any extra equipment for scanning. Figure1.8a shows the picture of Artec-Eva® 3D scanner. The scanner has a resolution of 0.5mm in 3D and a 0.10mm accuracy. Figures 1.8b and 1.8c provides the glimpse of the scanning process carried out using Artec-Eva® 3D scanner. As scanner is unable to capture the subject's hair, a wig cap has been used to cover up the subject's hair. The age of the subjects present in the database ranges from 18 and 60 years. Figure 1.7 shows the process of full profile scanning where the scanning process starts from the right side view of the subject and ends at the left side view of the subject. Figure 1.9 shows a few 3D face sample images from this database. In the database, we have 2 full-profile scans and 3-3 profile scans of the subject (left and right). The captured 3D images contain some amount of noise along with holes. Hence the images have been processed for noise removal and hole filling to make them clean.

Table 1.1: Summary of 3D face databases used in the evaluations.

Sr.No.	Database	Data Available	Acquisition Device	# Subjects	# Samples
1	University of Notre Dame,(ND-collection D)(21; 14)	480x640 range images	Minolta Vivid 900 3D scanner	277	953
2	Bosphorus(78)	both 3D co-ordinates and corresponding 2D image coordinates	Inspeck Mega Capturor II 3D scanner	105	4666
3	IIT Indore Phase-3 face dataset	Point Cloud,Mesh	Artec 3D EVA scanner	170	445

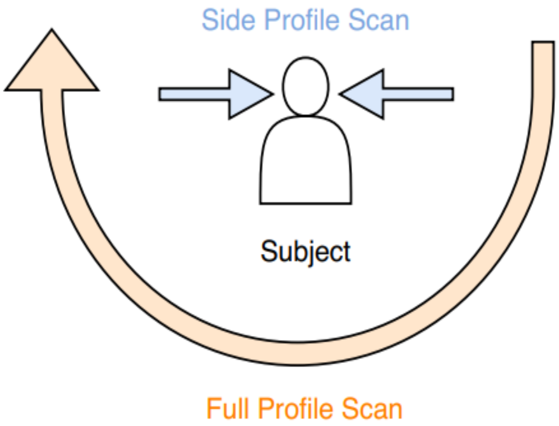


Figure 1.7: Full profile and side profile 3D scanning

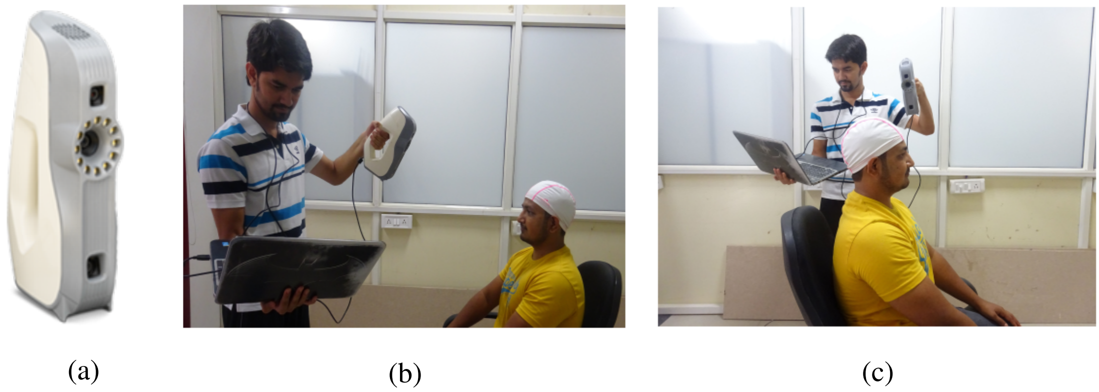


Figure 1.8: 3D face scanning process: (a) Artec-Eva®, 3D scanner used in data collection, (b) full profile scan of a subject, (c) side profile scan of a subject.

1.4 Key Contributions of the Thesis

The work presented in the thesis mainly deals with the face recognition by making use of 3D data. The key contributions of the thesis are highlighted below.

- **Handling Scarcity of 3D Data Through Data Augmentation:** The availability of limited 3D data causes overfitting when used for training in Deep Neural Networks. In this work, we propose three data augmentation techniques for 3D point cloud data that use sub-sampling from the existing point clouds to increase the size and the variability of the available data. For each 3D sample, we create sub-samples using 30% of the

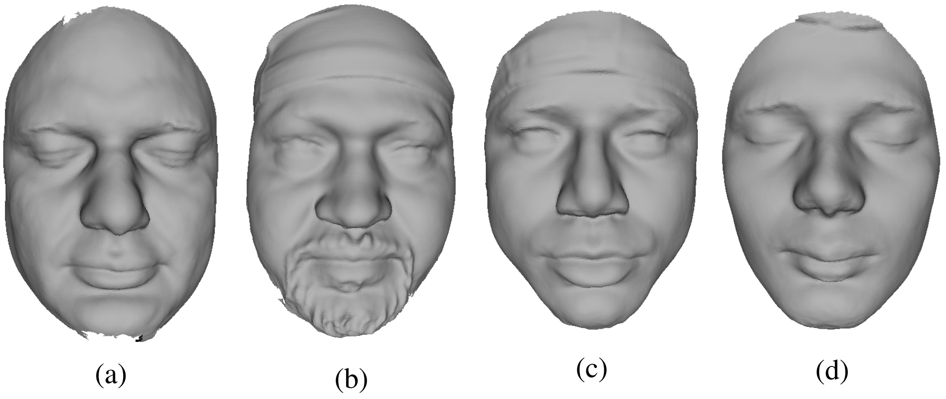


Figure 1.9: Few examples of 3D face samples from IIT Indore database.

points sampled randomly, systematically or in a stratified manner. We further show that the 3D samples created through data augmentation carry the same information by comparing the Iterative Closest Point Registration Error within the sub-samples, between the sub-samples and their parent sample, between the sub-samples with different parent and same subject and finally, between the sub-samples of different subjects. Subsequently, we show that the augmented sub-samples have the same characteristics and features as those of the original 3D point cloud by applying the Central Limit Theorem and comparing the original sample mean and standard deviation values with those of the average of the newly created samples through augmentation.

- **3D Face Recognition:** The performance of 2D object recognition suffers in presence of poor illumination, pose variation, and occlusion. The 3D object recognition is capable of handling these challenges and providing superior performance as it is unaffected by these issues and the 3D representation provides complete geometric features of the object which are rich in information. The available techniques of 3D object recognition performs well while comparing objects which are very distinct; however, does not hold well when comparison is being performed between very similar object classes like in case of biometrics. The PointNet architecture, a popular deep learning framework for 3D object recognition, provides remarkable performance when the object classes are very distinct; however, its performance deteriorates when it is used for the classification of objects which are very similar. In this thesis, a generic technique for 3D object recognition has been proposed that provides solution to this problem. The technique presents a deep neural network model that combines the architecture of PointNet with One-Shot Learning from Siamese Network for 3D object recognition. The model converts a multi-class classification problem of 3D object recognition into a binary class classification problem. the effectiveness of the proposed technique has been demonstrated on 3D face recognition.

- **Face Recognition using Multimodal Data and Transfer Learning:** As stated above, the challenges faced by 2D face recognition due to poor illumination, pose variation and occlusion are easily overcome by the 3D face recognition. However, it is observed that the 3D data has its own challenges. For example, processing of 3D data demands high computational infrastructure which is often not available in resource limited settings. To handle these challenges, we propose the use multimodal data where 2.5D representation (depth image) of 3D data is utilized along with registered 2D images. The use of 2.5D data carries the information inherent in 3D representation up to great extent and its use with the 2D images compensates for the performance. In terms of representation, the 2D and the 2.5D images have similarity; hence, we make use of transfer learning to expedite the training process. We first train the deep neural architecture which is ResNet-34 on 2D images and make use of obtained learning in the training of the network on 2.5D images. The overall recognition is performed by fusing the recognition results obtained for 2D and 2.5D images respectively.

1.5 Organization of the Thesis

The content described in the thesis has been organized in six chapters. The brief description of each chapter of the thesis is presented below.

Chapter 2 presents a detailed review of the techniques available in the literature for data augmentation and 3D face recognition. It discusses the 3D face recognition techniques based on feature key-point detection and description, and the techniques which use deep neural network architectures for 3D face recognition.

Chapter 3 proposes three data augmentation techniques for 3D point cloud data to overcome the problem of availability of limited 3D data. In the proposed data augmentation, new 3D samples are created by using three sampling techniques, *viz.* random, systematic and stratified samplings. To demonstrate that the newly created samples using sampling procedures carry the same information as their parent samples and are capable of differentiating the two objects, registration error between the newly created sub-samples and their parent sample, between the sub-sample and different parent sample and between the sub-samples of different subjects is computed. Further, the Central Limit Theorem is used to show that the information carried by the sub-samples is the same as that carried by their original samples, that is, they have the same discriminative power. Finally, the three sampling techniques are compared based on their results.

Chapter 4 presents a general technique for matching of highly similar 3D objects. The proposed approach combines PointNet architecture with Siamese Network's One-Shot learning capabilities to reduce multi-class classification to a binary classification problem. The proposed approach is capable of matching very similar 3D objects such as 3D human faces, very quickly and accurately. Considering its time and space efficiency in matching the objects, it can be used for creating real-time security solutions for different security applications.

Chapter 5 proposes a technique for robust face recognition using multi-modal data and transfer learning. The proposed technique uses the concept that instead of using 3D data directly, the converted 2.5D data (depth images) from 3D can be used, which are relatively easy to work with in terms of computation power and time requirements. The proposed approach is built on a residual network, ResNet-34, which is first trained on 2D face images. Later the trained model is reused and transfer learning is employed to get a trained model for 2.5D data. The two ResNet-34 models, one trained on 2D data and another on 2.5D data, are used on 2D and 2.5D images respectively to extract features from them. Further, these features are fed to the Siamese Network for verification.

Chapter 6 is the last chapter and it summarizes the work presented in the thesis. It also provides future directions for the research in the field of biometric recognition using 3D face data.

Chapter 2

Literature Review

2.1 Data Augmentation

This section reviews data augmentation techniques that are useful in handling of scarcity of 3D data. Data augmentation deals with increasing the amount of data from the limited set of original data by adding slight modifications in the copies of original data. It helps in reducing the overfitting while training a model by helping in producing enough data required by the model. There are many different strategies, such as padding, cropping, and flipping, which are used for augmenting 3D data in deep learning. These strategies improve performance of the underlying data-driven deep neural network model up to quite extent. Iwasaki et al. (42) have described one such method which involves the use of STereoLithography(STL) data of an object. The algorithm automatically generates a set of training data that covers various backgrounds and a continuous range of view angles. It applies two convolutional neural networks for improving the tolerance of the model against over-classification, increasing the performance over conventional methods. Recent work of data augmentation has also been done in the field of hand pose estimation. Even though, deep learning based methods have significantly improved the performance of hand pose recognition, some of the limitations still remain due to lack of large datasets. Data augmentation strategies used to solve this problem mostly apply image transformation methods such as translation, rotation, scaling, and mirroring (63; 96; 99). In case of color-based methods, training images have been augmented by a process of adjusting the hue-channel for the colored data in (99). Ge et al. in (24) have proposed a 3D transformation for data augmentation in depth-based methods. The transformation involves randomly rotating and stretching of the 3D point cloud for synthesizing the 3D data. Hinterstoisser et al. in (36) have generated augmented data by using training samples that are rendered from 3D models. The method has the limitation of over-fitting as the synthetic data does not have the distribution similar to that of the real data and hence, requires carefully designed training process. Zhang et al. in (106) have presented a method to synthesize image data for augmenting the training process of the neural networks called HandAugment. The method uses a scheme of two-stage neural networks to improve the performance. It also introduces an effective way for synthesizing the data by combining real and synthetic images in the image space together. Jiang et al. (58) have presented five approaches for facial image augmentation, including landmark perturbation and four synthesis methods (hairstyles, glasses, poses, illuminations). The suggested approaches successfully increase the size of the training dataset, which mitigates the effects of dislocation, pose variation, changes in illumination, partial blur, and the negative impacts of overfitting during training. B. Leng et al. (51) presented a strategy in which, before training a model, algorithms generally enlarge the entire dataset, particularly for subjects with inadequate samples. The purpose of this work is to offer a novel data augmentation approach for face classification problems.

3D object recognition is one of the essential applications of biometrics due to its significance in security. Hence, to create strong biometric security systems, robust data augmentation methods are required. In the domain of 3D biometrics recognition, Principal Component Analysis (PCA) in (84) is one of the earliest methods to extract features from a

3D surface to identify two-dimensional and three-dimensional biometric images. Ganapathi et al. (23) have proposed a technique that uses local feature detection and description for 2D and 3D ear biometric recognition. On the collection J2 (UND-J2) dataset acquired at University of Notre Dame, the suggested model achieves 98.69% accuracy. To maximize recognition performance on limited data, a technique with a integrated training scheme for training the classifier based on a right combination of generic and application-oriented data has been developed in (102). Patil et al. in (67) have provided a comprehensive analysis of recently adopted three dimensional face recognition databases, algorithms, features, and problems associated with expression, position, and occlusion changes. Lei et al. in (49) have presented a cost effective three dimensional face identification approach for addressing the problem of partial data such as incorrect data, blur data, or single sample for training. A depth-learning based strategy has been presented in (56) to make 3D object identification algorithms more resilient to external influences such as light, emotions, or attitude. Using the depth information of the 3D face scans, it lessens the impact of external influences. To circumvent the restrictions of the lack of a large number of 3D face scans to train the model, a deep twin neural network has been developed in (98). The suggested technique is accomplished by the use of a convolutional twin neural network that incorporates both 3D depth and 2D texture information from the face samples for 3D face recognition. Face recognition is improved by utilizing cartoons, which enhance distinguishing features of the face. Recently, a three-dimensional automated cartoon-based face recognition approach has been suggested in (62) that generates three-dimensional structures from two-dimensional pictures of face scans and produces promising results. Zulqarnain Gilani et al. in (25) have developed a method for creating massive collections of labelled 3D face scans for training the model as well as a method for merging existing 3D collections for the purpose of testing. Kim et al. in (45) have suggested a Deep CNN and a three-dimensional augmentation approach that combines a variety of various facial gestures from a single three-dimensional face scan.

Robotic manipulation and perception also necessitate the use of 3D object tracking and therefore data augmentation. This is accomplished by creating a synthetic dataset based on the local two-dimensional and three-dimensional features (101). To take full advantage of volumetric information, typically concealed in the depth image, a 3D model which is view-based is built from a single depth image presented in (11). When 3D features are translated to their corresponding representations in voxel form, it is common for some of the information to be lost. In order to address this issue, a novel rotation-invariant feature approach based on mean curvature has been offered in (8) as a solution. This approach significantly improves recognition on voxel CNNs and accuracy rate on the ModelNet10 dataset by 1%. Based on the VGG16 convolutional neural network (CNN), Porcu et al. (68) have assessed and compared the influence of well-known data augmentation approaches on the emotion detection accuracy of a Facial Expression Recognition (FER) system. To expand the number of training images, both geometric changes and Generative Adversarial Networks (GAN) have been used. The best results have been achieved by combining horizontal reflection, translation, and Generative Adversarial Networks(GAN), which led to an accuracy boost of about 30%. Valeska et al. (89) have suggested an approach for face recognition that utilises pre-trained Convolutional Neural Networks (CNNs) with data augmentation as well as transfer learning. The main focus of this research is on the effectiveness of data augmentation for face recognition systems when combined with CNN and transfer learning.

In Chapter 3, we have proposed the use of different sampling techniques for augmenting the data. The proposal is highly useful when it is not possible to collect enough data required for training of a deep neural network model. Taherdoost et al. in (85) have presented various types of sampling techniques and suggested the differences to select the proper sampling method for the research. We have proposed three sampling techniques to be used in augmenting 3D point cloud data, namely, random sampling, systematic sampling, and stratified sampling. Further, we make use of Iterative Closest Point (ICP) algorithm (15; 69; 91) and Central Limit Theorem (CLT) (35) to prove that the sub-samples created from our original samples for the purpose of data augmentation all carry the same information and that they have the same discriminative power as possessed by the original sample.

2.2 Conventional 3D Face Recognition

This section reviews generic techniques for 3D face recognition. Despite the fact that 2D face recognition has had a lot of success, changes in pose and lighting conditions still have a significant impact on accuracy (1; 108). Many researchers have shifted to 3D face recognition because it has the potential to cope up the fundamental restrictions and shortcomings of 2D face recognition. Furthermore, in the case of face recognition, when the pose and lighting conditions do not change, 3D face data is more accurate than 2D data due to the geometric information present in 3D face images (7; 97). Curvature-based algorithms have been tested on a Small 3D face database in the late 1980s by W. Yijun et al.(93), and they achieved 100% identification accuracy. In 1996, Gaile G. Gordon (27) has performed a face recognition experiment that combines frontal and side views of face in order to enhance the face recognition accuracy. Following that, progressive research in the field of 3D face recognition has been done because of the availability of new 3D scanning equipment that uses lasers and structured lights. The 3D deformation model (3DMM) synthesis technique has been established by Blanz and Vetter (6) in 1999, and this model has been utilized for 3D face recognition. Their 3D deformation model is being recreated from 2D photos due to the technological limits of 3D scanning types of equipment at the time. The reconstruction of the 3D model necessitates a significant amount of computing. Many researchers agree that

3DMM are useful for face recognition, even so the computational difficulty of the rebuilding process restricts their use (2; 38; 48). Wu et al. (64) have proposed 3D face recognition using facial range data to extract multiple horizontal profiles. One disadvantage of this approach is that the recognition accuracy drops dramatically as the head posture changes. Zhang et al. (103) have investigated techniques and algorithms for 3D face identification under posture changes, as well as the greatest angle that can be recognized while the posture varies. In 3D facial recognition, Chua et al. (16) have employed point signatures. Only the portions of the face that are rigid (below the eyebrow and above the nose) are used to handle with variations in facial expressions. In the experiment, images of six different subjects with distinguished expressions are employed, and 100% accuracy has been achieved. Hesher et al. (34) have tested the principle component analysis (PCA) approach, which employs a variety of feature vectors and image sizes. The image dataset consists of 37 subjects, each with six different face expressions. The recognition accuracy improves when multiple images are used in the gallery. Moreno et al. (61) have presented a technique where a feature vector on the segmented region of three-dimensional facial data has been constructed using Gaussian curvature method. The accuracy obtained by this technique is 78% on a sample of 420 faces from 60 subjects with diverse facial expressions. The face model has been partitioned by Martinez et al. (59) into small sections, and a probabilistic technique has been devised to locally match each area, and the matched results are integrated for facial recognition. Osada et al. (72) have suggested a method for determining the similarities between three-dimensional objects by calculating shape signatures for three-dimensional polygonal models. The suggested technique encodes the object signature as a sampled shape distribution derived from a shape function that quantifies the object's global geometric features. The technique is robust with respect to geometric changes such as rotations and translations and can be utilised as a pre-classifier in three dimensional object recognition systems.

Earlier, 3D object recognition has utilized the techniques like Iterative Closest Point algorithm, the differential geometry technique (26) and the techniques to calculate curved surfaces using spherical correlation (86). Prior to 2004, there have been a few freely available 3D face databases. Song et al. (82) have devised a three-dimensional face recognition algorithm that could withstand significant head displacement. To correct the head pose in the scanned image, the approach uses geometric information from feature points on the face. In 2006, Samir et al. (76) have presented a technique to compare facial shapes based on the curvature of the surface. The basic idea is to approximate the surface of a face with a limited level curve, and the curve is taken from the depth image. Using the combination of linear support vector machine (LSVM) and linear discriminant analysis (LDA), in 2007, Kin-Chung et al. (92) have suggested a three-dimensional face recognition system. By collecting local features from several regions, this approach obtains the sum of invariants. From the frontal face image, ten sub-regions and resultant feature vectors are retrieved. An additional approach for retrieving comparable shapes from an extensive 3D object database has been presented, which is called priority-driven search (PDS) (22). The objects are described in terms of three-dimensional feature sets. The outcome of the search produces a list of target objects with a rank that indicates how precisely any subset with *k* features matches for the probe item and the target item. Many research institutes have set up various types of three-dimensional face databases in recent years to to test and assess their in-house three-dimensional face recognition systems. Various 3D face recognition algorithms perform differently on different three-dimensional face databases. Many approaches are implemented on a particular 3D face dataset, and their achievement on other datasets may be different. Huang et al. (39) have proposed a multi-scale local binary model (MS-LBP) depth map as a novel 3D surface representation approach. This approach is used with combination of Shape Index (SI) map and Scale Invariant Feature Transform (SIFT). Using the approach, on the Face Recognition Grand Challenge (FRGC v2.0) database, the Rank-1 recognition rate has been achieved as 96.1%. This approach has been demonstrated to work with partly occluded facial probes. On the Bosphorus database, Li. Huibin et al. (52) have proposed a mesh-based three-dimensional face recognition technique using a novice local shape descriptor and a SIFT-like matching process. D. Smeets et al. (80) have developed the meshSIFT algorithm and its application for 3D face recognition. The technique extracts information on different scales from 3D surfaces, resulting in expression-stable 3D face recognition that has been verified against the FRGC and Bosphorus databases.

In order to obtain 3D geometric information, S. Soltanpour et al. (81) have used SIFT keypoint detection on pyramidal shape maps and combine it with 2D keypoints. In this work, the Face Recognition Grand Challenge (FRGC v2) database and Bosphorus database is used for experimentation. On FRGC v2, verification rate is obtained as 99% for all versus all case and on Bosphorus, it is 95.8% for neutral versus all case. Disadvantage of this SIFT method is that it is sensitive to changes in pose. To address the challenges like missing parts, occlusions and data corruptions, Y Lie. et al. (50) have presented a competent 3D face recognition method. In this method, significant facial expressions and pose variations are represented by a facial scan with a set of local Keypoint-based Multiple Triangle Statistics (KMTS), which is robust to partial facial data. They also proposed Two-Phase Weighted Collaborative Representation Classification (TPWCRC) scheme is used to carry out face recognition and its performance is analyzed on six databases, *viz.*, Bosphorus dataset, GavabDB, UMB-DB, SHREC 2008, BU-3DFE, and FRGC v2.0 datasets. In 3D Facial Expressions Recognition(FER), W. Hariri et al. (30) have explored the usage of co-variance matrices of descriptors, instead of the descriptors themselves. The BU-3DFE and the Bosphorus datasets are used to test the performance, and the results are compared to the best available techniques. X. Deng. et al. (18) have suggested a new 3D face recognition approach based on the local co-variance descriptor and Riemannian kernel sparse coding to precisely assess the inherent correlation of extracted features. FRGC v2.0 and Bosphorus datasets are used for experiments and the suggested method outperforms the recognition accuracy, and results are compared to current state-of-the-art techniques. Yi. You et al. (100) have proposed a rigid registration approach based on surface resampling and denoising, that reduces the influence of sampling difference and noise on registration residuals. Bosphorus and FRGC v2.0 databases have been used for the experiment, and proposed algorithm outperforms other benchmark algorithms. L. Shi et. al. (79) have proposed a 3D face recognition approach integrating Local Binary Pattern (LBP) and Support vector Machine (SVM) to increase the accuracy and the speed of 3D face identification. The trait information of the three-dimensional facial depth image is extracted using the Local Binary Pattern (LBP) technique, and then the feature information is classified using the Support vector Machine (SVM) algorithm. The experiment shows that the algorithm has a higher recognition rate and consumes less time by picking samples from the Texas 3DFRD three-dimensional face depth database and the self-made 3D face depth library.

2.3 3D Face Recognition using Deep Learning

Deep Learning is a branch of machine learning that relies on methods and techniques influenced by the brain's structure and function. An artificial neural network (ANN) comprises of a layer of nodes or neurons with several layers. The first layer is an input layer; the second layer is the hidden layer. In a network, there may be one or more hidden layers, and finally, an output layer. The neurons are interconnected, and each neuron has a weight and threshold value linked with it. A neuron is activated if its output value exceeds the predefined threshold value. In this situation, the neuron starts forwarding the data to the next layer of the network else data is not sent to the next layer. Using several layers of representation and abstraction, deep learning-based techniques extract high-level features from the raw data. In image identification and processing, a convolutional neural network (CNN or ConvNet) is a sort of artificial neural network that is specially intended to interpret pixel input. CNNs are used to recognize and process images. They are powerful image processing and artificial intelligence (AI) systems that use deep learning to perform both conceptual and informative tasks. CNNs are often used in conjunction with machine vision, which includes image and video recognition, medical image prediction systems, and natural language processing (NLP), among other things. The amount of pre-processing required by a ConvNet is much less than that of other classification techniques. While primitive approaches rely on hand-engineered filters/characteristics, ConvNets are capable of learning these characteristics with sufficient training. They have been particularly beneficial in (28) digit identification, Electroencephalogram (EEG) recognition (54), and general object recognition, among other applications. Hyper-spectral picture categorization is accomplished by the use of a multi-scale 3D deep convolutional neural network, as suggested in (33). A slice-based CNN strategy to classify three-dimensional objects in real-time has been developed by Gomez-Donoso et al. (19), and it has been demonstrated to have an accuracy of 94.34% in the ModelNet10 dataset. Asif et al. (3) have presented a methodology for object classification that is discriminative and structurally invariant to handle the difficulties of inter-class similarities, intra-class changes, as well as spatial variability in images.

Sales et al. (75) have suggested a unique 3D shape descriptor for detecting objects in 3D instances that may be used as input for the learning model. Another way to recognize 3D objects is to use a system that has an improved depth estimation algorithm (20). This algorithm uses statistical calculations to improve the depth image and lessen the effect of noise. Using spherical CNNs, a new approach for achieving rotation-invariance in 3D object recognition has been described in (104). For detecting rotated 3D objects, the study offers a novel rotation-invariant deep network that employs the rotation-equivariant spherical correlation notion. Volumetric CNNs are used in the majority of deep learning-based 3D data techniques. One form of input to 3D CNNs is voxelized shapes (60; 71; 95). In contrast, sparse data spaces limit these representations because convolution procedures are computationally costly. Capturing distinct face features necessitates a high voxel resolution, requiring a large quantity of memory. Compared to other representations, features in a point cloud are represented by the set of 3D points and are invariant to certain internal (4; 10) and external (74) modifications. The features may be local or global, but these features must be appropriately integrated optimally to create the best model. Vectorized 3D data in the form of vectors is utilized by feature-based DNNs (29; 41), which extract original features from shapes and identify those shapes using a fully inter-connected network to extract and classify such shapes. It has also been stated that the embedding patch method (105) in CNNs might help to enhance face representation. To solve 3D non-rigid shape retrieval difficulties in a large dataset, Amores et al. (9) have suggested a feature-based solution based on text search algorithms that employ a ”bag of features.” Using multi-scale diffusion heat kernels, the approach is able to create relevant and concise shape descriptors and the results obtained on a large-scale shape retrieval benchmark.

One of the most important uses of 3D object recognition is in the field of biometrics. 3D face and ear structures may be utilized to construct robust biometric security systems because of their unique anatomical identifiers. Principal Component Analysis was used in one of the earliest ways to extract characteristics from the 3D surface for 2D/3D face or ear recognition (84). One of the challenges in the processing of 3D faces is the extraction of correct facial landmarks from the 3D model. Terada et al. (88) use scanned 3D images to investigate facial shapes and look at the different methods to find facial landmarks. A complete summary of recently utilised 3D face identification algorithms, datasets, features, and the issues coupled with changes in expressions, positions, and occlusions is provided by Patil et al. (67). Y Lei et al. (49) provides an effective 3D face recognition system that addresses the issue of incomplete data, such as damaged data, occluded data, and a single sample data for training, among other issues. It is recommended in (56) that a depth-learning-based technique be used to make the 3D face recognition algorithm more resistant to external influences such as light, expression, and posture. This strategy also decreases the influence of extrinsic factors by employing the depth information of the 3D face images. A deep twin neural network is presented in (98) to address the constraint of a significant number of 3D face data being unavailable for training the model. Face identification is accomplished using a convolutional twin neural network that mixes the 3D depth and 2D texture of the faces. Caricatures, which exaggerate distinguishing aspects of the face, are used to improve the identification of faces in a more natural way. An automated 3D caricature-based face recognition system has been suggested in (62) that derives 3D shape structures from 2D images of the faces and produces a promising result.

While the majority of 3D face recognition methods that employ point clouds attempt to resolve expression variations, only a handful have been effective in resolving obstacles posed by pose variations and occlusions. Li et al. (53) have adapted the SIFT-like mechanism to mesh data and have provided a method for comparing 3D keypoint descriptors with fine-grained matching. Hu et al. (37) have suggested a complete parametric investigation of two CNN facial recognition models. The models differed in terms of combination of their hyper-parameters like activation functions, learning rates, and the size of filter. Despite recent advances in deep learning, the saturation of 3D datasets due to their restricted gallery size has hampered advancement in the field of 3D biometrics. Zulqarnain Gilani et al. (25)have provided a technique for producing a huge corpus of annotated 3D face images for training, including a solution for combining various 3D datasets for testing. Kim et al.

(45) have suggested a Deep CNN and a new 3D augmentation approach that reconstructs several facial expressions from a single 3D facial scan.

In addition, deep Siamese Neural Networks (94) have been employed for facial recognition. Joshi et al. (44) have also employed the Siamese Network to compare scanned and digitised facial images. Hayale et al. (31) have described one such strategy using a supervised loss function that maximizes the distance between features for distinct classes while decreasing intra-class variations, increasing inter-class variations.

The exploration of transfer learning is motivated by the notion that people can intelligently reuse previously learned information to solve new challenges more quickly and efficiently. In a session on ”Learning to Learn” presented at NIPS-95 (Neural Information Processing Systems), the core motives for transfer learning centered on the need of lifelong machine learning algorithms that store and reuse previously learned information discussed in (65). Small data and personalisation should be the emphasis of future machine learning research. It is possible to use a pre-trained model trained on an extensive reference dataset to solve a problem comparable to the similar set of domains we want to solve. Luttrell et al. (57) combine a pre-trained facial recognition model with transfer learning to create a network that can accurately predict on a considerably smaller dataset. In template adaption, VGG system is used for transfer learning. In this approach, the deep CNN features from pre-trained VGGNet are combined with template specific linear SVMs, outperforms the state-of-the-art by a wide margin (17). Hang Zhao et al. (107) have presented an instance-based transfer learning approach, which is a weighted ensemble transfer learning methodology with multiple feature representations. R. Sandip Kute et al. (46) have introduced a unique technique for component-based face recognition and association via transfer learning, demonstrating that knowledge gained from entire face images is used to classify face components. Cengil et al. (13) have developed a multiple classification model of flower images and have achieved a good performance with VGG16 model as a pre-trained network. Deep learning has improved recognition rate, but small sample sizes creates new problem. In 3D face recognition, an expression-invariant method has been proposed by Zhenye Li et al.(55). The main objective of the proposed method is to use transfer learning with Siamese Networks to overcome the issue of limited sample size. G. Vishnuvardhan et al. (90) have developed a faster and better way to train a facial recognition model used in banking operations. The method uses transfer learning to extract facial embeddings and Nearest Neighbors (NN) to identify the face without using big datasets or graphics processing unit (GPU) calculations to train the model. Using the Georgia Tech Face-Database (GTFD), 96.67% accuracy has been obtained in this technique that is near to human vision which has accuracy of 97.53%.

Chapter 3

Handling Scarcity of 3D Data

Nowadays, many of the object recognition techniques use 3D data instead of 2D. This is due to the fact that the object recognition performance on 3D data is significantly better than that on the 2D data. For example, in the case of face recognition, 2D face recognition is hindered by pose, expression, and illumination variations. These limitations are easily overcome when using 3D data as all the information about the face geometry is possessed by 3D data. Similar applications of 3D data can be seen in various other fields including robotic perception and manipulation.

Given the significance and vast applications of 3D data in areas like object recognition and biometrics, it becomes important to address the issues faced during the training of the deep neural network model. Although the 3D object recognition has achieved great accuracy, 3D data acquisition from objects takes time and hence often, there is a very limited data available for 3D objects. In the availability of limited data, the model learns the details and noise of these few samples so well that it negatively impacts the testing of the selected model on new data. To avoid this problem of overfitting, we need to increase the variability

The work presented in this chapter appears in the paper "Handling Data Scarcity through Data Augmentation in Training of Deep Neural Networks for 3D Data Processing", International Journal on Semantic Web and Information Systems (IJSWIS), Volume 18, Issue 1, Article 14, pages: 1-16 (2022). `https://doi.org/10.4018/IJSWIS.297038` of the 3D data by enlarging the size of the database by making use of data augmentation.

There are different ways to represent and input 3D data to a model. Some common and popular ways of representing an object in 3D include 3D voxel and point cloud. The 3D voxel representation is a highly regularized form of representation. In this representation, a 3D object is represented by discretizing its volume where the unit cubic volume is called a voxel. This representation has an advantage as it simplifies weight sharing and other kernel optimizations. However, it is bulky in nature with sparse data spaces and involves convolution operations that renders this representation computationally and spatially expensive. Further, capturing fine structures require a very high voxel resolution, consuming a massive amount of memory. On the other hand, point clouds are the most raw form of 3D data and are the direct outcome of the object scanning process. In point clouds, a 3D object is represented by digitizing its surface in the form of an unordered set of data points which can be directly consumed as inputs to any deep neural network instead of transforming them into regular 3D representations such as 3D voxels.

As stated above, the 3D input data for an object which is in the form of a point cloud, contains an unordered set of 3D points. It is seen that this original set of points for an object contains a huge number of 3D points; however, due to the computational and memory limitations of the system, often, we cannot use the entire point cloud of a single sample for processing. To mitigate this problem, usually, the original point cloud data is sub-sampled, and a reduced size cloud is used for processing. However, in this process, the number of samples for a subject remains the same as was available earlier before sampling. We exploit the use of sampling in a different way and propose its use in data augmentation by increasing the number of samples of the subjects.

This chapter presents three techniques for augmentation of 3D data which is originally available in point cloud format. The techniques use sub-sampling from the existing point clouds to increase the size and variability of the available data. We use the Iterative Closest Point (ICP) (15; 69; 91) algorithm to show that the samples created from the original data all carry the same information. We further use Central Limit Theorem (CLT) (35) to prove that the information carried by the sub-samples is the same as that carried by the original sample, *i.e.*, they have the same discriminative power. Finally, we compare techniques based on analytical results.

3.1 Proposed Technique

The proposed technique use three different types of sampling approaches for augmenting the 3D data. This generates different subsets or sub-samples from the original samples. As the original samples are in the form of 3D point clouds, number of points in the point cloud will be considerably less in the sub-samples, which will make the training of deep neural network model computationally and spatially efficient. We also prove, using Iterative Closest Point (ICP) and Central Limit Theorem (CLT), that no information is lost while sub-sampling the point clouds. Given a 3D point cloud, we use the following types of sampling approaches for data augmentation.

- **Random Sampling:** In this sampling, each member of the set has an equal unbiased opportunity of being chosen as a part of the sampling process. In random sampling, to create multiple sub-samples from a single sample, we randomly select a fixed proportion of points from the original 3D point cloud multiple times. This creates different unordered subsets containing a uniform number of 3D points. We are selecting one-third of the original number of points from each sample point cloud to carry out the sampling process to generate the sub-samples.

- **Systematic Sampling:** Systematic sampling is a probability based sampling technique where sample members from a population are selected from a random starting point but with a periodic and fixed sampling interval. This technique eliminates the chances of clustered selection. In this technique, we sort the point cloud of a sample by ordering the points in 6 possible arrangements - (x, y, z), (x, z, y), (y, x, z), (y, z, x), (z, x, y), (z, y, x). For each arrangement, we choose a random starting point in [0, k-1] and choose the subsequent points after skipping k, $2k$, $3k$... points where k lies in the range [3,5] depending on how crowded or sparse we want our sub-samples to be. Lower k results in a lower variance of points among different sub-samples while higher k results in less repetition but sparser point clouds. However, we need to ensure that the chosen k is not symmetric about the point cloud as this will result in the same sub-sampled point cloud irrespective of the ordering arrangement. We are making use of $k = 3$ so that the sub-samples use one-third of the point cloud.

- **Stratified Sampling:** Stratified sampling divides the total population into smaller groups for carrying out the sampling. These groups are formed based on some common properties existing in the population. After dividing the population into groups, random selection of the samples is performed proportionally. In this technique, we divide the entire point cloud of a object sample into cubical windows of fixed size and then select a proportionate number of points randomly from each window to create a single sub-sample. Hence, a higher number of points are selected from a dense region whereas a lower number of points are chosen from a sparse region thus maintaining localization. We are making use of a window of size $5 \times 5 \times 5$, and select one-third of the total number of points from each window to carry out the sampling. Sampling is performed multiple times to generate multiple sub-samples from the point cloud of an object.

We use ICP algorithm and Central Limit Theorem to prove that the sub-samples created from the original samples (the original 3D point cloud of the object) all carry the same information and that they have the same discriminative power as possessed by the original sample.

The ICP algorithm finds a transformation matrix between two point clouds by minimizing the square errors between them. One of the point clouds (target) is fixed, and the other one (source) is transformed to best match the target. The algorithm is iterative and improves the transformation matrix to minimize the error. Finally, it returns the final error after the transformation along with the transformation matrix. The error is essentially the registration error between the two point clouds which indicates how dissimilar the information carried by two point clouds is. For very similar point clouds, the registration error is very close to zero. The registration error can be used to find the similarity between the generated 3D sub-samples in the following ways for all the proposed three augmentation techniques.

- **Intra-sample Registration Error:** For a given sample, we find the registration error between the sub-samples created from that sample, as well as between each of the sub-sample and the original sample. Since, all the created sub-samples carry the same information, the error in first case should be very close to zero while in second case, it should be similar for all the sub-samples.

- **Inter-sample Registration Error:** For a given subject, we find the registration errors between the sub-samples created from the two different samples. For example, Subject 1 has two available samples - Sample 1 and Sample 2. We create 3 sub-samples each out of Sample 1 and Sample 2. Now, we find the registration error between the sub-samples of Sample 1 with each of the sub-samples of Sample 2. These should yield similar values for each combination which should be close to the original registration error between Sample 1 and Sample 2, verifying that the sub-samples carry the same features as well as they inherit the features of their parent sample.

- **Inter-subject Registration Error:** For two different subjects, we find the registration errors between the sub-samples created from a sample of each subject. This method is similar to the previous one except that we are using sub-samples of different subjects instead of sub-samples of different samples from the same subject.

The CLT (35) states that for any kind of data with a high number of samples, mean and standard deviation of the sampling distribution should be equal to the mean and standard deviation of the population divided by the square root of the total number of samples or the sampling size. We calculate mean and standard deviation for population as well as sampling distribution to verify the CLT. Using this, we prove that the discriminative power of samples, in our case the sub-samples created from the original point cloud of the object, is same as that of the population, *i.e.* the original point cloud of the object.

Therefore, using ICP and CLT, we show that no information is lost while sampling the data, and hence, sub-samples are effective to be used in training the deep neural network model. The proposed work attempts to increase the size of the limited 3D input data by using sampling techniques. The suggested augmentation techniques can be used on any class of 3D point clouds such as general objects, biometric modalities (faces, ears, etc.) since the sampling techniques are independent of the object classes to be recognized. The most obvious advantage of such data augmentation technique is that it overcomes the problem of overfitting due to limited training samples per subject. Moreover, due to multiple sampling from the same point cloud, not only do we increase the number of samples per subject but we also reduce the time and space complexity while training the model since the number of points in each sample is reduced to a third of the original sample. This also ensures that features of the original data are retained as evident from the ICP registration errors of the sub-samples and their results on CLT as discussed in the next section.

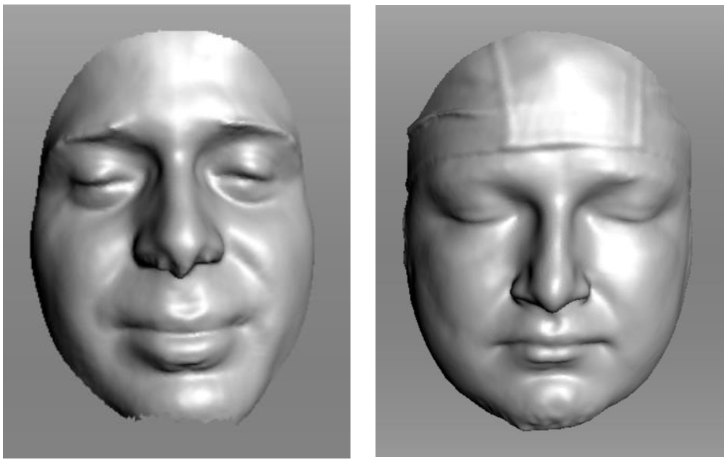


Figure 3.1: 3D face samples from IITI 3D database used in experimental evaluation.

3.2 Experimental Results

We use our in-house database, the IIT Indore Phase-3 (IITI) database, for the experiments. A few sample images from this database are shown in Figure 4.6. The IITI database contains 170 subjects, with a total of 445 samples where Artec EVA 3D scanner has been used to acquire the 3D facial scans. The database contains challenging samples where many of them are noisy and are not aligned properly. We augment these 3D facial samples using the proposed three sampling techniques and compare the results.

3.2.1 Intra-sample Registration Error

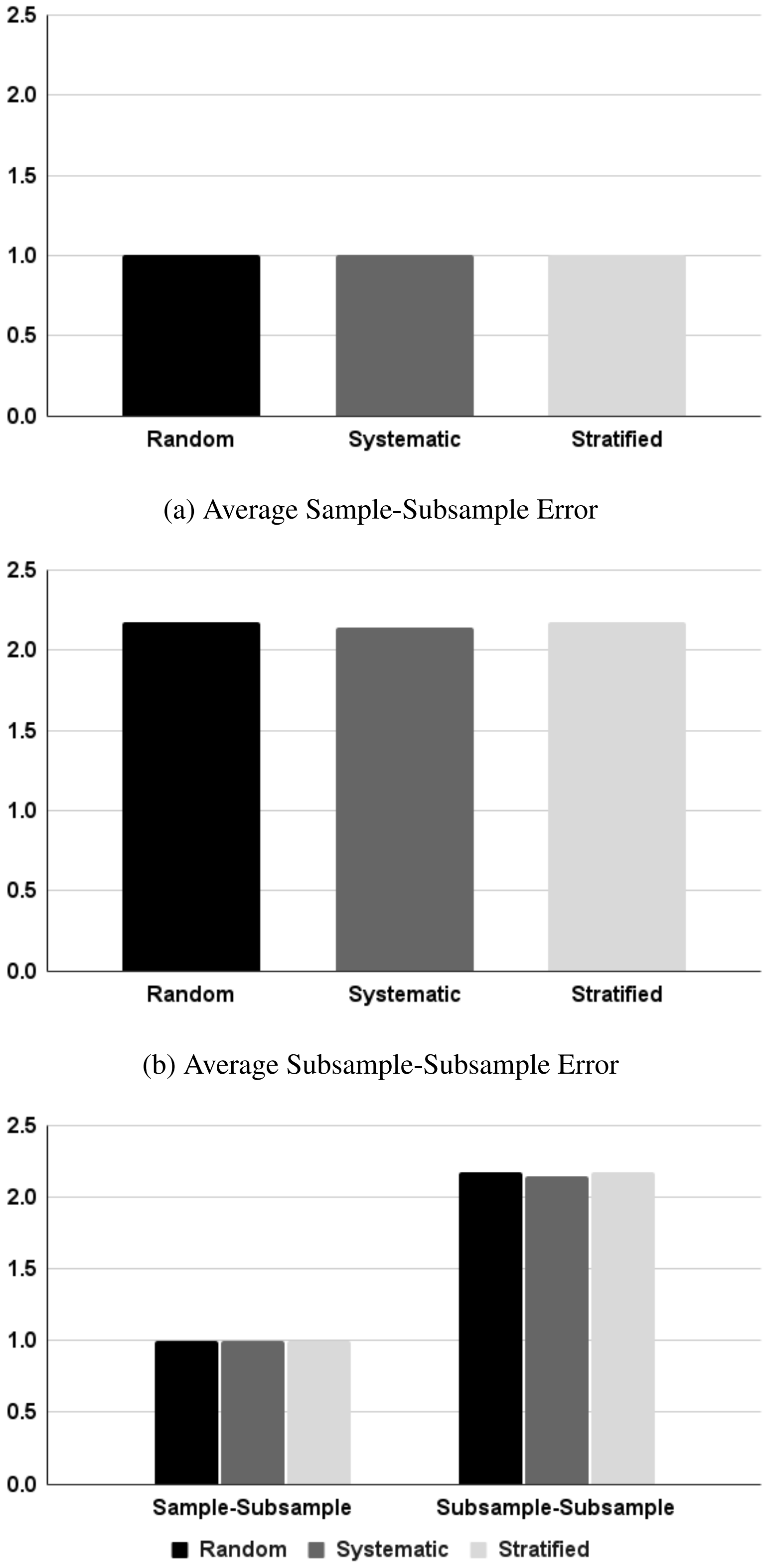
We use three samples of each subject and create a set of three sub-samples for each of them. Table 3.1 shows the ICP registration error between a given sample and its respective sub-samples for all three samples for a subject. Similar experiment is repeated for all the subjects. Table 3.3 shows the average of means of Sample - Sub-sample error for all the samples (*i.e.*, all samples of 170 subjects). Further, Table 3.2 shows the ICP registration error between all three pairs of sub-samples of each of the three samples for a subject. Similar experiment is repeated for all subjects and results are reported in Table 3.4 where it shows the average of means of Sub-sample - Sub-sample error for all the samples (*i.e.*, all samples of 170 subjects).

From Table 3.1, we can see that the registration error between the original sample and its sub-samples is very similar for all the sub-samples while from Table 3.2, we see that the registration errors are very close to zero, verifying that all the sub-samples of a particular sample carry the same information. From Figure 3.2(a) (values are plotted in exponential scale for clarity) which is the graphical representation of Table 3.3, it is evident that the sample - sub-sample similarity is relatively highest (that is, registration error is the least) in the case of stratified sampling which can be explained because of the use of localization in selecting points. Systematic sampling has the next best similarity owing to ordering of the points before selection. Random sampling has the highest error because of the absence of any ordering or localization. Further, from Figure 3.2(b) (values are plotted in exponential scale for clarity) which is the graphical representation of Table 3.4, we can see that the sub-sample similarity is highest (that is, registration error is the least) in the systematic sampling because in this sampling technique, there is a possibility of repetition as we might choose the same set of points which is not desirable. For an effective sampling, we need low sub-sample similarity for more variation in the data after the augmentation. We see that this desirable characteristic is achievable in case of stratified or random sampling. Figure 3.2(c) provides a combined comparison of Figures 3.2(a) and 3.2(b).

.

3.2.2 Inter-sample Registration Error

In this experiment, we are taking two different samples of the same subject, say Sample 1 and Sample 2. We also take the respective sub-samples, namely Sub-samples 11, 12, and 13 from Sample 1 and Sub-samples 21, 22, and 23 from Sample 2. Table 3.5 shows the



(a) Average Sample-Subsample Error

(b) Average Subsample-Subsample Error

(c) Intra-Sample Graphs Comparison

Figure 3.2: Intra-sample Registration Error for Random, Systematic and Stratified Samplings (values are plotted in exponential scale for clarity).

Table 3.1: Sample - Sub-sample Registration Error (demonstration for one Subject) for Random, Systematic and Stratified Samplings.

Sampling technique	Samples	Sub-sample 1	Sub-sample 2	Sub-sample 3	Mean
Random	Sample 1	4.18E-08	3.49E-08	3.66E-08	3.78E-0
	Sample 2	0.00E+00	0.00E+00	8.05E-09	2.68E-09
	Sample 3	1.83E-08	2.72E-08	2.17E-08	2.24E-08
Systematic	Sample 1	3.73E-08	3.73E-08	3.32E-08	3.59E-08
	Sample 2	1.13E-08	1.80E-08	0.00E+00	9.77E-09
	Sample 3	1.83E-08	1.41E-08	1.83E-08	1.69E-08
Stratified	Sample 1	0.00E+00	0.00E+00	0.00E+00	0.00E+00
	Sample 2	0.00E+00	0.00E+00	0.00E+00	0.00E+00
	Sample 3	0.00E+00	0.00E+00	0.00E+00	0.00E+00

Table 3.2: Sub-sample - Sub-sample Registration Error (demonstration for one subject) for Random, Systematic and Stratified Samplings.

Sampling technique	Samples	Sub-sample 1-2	Sub-sample 2-3	Sub-sample 3-1	Mean
Random	Sample 1	7.89E-01	7.89E-01	7.82E-01	7.86E-01
	Sample 2	7.82E-01	7.95E-01	7.85E-01	7.87E-01
	Sample 3	7.87E-01	7.92E-01	7.85E-01	7.88E-01
Systematic	Sample 1	7.47E-01	7.35E-01	8.17E-01	7.66E-01
	Sample 2	7.50E-01	7.38E-01	7.34E-01	7.74E-01
	Sample 3	7.43E-01	7.34E-01	8.28E-01	7.68E-01
Stratified	Sample 1	7.87E-01	7.89E-01	7.85E-01	7.87E-01
	Sample 2	7.84E-01	7.84E-01	7.84E-01	7.84E-01
	Sample 3	7.82E-01	7.80E-01	7.83E-01	7.82E-01

ICP registration error between Sample 1 along with its respective sub-samples and Sample 2 along with its respective sub-samples. We consider the value of ICP registration error between Sample 1 and Sample 2 as the original value (mean) and find RMS (root mean square) error for each row and column. Similar experiment is repeated for all the subjects and results are reported in Table 3.7 where it represents the average of RMS (root mean square) error of Inter-Sample registration over all subjects. From the above experiment, we can see that the registration error obtained between the sub-samples is very similar to that obtained for the original samples. From Figure 3.3 (values are plotted in exponential

Table 3.3: Average of Mean of Sample - Sub-sample Registration Error over 170 Subjects for Random, Systematic and Stratified Samplings.

Technique	Average Mean Sample - Sub-sample Error
Random	2.21E-08
Systematic	2.20E-08
Stratified	0.70E-08

Table 3.4: Average of Mean Sub-sample - Sub-sample Registration Error over 170 subjects for Random, Systematic and Stratified Samplings.

Sampling technique	Average Mean Sub-sample - Sub-sample Error
Random	0.7769
Systematic	0.7619
Stratified	0.7762

Table 3.5: Inter-Sample Registration Error (demonstration for a pair of samples) for Random, Systematic and Stratified Samplings.

Sampling technique	Samples	Sample 2	Sub-sample	Sub-sample	Sub-sample	RMS
Random	Sample 1	8.4114	8.3418	8.4907	8.44	0.0546
	Sub-sample 11	8.4472	8.3779	8.5268	8.4753	0.0704
	Sub-sample 12	8.4466	8.3765	8.5258	8.4756	0.0701
	Sub-sample 13	8.4452	8.3761	8.525	8.4735	0.0692
	RMS	0.0303	0.0459	0.1068	0.05674	
Systematic	Sample 1	8.4114	8.4109	8.343	8.3984	0.0348
	Sub-sample 11	8.4427	8.442	8.3743	8.4299	0.0301
	Sub-sample 12	8.4421	8.4416	8.3738	8.4293	0.0300
	Sub-sample 13	8.4428	8.4418	8.3739	8.4299	0.0302
	RMS	0.0270	0.0263	0.0471	0.0171	
Stratified	Sample 1	8.4114	8.4079	8.3954	8.3869	0.0147
	Sub-sample 11	8.4470	8.4438	8.4308	8.4227	0.0266
	Sub-sample 12	8.4453	8.4411	8.4290	8.4212	0.0247
	Subs-ample 13	8.4446	8.4409	8.4283	8.4202	0.0242
	RMS	0.0296	0.0265	0.0175	0.0150	

scale for clarity) which is the graphical representation of Table 3.7, we can see that the error is lowest in the case of stratified sampling as this sampling makes use of localization while selecting the points. Random sampling has the highest error values among the three

Table 3.6: Inter-Subject Registration Error (demonstration for a pair of subjects) for Random, Systematic and Stratified Samplings.

Sampling technique	Samples	Subject 2	Subject 21	Subject 22	Subject 23	RMS
Random	Subject 1	17.9131	17.8738	17.8848	17.8028	0.0602
	Subject 11	17.9340	17.8950	17.9056	17.8241	0.0467
	Subject 12	17.9335	17.8947	17.9052	17.8238	0.0469
	Subject 13	17.9337	17.8949	17.9051	17.8236	0.0470
	RMS	0.0179	0.0252	0.0156	0.0950	
Systematic	Subject 1	17.9131	17.8630	17.9607	17.9561	0.0406
	Subject 11	17.9331	17.8829	17.9807	17.9761	0.0496
	Subject 12	17.9321	17.8820	19.9799	17.9752	1.0340
	Subject 13	17.9317	17.8819	17.9792	17.9743	0.0486
	RMS	0.0166	0.0366	1.0348	0.05792	
Stratified	Subject 1	17.9131	17.9367	17.9050	17.9156	0.0125
	Subject 11	17.9350	17.9588	17.9276	17.9373	0.0290
	Subject 12	17.9326	17.9564	17.9245	17.9351	0.0268
	Subject 13	17.9338	17.9582	17.9260	17.9363	0.0281
	RMS	0.0179	0.0405	0.0120	0.0950	

Table 3.7: Average of RMS of Inter-Sample Registration Error over 170 Subjects for Random, Systematic and Stratified Samplings.

Sampling technique	Average RMS of Inter-Sample Error
Random	0.1142
Systematic	0.0922
Stratified	0.0878

as the points are selected randomly from the point cloud without any specific ordering or localization.

3.2.3 Inter-subject Registration Error

In this experiment, we take two different subjects, say Subject 1 and Subject 2. We also consider one sample and its respective sub-samples from each of the subjects. Table 3.6 shows the ICP registration error between Sample 1 along with its respective sub-samples from subject 1 and Sample 1 along with its respective sub-samples from Subject 2. Simi-

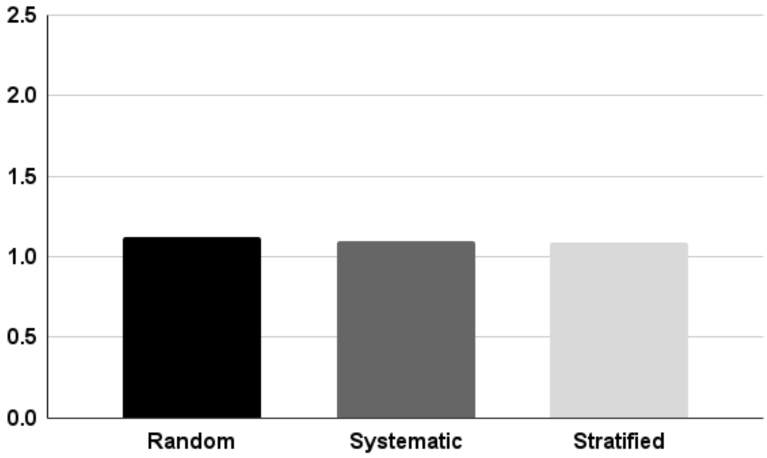


Figure 3.3: Inter-sample Registration Error for Random, Systematic and Stratified Samplings (values are plotted in exponential scale for clarity).

larly, we do the same experiment by comparing every subject with five different subjects and report the results in Table 3.8 where it represents the average of RMS error of Inter-Subject distance over all subjects. From the table, we can see that the registration error using the sub-samples is very similar to that obtained using the original samples. Further, from Figure 3.4 (values are plotted in exponential scale for clarity) which is the graphical representation of Table 3.8, it can be inferred that the stratified sampling and systematic sampling are comparable to each other and are better as compared to random sampling while comparing the inter-subject samples.

Table 3.8: Average of RMS of Inter-Subject Registration Error over 170 Subjects (each compared with 5 other subjects) for Random, Systematic and Stratified Samplings.

Sampling technique	Average RMS of Inter Subject Error
Random	0.1297
Systematic	0.0969
Stratified	0.0988

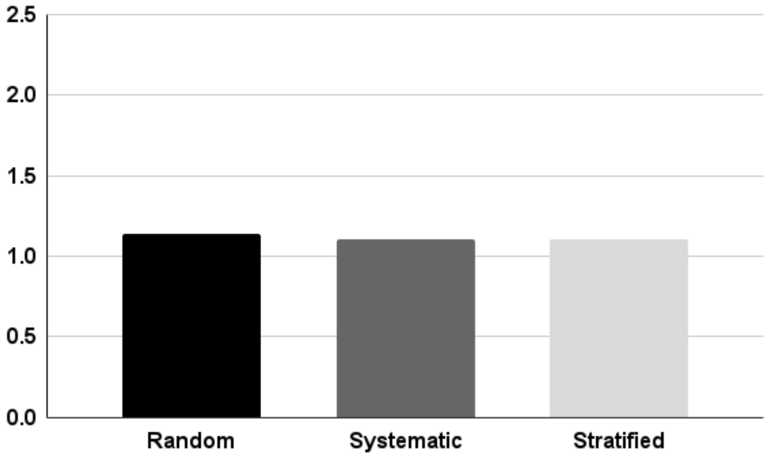


Figure 3.4: Inter-subject Registration Error for Random, Systematic and Stratified Samplings (values are plotted in exponential scale for clarity).

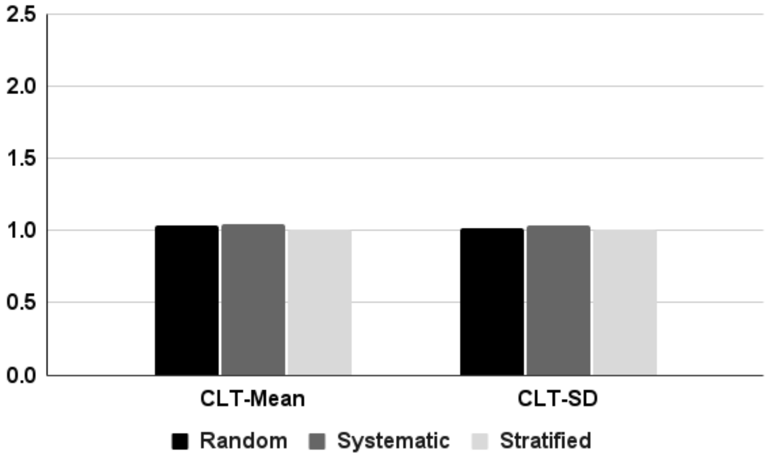


Figure 3.5: Average CLT Mean and Standard Deviation error for Random, Systematic and Stratified Sampling Techniques (values are plotted in exponential scale for clarity).

Table 3.9: Central Limit Theorem on Mean (demonstration for five samples) for Random, Systematic and Stratified Samplings.

	Original	Random	Systematic	Stratified	Error (Random)	Error (Systematic)	Error (Stratified)
Sample 1	88.453	88.417	88.470	88.442	0.035	0.017	0.010
Sample 2	81.863	81.831	81.868	81.861	0.031	0.004	0.001
Sample 3	79.607	79.591	79.599	79.615	0.015	0.007	0.008
Sample 4	94.382	94.343	94.416	94.371	0.038	0.033	0.011
Sample 5	95.836	95.826	95.898	95.853	0.000	0.072	0.027

Table 3.10: Central Limit Theorem on Standard Deviation (demonstration for five samples) for Random, Systematic and Stratified Samplings.

	Original	Random	Systematic	Stratified	Error (Random)	Error (Systematic)	Error (Stratified)
Sample 1	70.075	70.086	70.111	70.082	0.011	0.036	0.006
Sample 2	53.880	53.860	53.865	53.884	0.019	0.014	0.004
Sample 3	60.754	60.781	60.783	60.765	0.026	0.029	0.011
Sample 4	61.548	61.525	61.533	61.548	0.023	0.015	0.000
Sample 5	55.900	55.896	55.916	55.884	0.004	0.016	0.016

Table 3.11: Average of MSE of CLT Mean and CLT Standard Deviation over 170 Subjects for Random, Systematic and Stratified Samplings.

Sampling technique	Average MSE of CLT Mean	Average MSE of CLT Standard Deviation
Random	0.0324	0.0176
Systematic	0.0473	0.0297
Stratified	0.0088	0.0096

Table 3.12: Performance Ranking of Random, Systematic and Stratified Sampling Techniques.

Sampling technique	Computational Time	Sub-sample Similarity	Sample-Sub-sample Similarity	Inter-sample Difference	Inter-subject Difference	CLT
Random	1	3	3	3	3	2
Systematic	3	1	2	2	1	3
Stratified	2	2	1	1	2	1

3.2.4 Analysis using CLT

By choosing five samples from a subject, we create 30 sub-samples for each subject to apply CLT. As stated above, according to CLT, the average mean and the average standard deviation of the samples is similar to the mean and standard deviation of the original population. Tables 3.9 and 3.10 verifies the CLT on the original samples and their sub-samples demonstrated using a set of five demonstrative samples. Further, the average of MSE of CLT mean and CLT standard deviation for different sampling techniques when all 170 subjects of the database are used, are shown in Table 3.11. Results from this experiment prove that the sub-samples created from the original sample have the same discriminative power as the original sample. From the graphs in Figure 3.5 (values are plotted in exponential scale for clarity), we can infer that the stratified sampling is the best technique as it shows the least errors in mean and standard deviation for almost all cases.

3.2.5 Overall Assessment of the Sampling Techniques

An experimental evaluation is performed to rank the three sampling techniques with respect to different criteria such as computational time, sub-sample similarity, sample - sub-sample similarity, coherence with the CLT etc. The results of this experiment are presented in Table 3.12. It is evident from the table that the stratified sampling is the best overall out of the three sampling techniques. Hence, it can be termed as the best suited sampling technique for the data augmentation in various applications such as 3D object recognition and 3D biometric recognition.

3.3 Conclusions

This chapter has proposed the use three different sampling approaches for data augmentation and have proved that the sub-samples created using these approaches all carry the same information and have the same discriminative power as in the original dense samples. These sampling approaches decrease the number of points in the point clouds of the 3D objects, thus increasing the computational and spatial efficiency of the deep neural network used for training without loss of any information. Further, the data augmentation achieved through the sampling process overcomes the problem of overfitting of the deep neural network due to limited training samples per subject. At the end, experimental analysis has been carried out to rank the three sampling approaches with respect to different criteria. It is evident from the analysis that the stratified sampling technique is the best overall out of the three approaches that are being analyzed and have potential to be used for data augmentation in applications such 3D object recognition, 3D biometrics etc.

Chapter 4

3D Face Recognition Using Deep Learning

Most of the proposed object recognition techniques (8; 75; 20; 101) successfully classify and recognize dissimilar unknown objects, and their performance declines when applied for the recognition of object classes which are very similar to each other. In this chapter, we propose a solution to this problem by creating a generalized model for 3D object recognition, which can also be extended to the problem of matching highly similar 3D objects such as biometric recognition using 3D faces or ears. We are proposing a technique that uniquely combines an efficient object recognition architecture with a one-shot learning network. Face recognition is one of the use cases of object recognition for our proposed technique for classification of objects belonging to different human classes, which are very similar to each other. Hence, we demonstrate the effectiveness of our proposed technique by performing 3D face recognition.

One of the most active research fields in Computer Vision is 3D object recognition. It is observed that the recognition is the first step in the semantic analysis of an object. The

The content of this chapter has been taken from the paper titled " A Technique to Match Highly Similar 3D Objects with an Application to Biomedical Security", Multimedia Tools and Applications, Volume 81, Issue 10, pages: 13159–13178 (2022). https://doi.org/10.1007/s11042-020-10161-8 main purpose of object recognition is to recognize previously unknown objects in digital images and in 3D spaces. Object recognition techniques typically use matching, learning, or pattern recognition algorithms based on appearance or feature. With the advent of new algorithms, model, and approaches, 3D object recognition is becoming increasingly effective. The manufacturing industry, autonomous driving, video surveillance, urban planning, control and safety, and augmented reality extensively use 3D object classification and recognition. With rapid advancements in telecommunication technologies, there is significant information circulation over the internet, most of which is confidential and requires authenticated and authorized access. Biometrics, which authenticates people by their physical characteristics, is a useful technology for positive personal identification that can be extended to telecommunication systems over the Internet, thus enhancing the reliability of network services.

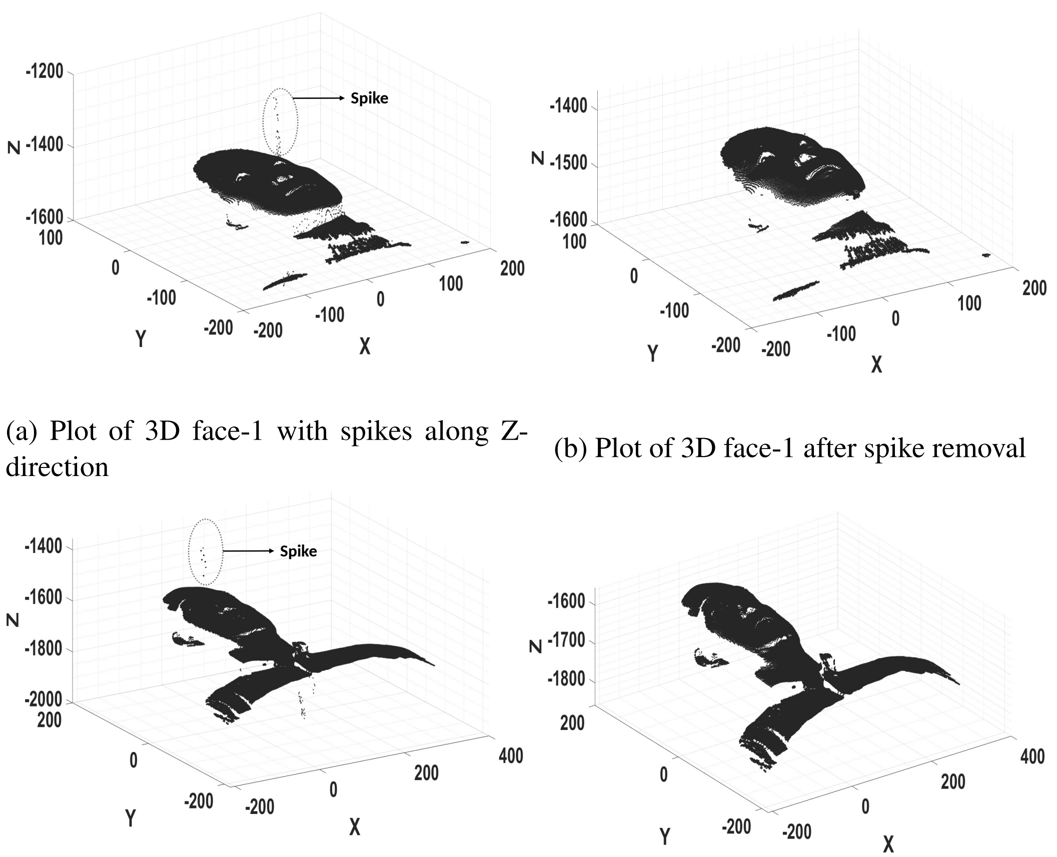
Most of the recent 2D and 3D object recognition works use conventional neural networks (CNN), which successfully extracts features from the data and perform recognition. We propose a deep learning technique that uses a generic solution for 3D object recognition and can be extended to object classification of very similar object classes, like in biometrics. The given model uses the point cloud representation of the 3D objects as input.

4.1 Proposed Technique

The chapter presents a generic solution for 3D object recognition that can be extended to object classification of the very similar classes like in biometrics. The given model uses the point cloud representation of the 3D objects as input. The point cloud of each 3D object contains 50000 points. In order to make the model efficient in time as well as in space, we use multiple random subsets of each point cloud to feed into the model. This also increases the size of our database, which is otherwise limited, containing only 3-4 samples per subject on average hence preventing overfitting. The model uses two networks - PointNet Architecture for feature extraction and Siamese Network for verification. Once the PointNet model is trained on the training samples of all the subjects, we extract features from the second-last dense layer of the architecture. The extracted features are then used to train the Siamese Network, which calculates the similarity score between pairs of feature vectors and finally predicts whether the two objects supplied to it belong to the same or different classes. We apply our model to 3D face recognition to show its robustness.

4.1.1 Preprocessing

Preprocessing of 3D scans is required to eliminate variations in poses as well as 3D noise, which can impact the performance of the object classification model. Generally, this elimination is achieved using frontalization and denoising techniques, respectively. Further, the 3D objects to be classified need to be normally aligned to a base reference image to make the database uniform in accordance with some ground truth. The algorithms such as Iterative Closest Point (15) can be used to align objects in the database to the reference object. However, the PointNet architecture is invariant to geometric transformations such as rotation and translation (70); thus, eliminating the need to frontalize the objects. The point clouds of the 3D objects may get contaminated with spikes due to sensor noise, thus affecting the feature extraction process. Hence, this needs to be handled before object 3D data is used for feature extraction. We make use of a standard spike removal technique that uses a moving averaging filter for denoising of the images. In this method, a sliding window is moved across the object, and the offset along the Z-coordinates is calculated. Further, if the obtained value is found to be greater than a threshold, then the center of the window is translated to the mean offset. This process denoises the 3D scan by limiting the spikes to the given threshold as shown in Figure 4.1.



(a) Plot of 3D face-1 with spikes along Z-direction

(b) Plot of 3D face-1 after spike removal

(c) Plot of 3D face-2 with spikes along Z-direction

(d) Plot of 3D face-2 after spike removal

Figure 4.1: Preprocessing using spike removal

4.1.2 Augmentation

Since 3D data acquisition from objects takes time and hence often, there is very limited data available for 3D objects. The three databases that we have used to train our model, viz. IIT Indore (IITI) Phase-3 database, Bosphorus database, and University of Notre Dame (UND) database, have very few samples per subject. On average, the number of samples per subject is 3 to 4. Even if we use a single sample for testing and the remaining samples for training, the model will not get trained or extract features properly with such a limited number of examples per class. Hence, augmentation is necessary to train the architecture with sufficient samples. In the availability of limited data, the model learns the details and noise of these few samples so well that it negatively impacts the testing of the selected model on new data. To avoid this problem of over-fitting, we increase the variability of the 3D data by enlarging the size of the database by making use of data augmentation.

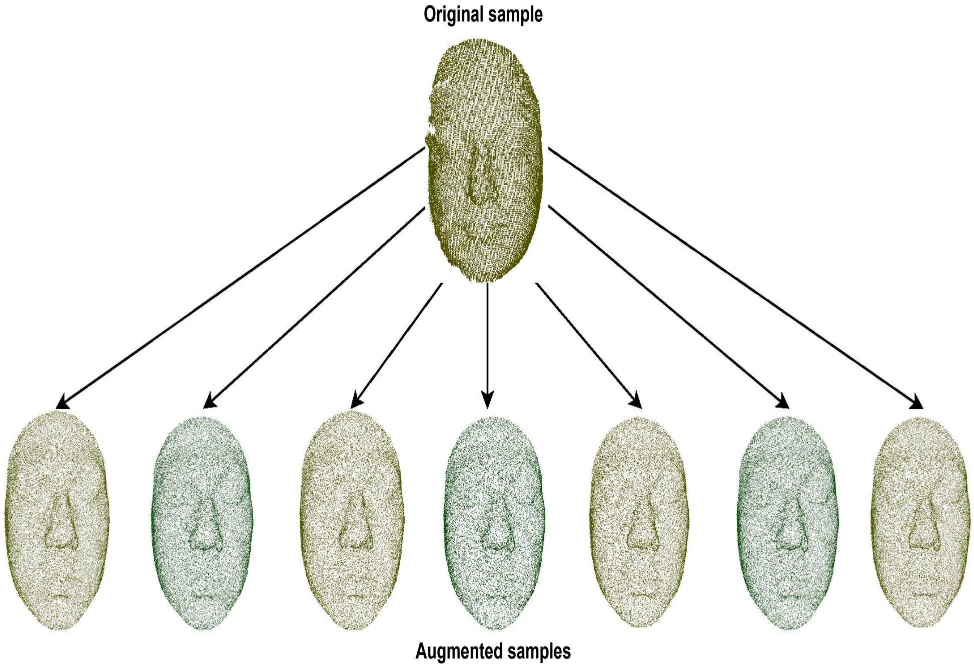


Figure 4.2: Result of the proposed augmentation technique where seven augmented samples are created from an original sample

3D input data for an object, which is in the form of a point cloud, contains an unordered set of 3D points. It is seen that this original set of 3D points for an object contains a huge number of 3D points; however, due to the computational and memory limitations of the system, often, we cannot use the entire point cloud of a single sample for processing. To mitigate this problem, usually, the original point cloud data is sub-sampled, and a reduced size cloud is used for processing. However, in this process, the number of samples for a subject remains the same as was available earlier before sampling. We exploit the use of sampling in a different way and propose its use in data augmentation by increasing the number of samples of the subjects. In our proposed augmentation technique, a fixed number of points are randomly selected from the points in the point cloud of the original sample, creating different unordered subsets. These unordered subsets containing a uniform number of points become our new samples for training the proposed model. The augmentation technique can be explained better with the help of an example. Figure 4.2 shows one such example of augmentation of data from a 3D face point cloud. In this example, the original face contains around 55000 points in the point cloud, from which we are randomly creating seven different subsets of 25000 points each. These subsets will be our new samples in place of the original sample. We can visualize from the figure that the overall geometry and the structure of the face are maintained even after reducing the number of points in the point cloud by approximately half. Our proposed data augmentation method also ensures that the information from the entire point cloud is getting utilized without overshooting the computational and memory requirements.

Further, we validate our proposed data augmentation technique by calculating a similarity score, as discussed later in Section 4.2.2 between the sub-samples created from the same original sample. From the similarity scores shown in Figure 4.7, we observe that the sub-samples show a stark similarity with each other, implying that the features remain intact even after scaling down the number of points. Using this proposed novel technique, we create two different kinds of augmented databases, as explained below.

- **Random Point Cloud Augmentation (Type I)**: In this method, we randomly select a specific number of 3D points from the entire point cloud of the 3D scans repeatedly to generate multiple samples from a single sample of a subject. We do this in a round-robin fashion for each subject to create a uniform number of augmented samples. It is to note that in this case, individual samples may or may not be uniformly used; however, it is guaranteed that each subject will have the same number of samples in the augmented data.

- **Random Point Cloud Augmentation (Type II)**: This technique differs from the previous one in the sense that every available sample of a subject is used uniformly to create the augmented data. This means, from the given samples of the subjects, a fixed number of augmented subsamples are created for each sample. In this case, the total number of augmented subsamples for each subject may or may not be the same. Effectively, the subject for which originally more samples are available will have more subsamples in the augmented data set as well. For example, if a subject has 4 samples, then we create 20 subsamples from each of the 4 samples thus creating 80 augmented sample faces for a single subject.

4.1.3 Proposed Model

Our proposed model combines PointNet architecture (70) with Siamese Network (94) as demonstrated in Figure 4.3. We use PointNet architecture for feature extraction, whereas the Siamese Network for recognition based on these extracted features. The detailed architecture of our proposed model is shown in Figure 4.4.

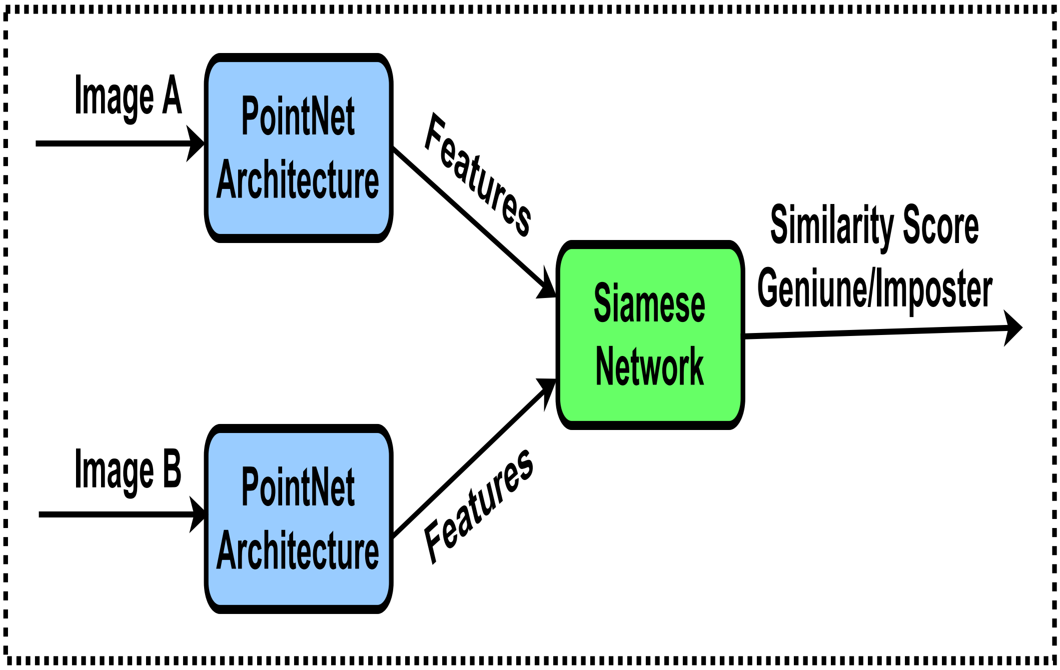


Figure 4.3: Block diagram of the proposed model

4.1.3.1 Feature Extraction using PointNet Architecture

The PointNet architecture is inspired by three essential properties of the point clouds. First, being a set of points, the point clouds are invariant to their *N!* Permutations. Hence, they are unordered. Second, there are interactions among points, meaning that in spite of being in a set, the neighboring points in the space form meaningful subsets that represent a local structure. Lastly, the point clouds are invariant under transformations. Rotating or translating the points of a point cloud altogether does not modify the point cloud itself as it is a geometric object.

The PointNet architecture contains two alignment networks and a max-pooling layer. The first alignment network aligns the input points while the second one is used to align the point features generated by the architecture. An affine transformation matrix is predicted by the network, which is directly applied to the input point clouds, thus aligning all input sets to a canonical space before extracting features. The same alignment technique is extended

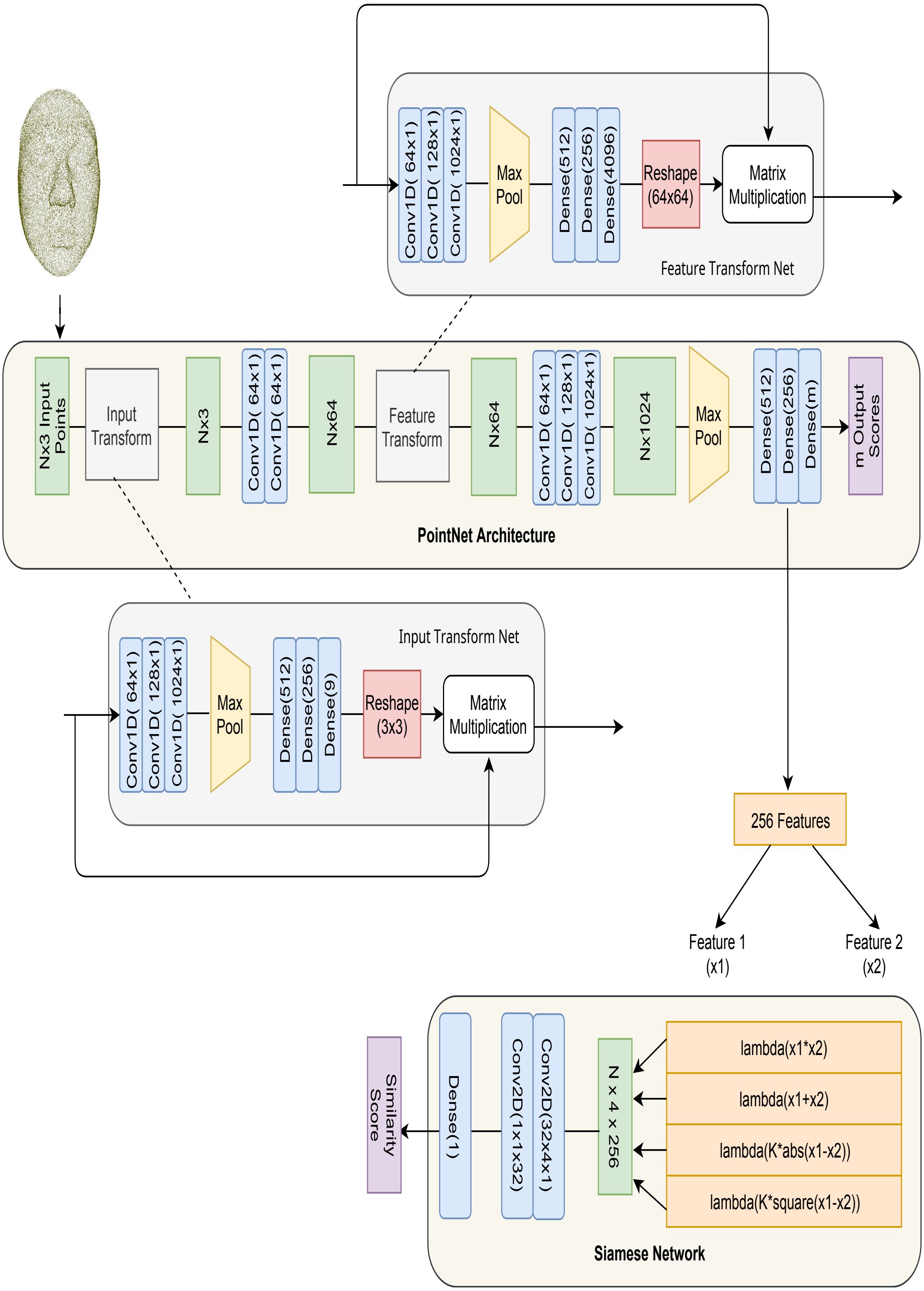


Figure 4.4: Detailed architecture of the proposed model

for aligning the feature space. The two joint alignment networks are used to maintain the geometric invariance property of the point clouds. Further, the max-pooling layer is used to aggregate the information extracted from all the points.

For training of our model, first, we split the database into two sets - the train set and the test set. PointNet Architecture is trained over the train set of the data. Typically, PointNet gives the class of the input sample as the output of its last layer. Instead, we are using PointNet till its second-last dense layer. This layer outputs the feature vectors for the given samples. We extract these features of our train set to then train the Siamese Network.

4.1.3.2 Recognition using Siamese Network

In standard classification problems, a probability distribution over all the classes is generated after feeding the input image to a series of layers. However, the Siamese Network uses a similarity score between the test image and a reference image to check if they belong to the same or different classes based on a threshold value to train itself. This similarity score lies in the range 0 to 1, 0 indicating no similarity and 1 indicating full similarity. Thus, the Siamese Network learns a similarity function which takes in two inputs and expresses how similar they are to each other.

For training the Siamese Network, we need two sets of pairs from the extracted features - genuine pairs that contain features from the samples belonging to the same class and imposter pairs, which contain features from the samples belonging to different classes. The Siamese Network uses four functions to determine the relationship between the features in the pairs, namely, addition, multiplication, absolute difference, and the square of the absolute difference between the two features as shown in Figure 4.4. The results of these four operations are concatenated and passed on to the Convolutional Layers for training. The network gets trained on these feature pairs to predict whether they are genuine or imposter.

We create all possible genuine pairs for each class and all possible imposter pairs by taking combinations of each class with every other class to make the training more robust.

Finally, we test our model by first passing our test set through the trained PointNet Architecture that generates feature vectors for the test set and then compares each of these test features with train features of all classes to predict whether the combination is genuine or imposter. The output of the Siamese Network is the probability that the pair is genuine. We use a threshold on the probability to decide if the pair is genuine or imposter. By pairing the test image with a reference image from each class, the Siamese Network calculates a similarity score for each pair and predicts the subject class with the highest similarity score as the class of the test image.

4.1.4 Novelty in the Proposed Technique

The proposed work attempts to use a generic 3D object recognition model and extend its use-case to match highly similar 3D objects. The suggested model can be trained on any class of biometrics like 3D faces, ears since the structure of the model is independent of the recognition classes. Further, to overcome the problem of overfitting due to limited samples per subject, we have used novel data augmentation technique. The Point Cloud Augmentation technique not only increases the number of samples for each subject but also reduces the time and space complexity while training the model since the number of points in each sample is reduced by almost half. Finally, we efficiently use PointNet Architecture that directly consumes the less-bulky point clouds as inputs keeping them invariant to affine transformations thus, making preprocessing minimal, for feature extraction and pass these feature vectors to the Siamese Network that calculates the similarity score between the feature vectors of the input images to label them as a genuine or imposter pair.

4.1.5 Application of the Proposed Technique

Biometrics uses measurable physical and behavioral characteristics such as fingerprints, iris scans, palm prints, or facial recognition to enable the establishment and verification of an individual's identity. 3D biometric-based authentication systems are more reliable, convenient, and faster than the established password systems. Different organizations such as medical organizations use different mechanisms for verifying the patient identity. The process is required to ensure privacy, security, usability, and high-performance. Our proposed 3D biometric security architecture has high speed and accuracy (refer Section 4.2), thus making it deployable by hospitals, healthcare organizations and many other organizations which require similar applications, for identity verification solutions. For example, it can be used for following different applications in the biomedical and healthcare sector.

- **Enrollment:** New subjects with sufficient 3D samples can be added to our model, which pre-processes and augments them. Feature extraction is then performed using our trained PointNet architecture. The Siamese Network is retrained using extracted features to add the new subjects to the database.

- **Identification and Verification:** The subjects can be passed through our architecture, which extracts features from the input samples using PointNet architecture and matches them with existing subjects using the Siamese Network. Biometric technology can lower healthcare fraud instances and increase privacy and security. 3D Facial recognition effectively serves as a modality that is interpretable by a human and acquired using common devices such as mobile apps, Kinect, or 3D scanners. Biometric technology has the :

- **Healthcare Fraud:** Complex and fragmented health insurance regimes are highly susceptible to frauds like filing incorrect healthcare claims or accessing services illegitimately, thus driving the deployment of biometrics. The proposed solution can determine if someone is trying to impersonate a false identity to receive healthcare without payment or picking up prescriptions at a pharmacy.

- **Personnel Authentication and Authorization:** Secure identification in healthcare is critical in controlling logical access to confidential data as well as physical access to wards and buildings by hospital staff via verification. Unlike biometric solutions such as 3D face recognition, authentication schemes such as password resets, cards, and PIN programs can get lost, stolen, forgotten, or shared.

- **Patient Identification:** The solution has applications in client registration, treatment tracking, and health insurance cards, thus avoiding patient's wrong identification due to patient registration errors or duplicate medical records in the system. Verified patients obtaining the correct treatment ensures safety. With the use of 3D biometrics, one can also identify proper insurance status, thus increasing fraud protection. Biometrics are quick and efficient, eliminating the need for manual input of data, which can be unreliable. The application also work for unresponsive patients, thus combining high security with convenience.

- **Mobile Healthcare Applications:** Since mobile biomedical apps collect community and clinical health data and deliver medical information to doctors, researchers, and patients, biometric technology can be highly leveraged for more secure and convenient login to telemedicine portals. Figure 4.5 summarizes the application of our proposed technique, particularly to Biomedical Security. The proposed solution can be deployed in several other similar applications.

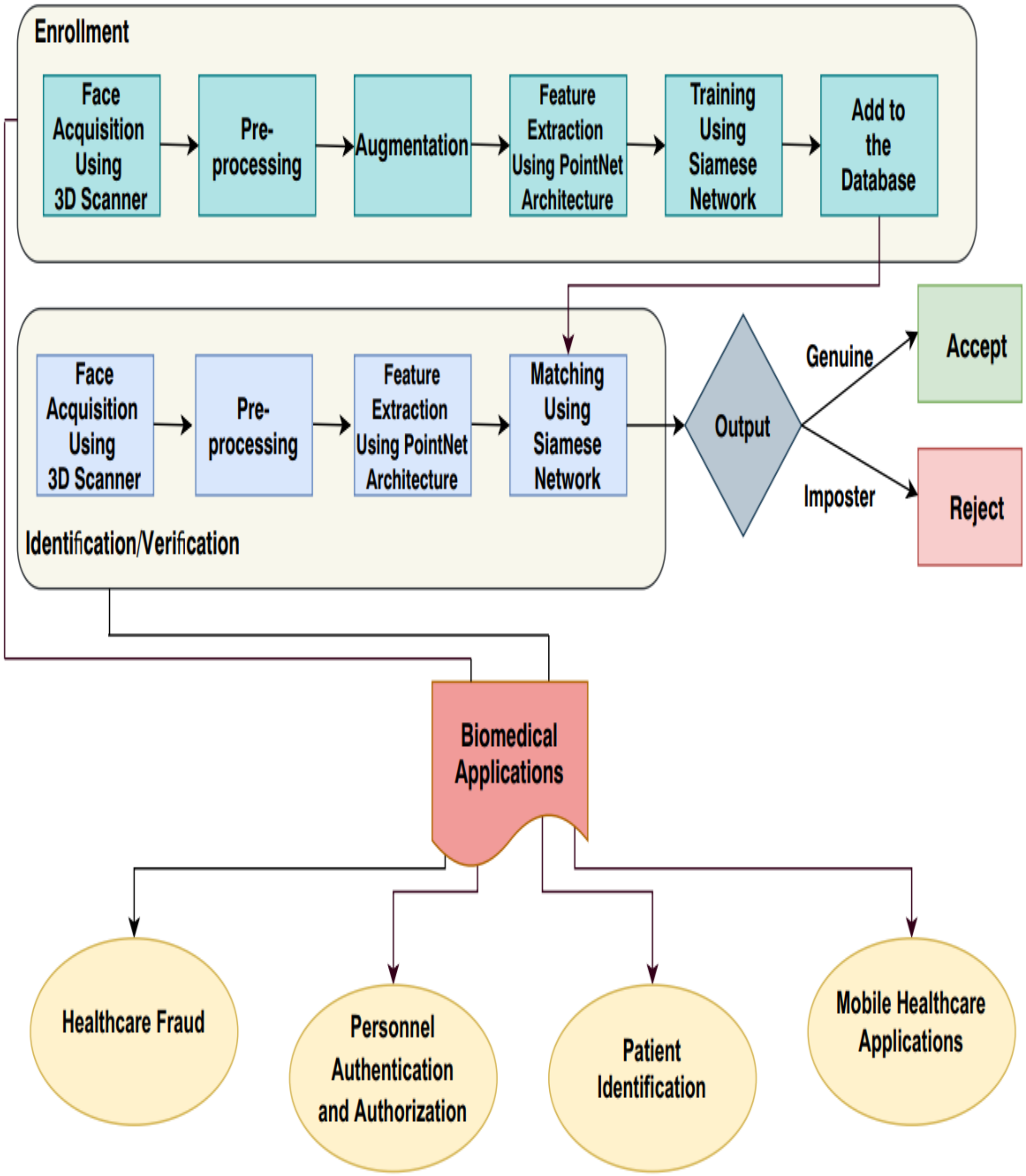

Enrollment
Face Acquisition Using 3D Scanner
Pre-processing
Augmentation
Feature Extraction Using PointNet Architecture
Training Using Siamese Network
Add to the Database
Face Acquisition Using 3D Scanner
Pre-processing
Feature Extraction Using PointNet Architecture
Matching Using Siamese Network
Identification/Verification
Output
Genuine
Imposter
Accept
Reject
Biomedical Applications
Healthcare Fraud
Personnel Authentication and Authorization
Patient Identification
Mobile Healthcare Applications

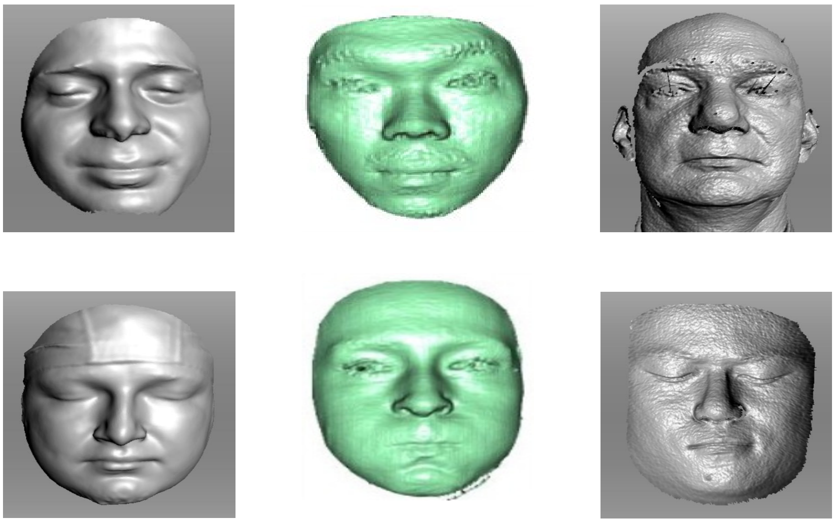
Figure 4.5: Application of the proposed technique in biomedical security

4.2 Experiments and Results

To demonstrate the effectiveness of the proposed recognition model, we now apply it to the problem of face recognition. We expand these databases using data augmentation to improvise feature extraction and prevent overfitting. These databases are randomly split into two sets: 70% for training and 30% for testing. We evaluate classification performance based on verification accuracy, Equal Error Rate (EER), Receiver Operating Characteristics (ROC) curve, and Rank-1 accuracy.

4.2.1 Databases used

We use three 3D face databases to test our model - our in-house database called IIT Indore (IITI) Phase-3 database, Bosphorus database, and University of Notre Dame (UND) database. Few sample images from these databases are shown in Figure 4.6. The IITI database contains 170 subjects, with a total of 445 samples. An Artec 3D EVA scanner has been used to acquire these 3D face scans. Many of the samples in the database are unaligned. However, we do not need to align the images since PointNet architecture is invariant to geometric transformations. The Bosphorus database (78) contains 105 subjects with 4666 3D facial scans. This database has 3D face samples with a rich set of expressions, systematic pose variations, and various occlusions. Out of these, we use 299 neutral faces in our experimentation to compare the performance with the other databases. An Inspeck Mega Capturor II 3D scanner has been used to acquire these facial data. The face scans in the database are aligned and contain minimal noise as the noise reduction is already made at the time of data acquisition by experimentally optimizing the acquisition setup. The University of Notre Dame database (ND-collection D) (14; 21) contains 277 subjects with 953 aligned 3D face scans. A Minolta Vivid 900 3D range scanner has been used to acquire these images. These face scans contain considerable noise in the form of spikes. Hence, we are using spike removal to denoise the 3D scans. The summary of these databases is provided in the Table 4.1.



(a) IITI database

(b) Bosphorus database

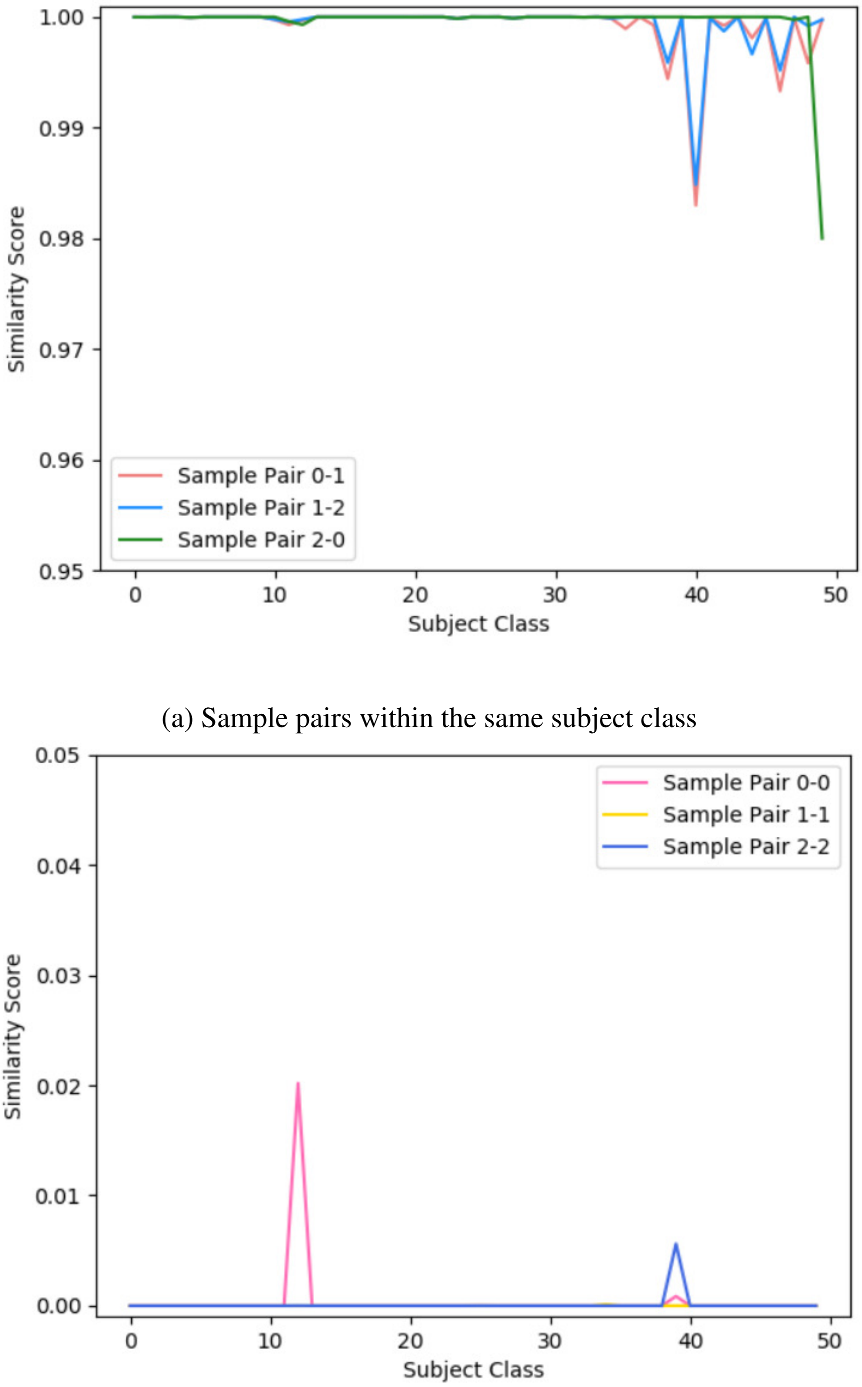
(c) UND database

Figure 4.6: A few sample scans from different databases used in the experimental evaluation of the proposed model

4.2.2 Augmentation of Databases

Since the number of samples per subject is quite low for the given databases, we augment them to increase the sample size. As discussed in Section 4.1.2, we are applying two types of point cloud augmentation. Type I augmentation creates 21 samples per subject, whereas Type II augmentation creates 21 samples for each original sample present in a subject.

In the Siamese Network, similarity scores are calculated, such that the imposter pairs have a score close to 0, and genuine pairs have a score close to 1. When we create multiple augmented samples from the available samples of a subject using augmentation, we need to validate that the features of the original samples are not compromised. From Figure 4.7, we can see that the similarity scores for two random samples within the same subject class for 50 subjects are close to 1, and those for the same samples in two random subject



(a) Sample pairs within the same subject class

(b) Sample pairs of different subject classes

Figure 4.7: Similarity scores distribution for sample pairs within the same and different subject classes

Table 4.1: Details of the 3D databases used in the experimental evaluation of the proposed model

Database	Number of subjects	Original number of samples	Number of samples after Type I augmentation	Number of samples after Type II augmentation
IITI	170	445	3570	9345
Bosphorus	105	299	2205	6279
UND	277	953	5817	20013

classes is close to 0, thus implying that the intra-class similarity and inter-class dissimilarity is maintained in spite of the reduction in points during augmentation.

4.2.3 Performance on 3D Databases

We split each of the IITI, Bosphorus, and UND databases into a 70-30 ratio for each subject to create train and test sets. Table 4.2 gives the training accuracy on the train set when it is passed through the PointNet architecture. As the output of the PointNet architecture gives the subject class of the given sample, the training accuracy calculated is the recognition accuracy. This accuracy indicates how well our model is extracting features from point clouds. To further improve this accuracy, we train the extracted features on the Siamese Network.

Table 4.2: Recognition accuracy of training on PointNet architecture (values in percentage)

Database	Type I Augmentation	Type II Augmentation
IITI	87.50	88.86
Bosphorus	84.60	83.97
UND	79.95	77.08

For the Siamese Network, the distribution of the similarity scores for imposter (Class 0) and genuine (Class 1) classes for each of the three databases is shown in Figure 4.8. We observe that the maximum concentration of similarity scores occurs near 0 (for an imposter

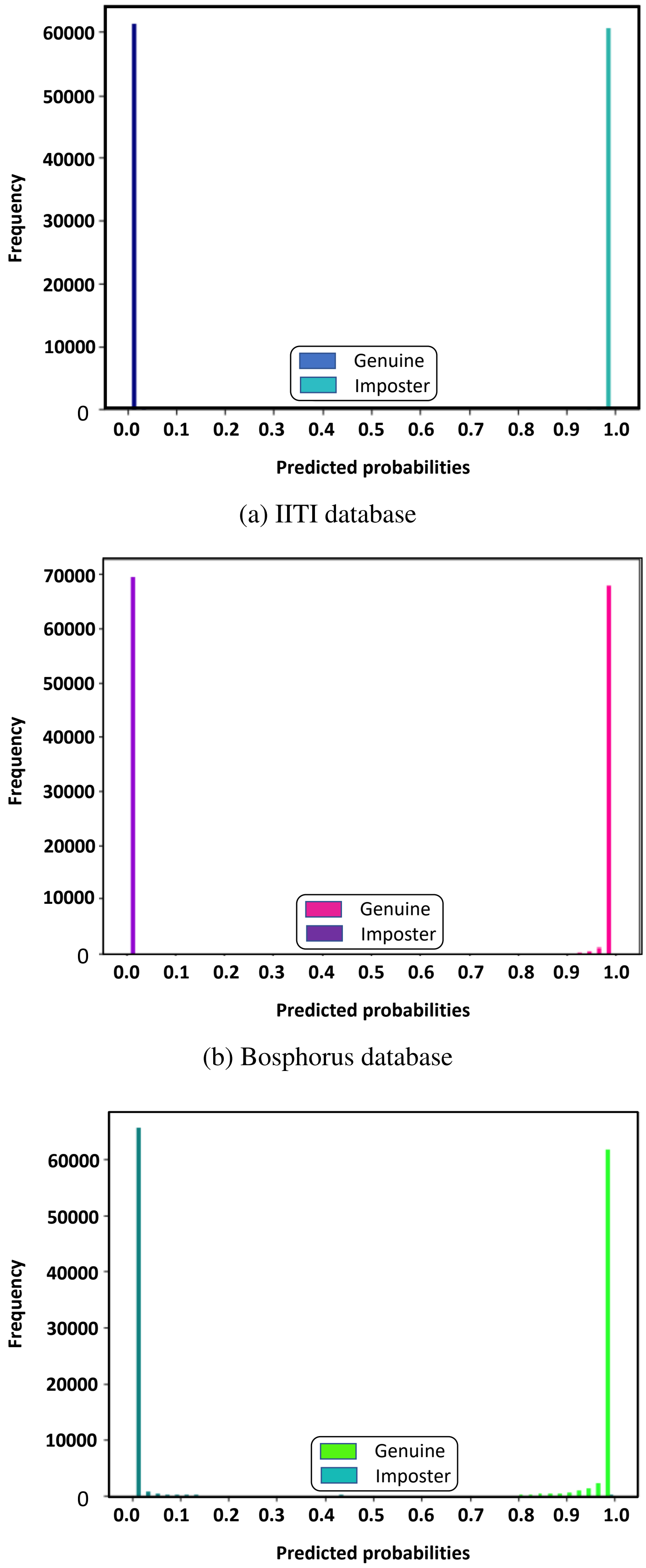
Table 4.3: Results on the proposed model in terms of verification accuracy (values in percentage)

Database	Type I Augmentation (for train-test pairs)	Type I Augmentation (for test-test pairs)	Type II Augmentation (for train-test pairs)
IITI	99.91	99.95	99.21
Bosphorus	99.66	99.69	98.30
UND	98.60	98.53	96.90

Table 4.4: Inference time on the proposed model (values in seconds)

Database	Feature extraction time	Matching time
IITI	2.9258	0.3062
Bosphorus	2.7765	0.3935
UND	2.6326	0.4659

pair) and 1 (for a genuine pair). It is evident from the graphs that the similarity scores are quite accurate for both genuine and imposter classes, *i.e.* the score for most of the imposter pairs is close to or equal to 0 and that for genuine pairs is close to or equal to 1. This means that our proposed model is capable of creating very clear segregation among the genuine and imposter pairs. Also, we use a train-test split of 70-30 while creating genuine and imposter pairs for the Siamese Network, *i.e.* 70% of the total pairs are created from train features obtained from PointNet architecture which are used to train the Siamese Network and remaining 30% are created from the test features extracted from the trained PointNet architecture which are used for testing. After training the Siamese Network on these 70% of the pairs, we can test our model in two ways. One way is to create pairs such that one sample is from the test set, and the other is from the train set (test-train pairs). This testing shows how well the proposed model compares an unknown sample with the known samples to determine the unknown sample's subject class. Another way is to create pairs from the test set itself (test-test pairs) to examine the ability of the model to recognize genuine and imposter pairs from the unknown set of samples.

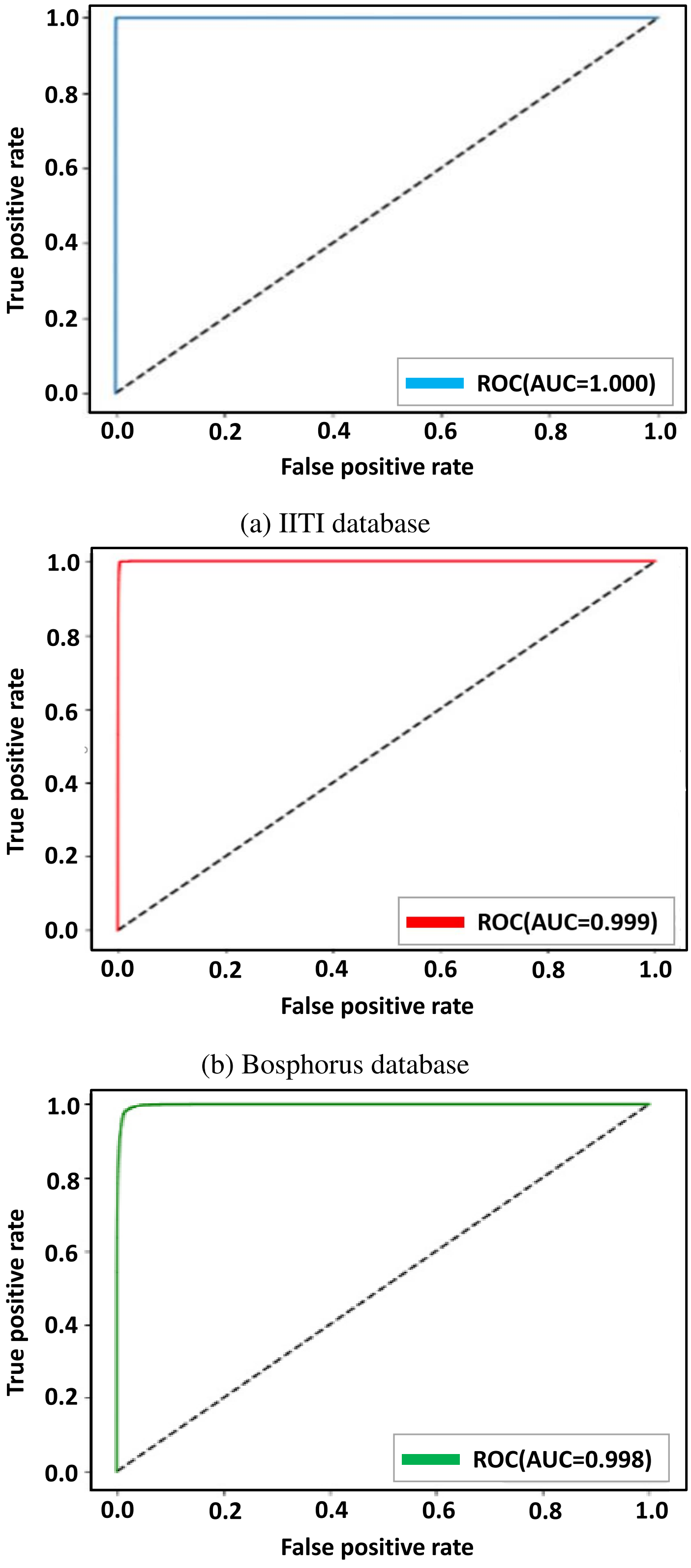


(a) IITI database

(b) Bosphorus database

(c) UND database

Figure 4.8: Similarity score distribution for IITI, Bosphorus and UND databases

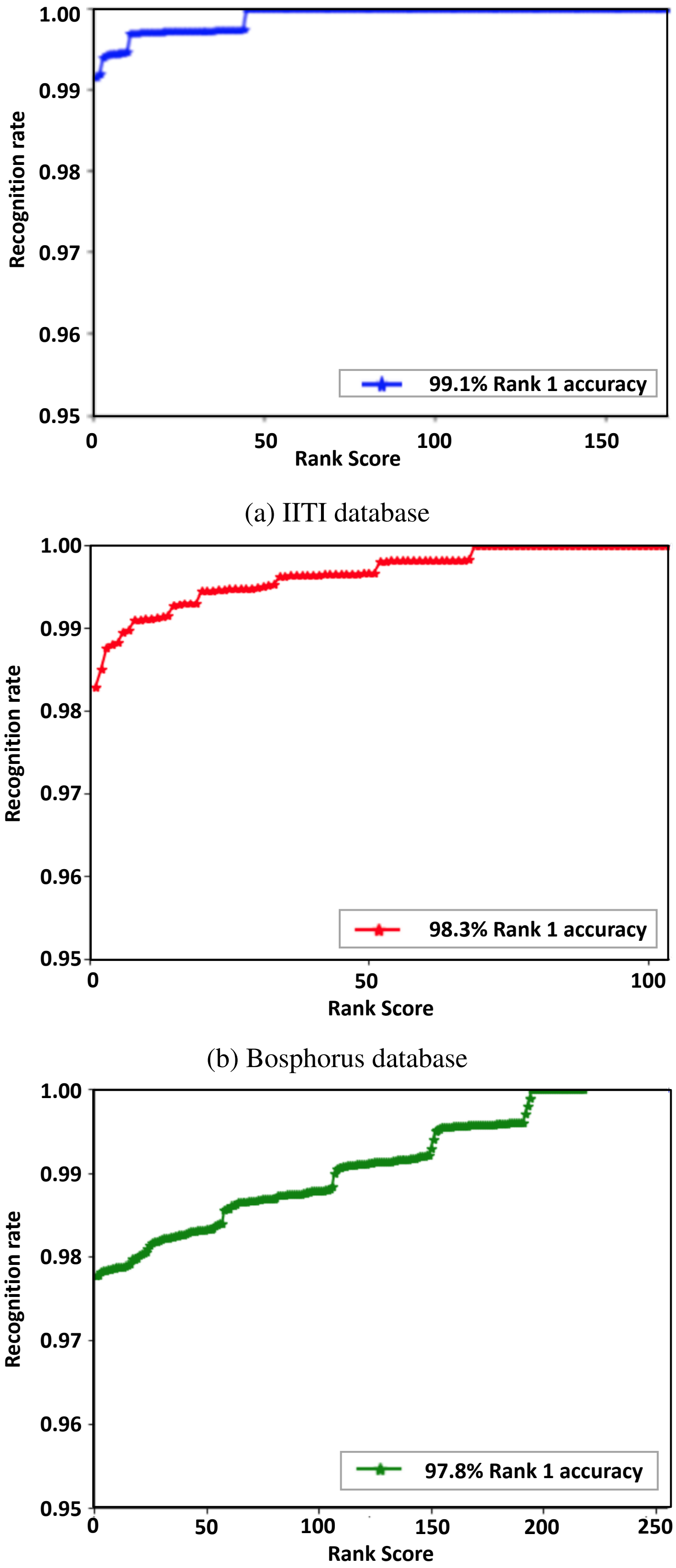


(a) IITI database

(b) Bosphorus database

(c) UND database

Figure 4.9: ROC curves for different databases



(a) IITI database

(b) Bosphorus database

(c) UND database

Figure 4.10: CMC curves for different databases

Table 4.5: Verification accuracy, Rank-1 accuracy, AUC, and EER values for the proposed model in case of Type I augmentation (values in percentage)

Database	Verification accuracy	Rank 1 accuracy	AUC	EER
IITI	99.91	99.10	1.000	0.0004
Bosphorus	99.66	98.30	0.999	0.0015
UND	98.60	97.00	0.998	0.0080

We evaluate the proposed model based on verification accuracy, Rank-1 accuracy, the area under the ROC curve (AUC), and EER. Verification accuracy is calculated as the probability of two input genuine pairs being predicted as genuine and two input imposter pairs being predicted as imposter. Rank-1 accuracy is defined using the Cumulative Match Characteristic (CMC) curve. The CMC curve provides the detection precision for each rank, which is either genuine or imposter in our case. The Y-intercept of the curve gives the Rank-1 accuracy of the model. AUC is a performance metric that measures the degree of separability between classes at various thresholds settings. Higher the AUC, the better the model is at predicting a class correctly. An ideal classifier has AUC equal to 1. The point on the ROC curve corresponding to an equal probability of wrongly classifying a positive or negative sample gives the EER. It is obtained by intersecting the ROC curve with the diagonal of the unit square. Table 4.3 shows the verification accuracy on the proposed model for Type I and Type II augmentations. For Type I augmentation, accuracy values are shown for both train-test and test-test pairs. From the experiments, we find that the best results are achieved for Type-I augmentation, as demonstrated in Figure 4.9 and Figure 4.10 based on the following performance metrics. The performance of the proposed model on different metrics for different databases is shown in Table 4.5 for Type I augmentation. All three databases show remarkable results on the given performance metrics. The efficiency of the proposed model in feature extraction and matching is shown in Table 4.4. It can be seen from the table that for all the three databases, the time required by the proposed model to extract the features from a pair of testing samples and to identify whether the pair is genuine or imposter is only a few seconds, making the model remarkably efficient. Moreover, while working with the UND database, we also used the unprocessed samples, which gave a verification accuracy of 79.8% on our proposed model. This is significantly less than the expected accuracy and hence, validates the requirement to preprocess the UND database by removing spikes.

The three databases that we have used for testing of our model are IIT Indore, Bosphorus, and the University of Notre Dame (UND) databases. Our in-house database (IIT Indore) has been used for the first time, and due to this we could not compare our results on this database with other papers. For the Bosphorus database, only neutral faces for testing have been used as the other two databases only contain neutral faces and we wanted to compare these results. The recently reported techniques have given the performance on the complete Bosphorus database. In our future study, we would evaluate our model on the complete database. Also, none of the papers have reported the performance on the UND database (ND-collection D) that we have used, solely based on 3D face scans and using all the subject classes available. Hence, we have not been able to compare the performance of our method with other reported techniques.

4.3 Conclusions

In this work, a generic 3D object recognition technique is proposed that gives remarkable accuracy even on highly similar objects. In the technique, we construct a model that improves over the existing PointNet architecture by combining it with the Siamese Network with minimal preprocessing. To overcome the problem of limited 3D data samples, we propose the use of data augmentation for which point sampling is carried out on the point clouds of the available 3D image samples. We evaluate the proposed model on three 3D face databases, namely, IITI, Bosphorus, and UND databases, achieving Rank-1 accuracy of 99.1%, 98.3%, and 97.0% with Equal Error Rates (EER) 0.0004, 0.0015, and 0.0080 respectively. Our experimental results show that the best performance is achieved when Type I augmentation is used. The graphical analysis of these results also verifies that the proposed model achieves high accuracy, implying perfect segregation between genuine and imposter pairs. The time required by the model for feature extraction and matching is only a few seconds, making the model remarkably efficient. Because of the high efficiency and high accuracy of our model, it can be effectively used for biometric authentication in different applications.

Chapter 5

Multimodal 3D Face Recognition Using Transfer Learning

Face recognition applications mostly use 2D data (2D face images) as an input. In most of the cases, the 2D data provides remarkable results, however, its performance decreases when there is poor illumination, change in orientation of the face and presence of noise. Another major drawback of using 2D data for face recognition is that when the model is used for biometric authentication, it can be easily forged. The above problem limits the usage of 2D data for authenticity based applications. To resolve these issues, 3D data can be used. However, 3D data has its own limitations; for example, the computational power and resources required to process 3D data are very high. Therefore in this chapter, we propose the use of 2.5D data (depth images) in place of 3D data for face recognition where the 2.5D data is obtained from 3D data. Here, 2.5D data represents 3D face data in terms of 2D depth images. There are more number of state-of-the-art object recognition models using 2D data as compared to 3D data. By using 2.5D data, we can extract features and use

The content of this chapter has been taken from the paper titled ”Robust Face Recognition using Multimodal Data and Transfer Learning”, Journal of Electronic Imaging, Volume 32, Issue 4, SN. 042105 (2022). `https://doi.org/10.1117/1.JEI.32.4.042105` transfer learning from some of the most powerful neural networks proposed for 2D data.

Face recognition is the technique of recognizing a person's face in an image and determining to whom it belongs. It is thus a form of personal identification. Early face recognition systems relied on an earlier version of facial landmarks that are extracted from images, such as the relative position and size of the eyes, nose, cheekbone, and jaw. However, because these face quantifications were retrieved manually by the computer scientists and administrators using the face recognition software, these systems were highly subjective and prone to error (73). Face recognition software generally uses computer algorithms to extract unique features from a person's face and use them for recognition. Facial details, such as eye distance or outline of face, are then transformed into a mathematical representation and compared to data from other faces in a face recognition database. A face template contains some unique features extracted from the face image of an individual, which can be used to differentiate a person from others. Face recognition can also be considered a type of biometric authentication. While there is growing interest in other biometrics, the face recognition platform is still used in several applications such as surveillance and law enforcement. The 2D face recognition systems have attained good performance in controlled environments since last few decades. The accuracy of 2D face recognition has improved dramatically, particularly since the advent of deep learning. However, the inherent limitations of 2D images, such as pose, expression, illumination variations, occlusion, and image quality, continue to provide a challenge to these systems (110).

3D face recognition has become an active research topic in recent years as it is not affected by the shortcomings of 2D face recognition like pose, illumination, and expression (7). 3D face images provide rich geometric information that gives more discriminative features (40). 3D face models include more shape information than 2D images. Furthermore, in terms of scale, rotation, and lighting, 3D face images are relatively unchanged (12). In 3D face recognition, 3D face models are commonly used for training and testing purposes. 3D scanners can capture both 3D mesh/point cloud and corresponding information regarding texture. One of the most challenging aspects of 3D face recognition is acquiring 3D images, which requires specialized hardware. The acquisition system actively generates invisible light (e.g., an infrared laser beam) to illuminate the target face and measure the reflectance to determine the target's shape features.

3D face recognition systems may be classified into traditional and deep learning-based approaches based on feature extraction methods. Traditional techniques, such as iterative closest point (ICP) (15; 91), principal component analysis (PCA), linear and nonlinear algorithms, are being used to extract face features in conventional approaches. In terms of deep learning-based methodologies, all of them practically rely on pre-trained networks that are subsequently fine-tuned using the converted data (e.g., 2D images from 3D faces). Visual Geometry Group (VGGNet) (66), Residual Neural Network (ResNet) (32), Artificial Neural Networks(ANN) (47), and recent light weight CNNs like MobileNetV2 (77) are popular deep learning-based facial recognition networks. A 3D face image is an abstract representation of face and can be represented as depth image, point cloud, polygon mesh and voxel. The Figure.1.1 has shown examples of these face representations. These representations have been used in the literature to extract the features, and then to perform 3D face recognition. A depth image gives us the "depth" of the object or the "z" information of the object in the real world and the intensity values in the image represent the distance of the object from a viewpoint. In surface modelling methods, like mesh, the topological information (connectivity between the points) of the points can be obtained, while in the point cloud, the data is unstructured and the topological information is absent. The voxel image is a volumetric representation of each point where the change in voxel size affects the resolution of the 3D image.

5.1 Proposed Technique

In our proposed technique, we convert the 3D data into 2.5D data and preprocess it as presented in Figure 5.1. Further, we augment the data to avoid overfitting, as the number of samples is less in 3D data. This increases the size of our dataset which helps in getting better results. We use pre-trained ResNet-34 (32) architecture for feature extraction and Siamese Network (94) for face verification. In this section, we discuss the network architecture and methodologies used in our approach.

5.1.1 Preparation of 2.5D Dataset

The UND 2D dataset contains RGB face images of size 640 x 480 pixels as shown in Figure 5.2a. There are a total of 277 subjects with 953 samples in this dataset. The UND 3D dataset contains (x,y,z) coordinates of the human face in the excel file. We convert this 3D data to 2.5D Image (depth image). The (x,y) coordinates of a point correspond to the pixel location and the z coordinate corresponds to the gray scale intensity of the pixel as shown in Figure 5.3a. The UND 3D dataset also has 953 samples, each corresponding to its respective sample in the 2D dataset.

5.1.2 Preprocessing

In our experiment, we have used UND 2D and 3D datasets. These datasets are partially preprocessed, but we need to further preprocess them as per our network requirements. The UND 2D dataset images are further cropped to contain only the faces and remove the unnecessary data (like shoulders and the background) to improve the accuracy as shown in Figures 5.2b and 5.3b respectively. Images are cropped using pre-trained Haar cascade model. These images are then resized to 224 x 224 pixels to facilitate the network requirement. We crop the 2.5D images similar to the way we cropped 2D data as mentioned above.

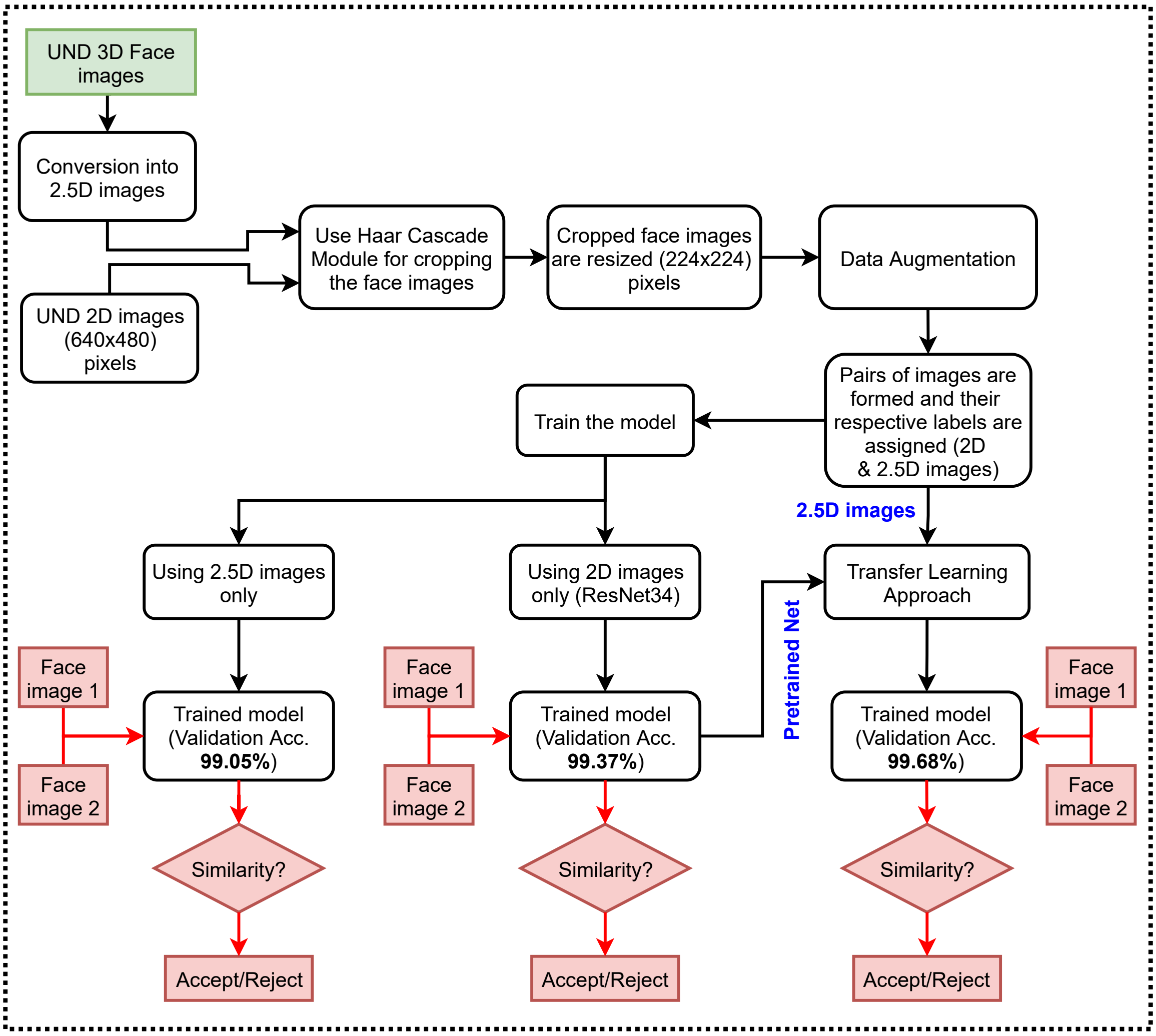

UND 3D Face images
Conversion into 2.5D images
UND 2D images (640x480) pixels
Use Haar Cascade Module for cropping the face images
Cropped face images are resized (224x224) pixels
Data Augmentation
Pairs of images are formed and their respective labels are assigned (2D & 2.5D images)
Train the model
2.5D images
Using 2.5D images only
Using 2D images only (ResNet34)
Transfer Learning Approach
Pretrained Net
Face image 1
Face image 2
Trained model (Validation Acc. 99.05%)
Trained model (Validation Acc. 99.37%)
Trained model (Validation Acc. 99.68%)
Similarity?
Accept/Reject

Figure 5.1: Block diagram of the proposed technique

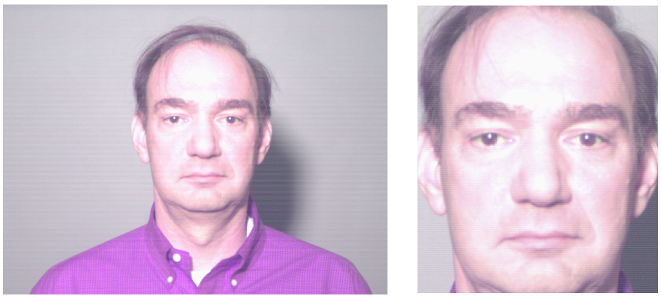
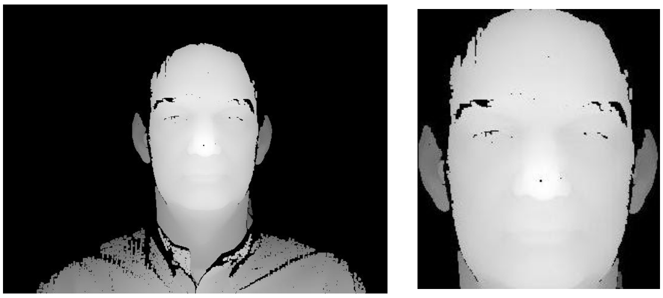

(a) (b)

Figure 5.2: A sample image from UND 2D face database and its cropped version



(a) (b)

Figure 5.3: A sample image of converted 2.5D data and its cropped version

5.1.3 2D Data Augmentation

Due to the lack of data, neural networks often suffer from overfitting and/or underfitting during training, which has a significant impact on the model's efficiency. Data augmentation is a technique for generating new training data from the existing data. This is achieved by transforming samples from the training data into new and unique training samples using domain-specific approaches. One of the most often used data augmentation methods is image data augmentation, which modifies the original image into a new image of the same class by performing multiple operations. These operations can be shift, twist, zoom, and other image manipulation operation.

Convolutional Neural Networks (CNN) are a type of modern neural network that can extract features regardless of where they appear in the picture. The model will learn characteristics from newly augmented images obtained by shifting, flipping, rotating, and adding other transform operations to the original images, which will make the model more robust. Image data augmentation is usually applied only to the training dataset, and not on the evaluation data.

Data augmentation differs from image preprocessing techniques like pixel scaling and image resizing, where it must be implemented equally throughout the whole dataset. We have used scaling, random rotation, brightness adjustment, and Gaussian noise for data

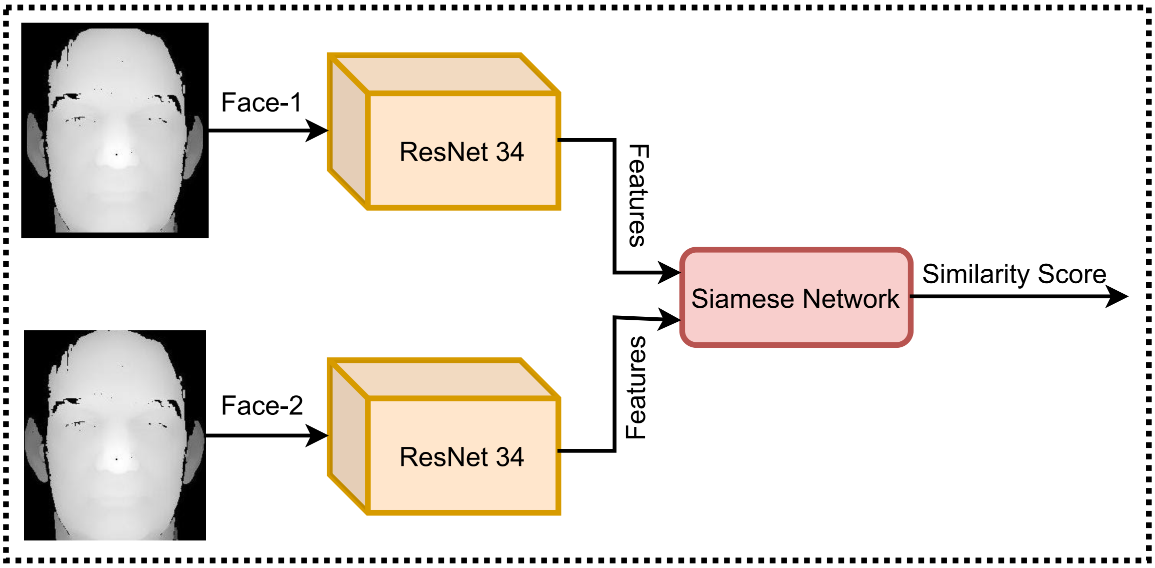


Figure 5.4: Network architecture of the proposed model

augmentation. For each sample in the dataset, we create five other samples by using combination of one or more of the augmentation techniques mentioned above. Table 5.1 contains details about the final number of samples.

Table 5.1: Details of 2D and 2.5D images used in the experimental evaluation of the proposed model

Dataset	No.of images	No.of images after augmentation
UND 2D images	953	5718
Converted 2.5D images	953	5718

5.1.4 Network Architecture

Over the last few years, there have been a series of breakthroughs in the field of Computer Vision. Especially with the introduction of deep convolutional neural networks, we are getting state-of-the-art results on problems such as image classification and image recognition. So, over the years, researchers tend to make deeper neural networks (adding more layers) to solve complex tasks and to also improve the classification and recognition accuracies. However, it has been seen that as we keep on adding on more layers to the neural network, it becomes difficult to train them and the accuracy starts saturating and then degrades also. This is not due to overfitting or underfitting, but due to a problem called vanishing gradient problem. This is due to the fact that when the network is too deep, the gradients of the loss function is easily shrink to zero after several chain rule applications. As a result, the weights are never updated, and thus no learning occurs.

5.1.4.1 Feature Extraction with ResNet (Residual Network)

Our proposed network architecture is shown in Figure 5.4. In our proposed approach, we use pre-trained ResNet-34 for feature extraction. The Resnet-34 is a state-of-the-art image classification model with 34 layers of convolutional neural networks. This is a model that has been pre-trained on the ImageNet dataset, which has 100,000+ photos divided into 200 classes. However, it is different from standard neural networks in the sense that it uses the residuals from each layer in the succeeding connected layers. The skip connections in ResNet solve the problem of vanishing gradient in deep neural networks by allowing an alternate shortcut path for the gradient to flow directly through the skip connections backward from later layers to initial filters. The other way that these connections help is by allowing the model to learn the identity functions which ensures that the higher layer will perform at least as good as the lower layer, and not worse.

5.1.4.2 Matching using Siamese Network

A Siamese Neural Network (SNNs) is a type of neural network architecture that has two or more subnetworks that are identical. The term ”identical” refers to the fact that they have the same setup, including the same parameters and weights. The modification of parameters is repeated in all sub-networks. These networks are used in a variety of applications to determine the similarity of inputs by matching feature vectors. Generally, a neural network is trained to predict multiple classes. When we need to add or remove new classes in the dataset, this causes an issue. We must upgrade the neural network and retrain it on the whole dataset in this situation. Deep neural networks often require a vast amount of data to learn on and by doing this a lot of time is wasted. In contrast, SNNs learn a similarity function. As a result, we will teach it to detect if the two images are identical (which we will do here). This allows us to identify new types of data images without having to retrain the network.

5.1.4.3 Transfer Learning

Transfer learning is a machine learning technique in which a model is created and trained for a specific task, and its weights and architecture are used as the basis for a model working on a different task. It is very commonly used in deep learning. Generally for more accuracy, the number of layers in a network are increased but with increase in layers, computation and time resources required to train the network also increase drastically. The availability of resources and a huge dataset to train large networks may not be always feasible. However, in transfer learning, we can use a pre-trained model as a basis for the new problem and then fine tune it on a new dataset even with limited resources.

In our case, we have selected ResNet-34 for feature extraction. Since, we do not have the resources to train this network from scratch, we have used the weights of pre-trained ResNet-34 network (trained on ImageNet database) and then fine-tuned it on our database using transfer learning. In one of our experiments, we initially train the model on only UND 2D data and use this as a starting point and fine-tune it on UND 3D (converted to 2.5D data) dataset using transfer learning. It is explained further in Section 5.2 in more detail.

5.2 Experimentation and Results

In our model, we use pre-trained ResNet-34 for feature extraction and Siamese for calculating similarity score using feature vectors. A pair of images is first pre-processed to get the images of size 224×224 pixels and passed onto the network. The ResNet-34 extracts useful features from the images and provides feature vectors. These feature vectors are passed through remaining layers of the Siamese Network and are used to calculate the similarity score between the two input images. Further, the discussion on four different experiments has been provided in following subsections

5.2.1 Experiment-1: Recognition using only 2D Data

After augmentation, there are a total of 5718 images in the dataset. We split this dataset into 3 parts, namely training data that is 70% of the total dataset, validation data that is 15% of the total dataset, and testing data which is 15% of the total dataset. Further, pairs of images are formed and their respective labels are assigned. These are passed to the proposed network and are used to train the network until a satisfactory validation accuracy is achieved. After training, the model and its weights are stored for testing. We get a validation accuracy of 99.05% after training the network for 120 epochs. The testing accuracy for the same is obtained as 98.30%. The graph for validation and training loss *vs.* epoch are shown in Figure 5.5 whereas the graph for validation accuracy *vs.* epoch is shown in Figure 5.6.

5.2.2 Experiment-2: Recognition using only 2.5D Data

In this experiment as well, there are a total of 5718 images in the dataset after augmentation. The dataset is further divided into 3 parts, namely training data that is 70% of the total dataset, validation data that is 15% of the total dataset, and testing data that is 15% of the total dataset. Pairs of images are formed and their respective labels are assigned as done

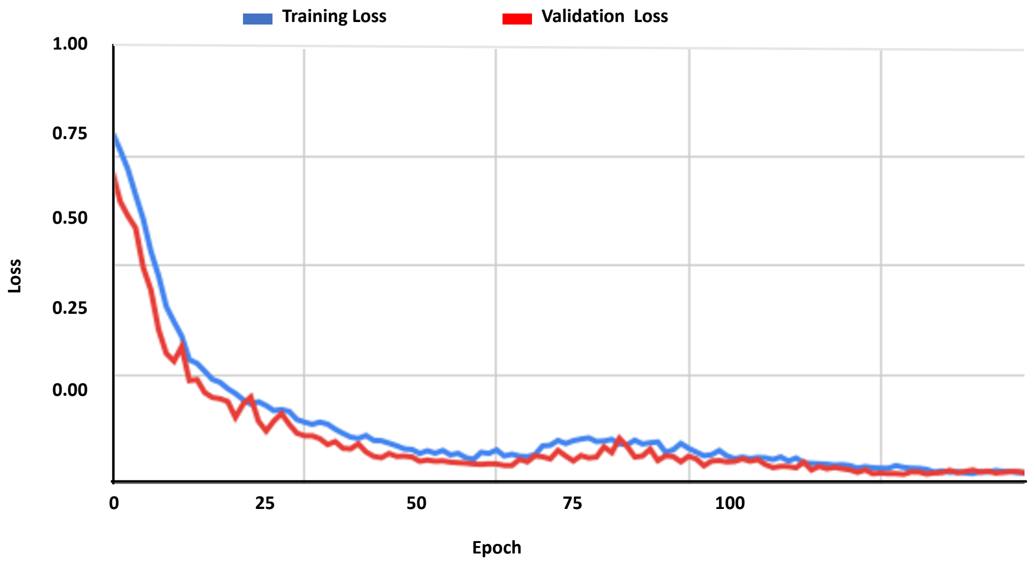


Figure 5.5: Plot of training and validation loss *vs.* epochs when only 2D data is used

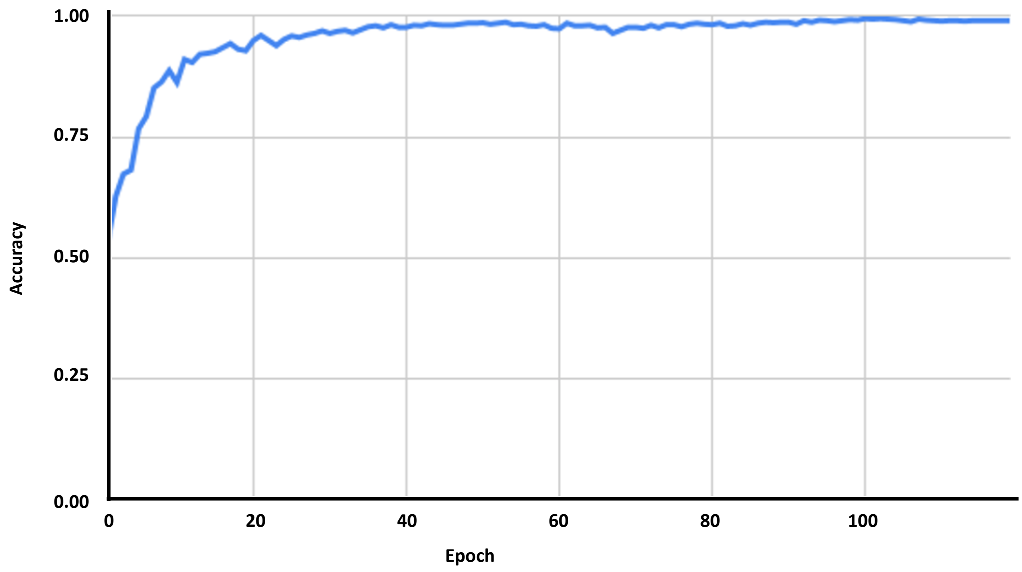


Figure 5.6: Plot of validation accuracy *vs.* epoch when only 2D data is used.

in previous experiment. These are passed to the proposed network and are used to train the network until a satisfactory validation accuracy is achieved. After training the network for 150 epochs, we get a validation accuracy of 99.37%. After training, the model and its weights are stored for testing and the testing accuracy for the same is achieved as 99.10%. The graphs for validation and training loss *vs.* epoch are shown in Figure 5.7 whereas the graph for validation accuracy *vs.* epoch is shown in Figure 5.8.

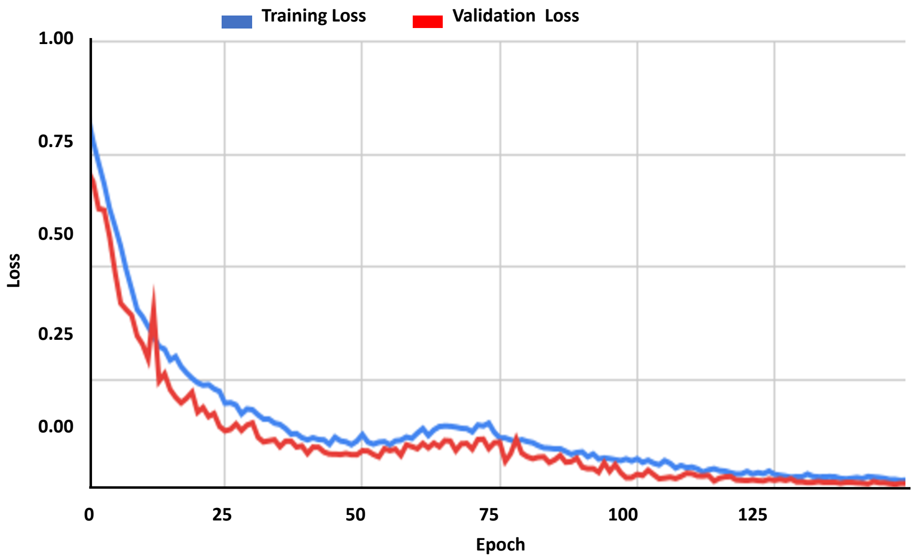


Figure 5.7: Plot of training and validation loss *vs.* epochs when only 2.5D data is used

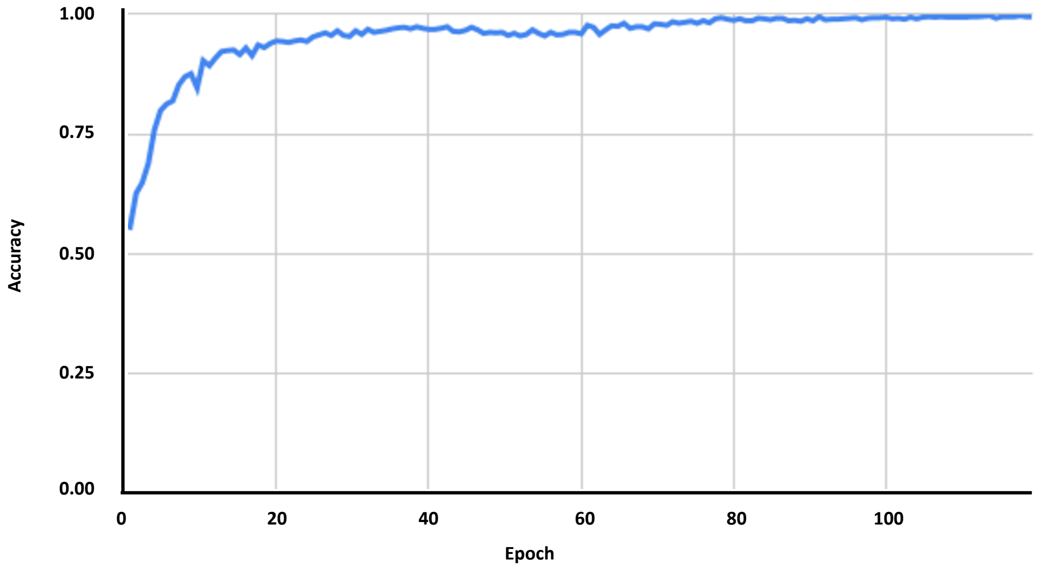


Figure 5.8: Plot of validation accuracy *vs.* epoch when only 2.5D data is used

5.2.3 Experiment-3: Recognition using Transfer Learning Approach (for 2.5D data)

Similar to previous experiments, there are a total of 5718 images in the dataset after performing yhe data augmentation in this experiment as well. The dataset is further divided into 3 parts, namely, training data that is 70% of the total dataset, validation data that is 15% of the total dataset, and testing data that is 15% of the total dataset. Here, we first load the model saved in experiment 1 that is the model trained on UND 2D data (discussed in Section 5.2.1). Further, pairs of 2.5D images are formed and their respective labels are assigned. Afterwards, as followed in earlier experiments, these pairs are passed to the pre-trained network, which is loaded in the previous step and are used to further train it until a satisfactory validation accuracy is achieved. After training, the model and its weights are stored for testing. In this experiment, after training the network for 130 epochs, we get a validation accuracy of 99.68%. The testing accuracy for the same is obtained as 99.24%.

The graph for validation and training loss *vs.* epoch is shown in Figure 5.9 and validation accuracy *vs.* epoch is shown in Figure 5.10. We observe from Figure 5.11 that testing accuracy is almost similar to the one achieved in experiment 2 (when 2.5D data is only used), however, the training converges faster in this case, which saves the training time. This can be observed from Figure 5.11 where the validation accuracy starts at a higher value when transfer learning is used.

5.2.4 Experiment-4: Multimodal Recognition by Combining Results of 2D and 2.5D Models

In this experiment, we consider a fused model composed of the previously pre-trained models on 2D and 2.5D data, and perform 'AND' and 'OR' operations to get the final result of fusion. This model takes 4 images as input, where 2 images are from 2D dataset and

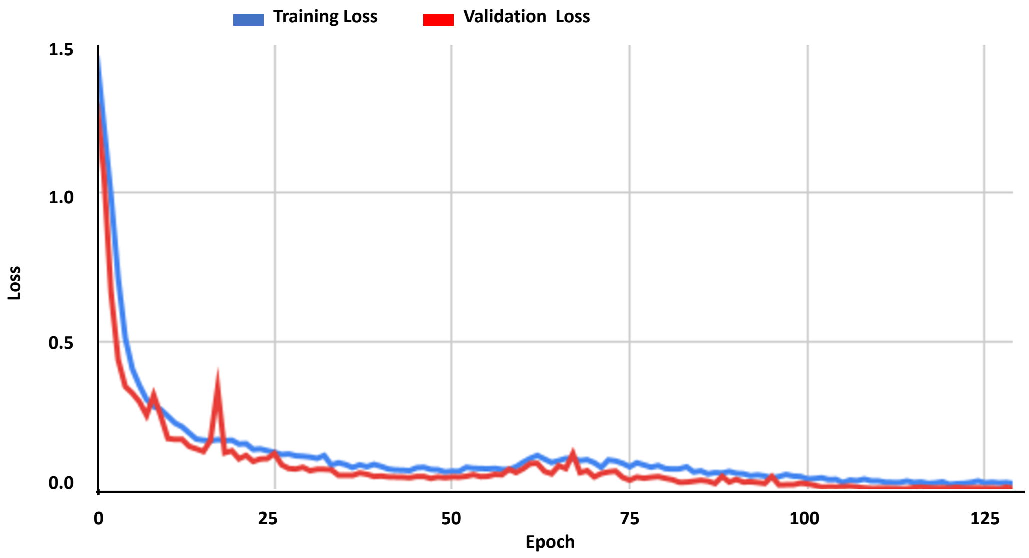


Figure 5.9: Plot of training and validation loss *vs.* epoch when transfer learning approach is used

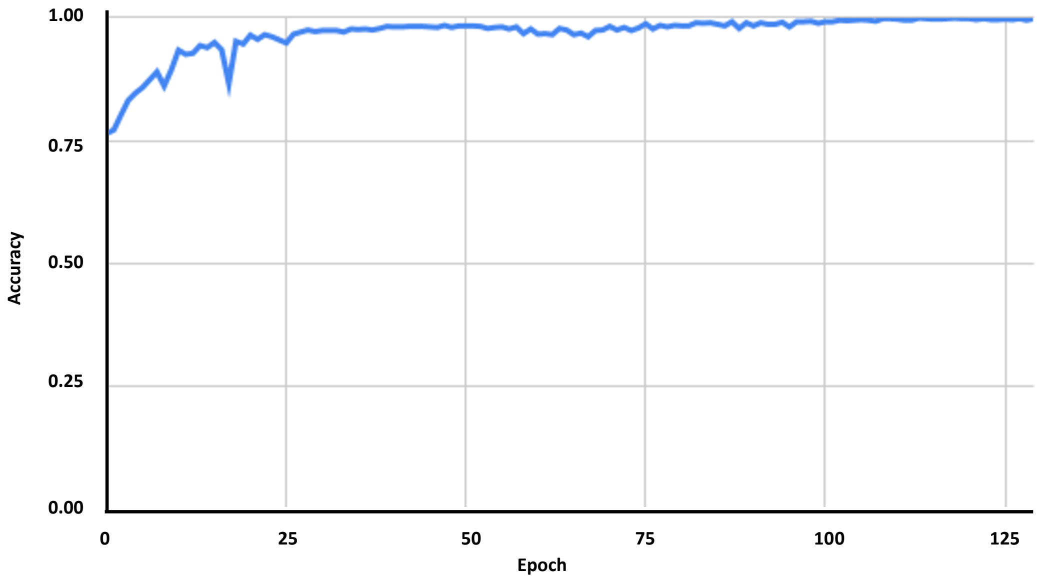


Figure 5.10: Plot of validation accuracy *vs.* epoch when transfer learning approach is used

2 corresponding images are from 2.5D dataset. The 2 images of 2D dataset are passed to the 2D model whereas other 2 images of 2.5D dataset are passed to the 2.5D model. We take outputs from both these models and perform ’AND’ and ’OR’ operations to get a final output as shown in Figure 5.12. The experiment with ’AND’ operator has produced a test

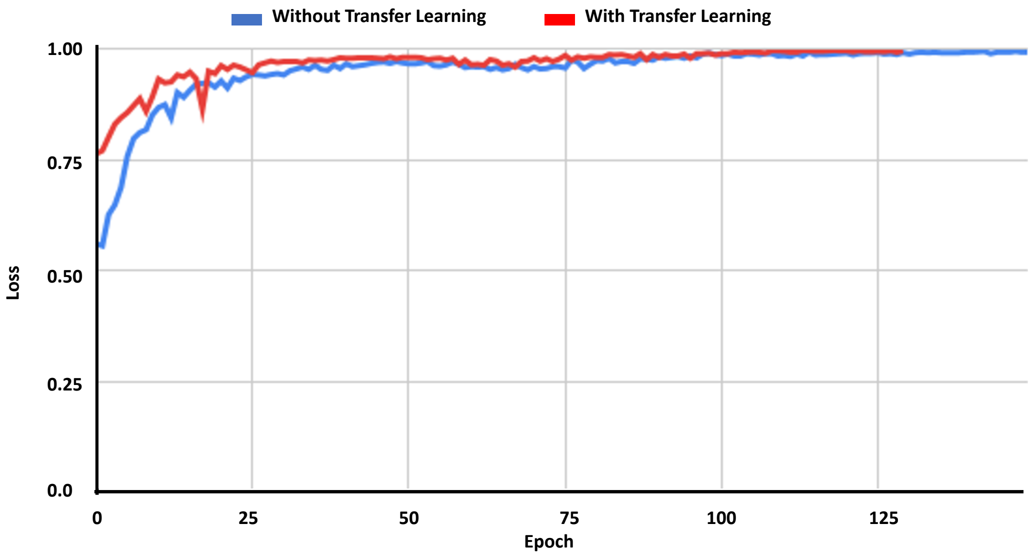


Figure 5.11: Comparison of validation accuracy values obtained using with and without transfer learning

accuracy of 99.49% while the experiment with 'OR' operator has produced a test accuracy as 97.05%. The performance of all the experiments is summarized in Table 5.2.

Table 5.2: Results of the proposed model in terms of validation and testing accuracies for different experiments

Experiment Name	Validation Accuracy	Testing Accuracy
Using 2D Data only	99.05%	98.30%
Using 3D Data only	99.37%	99.15%
Transfer Learning Approach	99.68%	99.24%
Combining 2D and 3D models (OR operator)	NA	97.05%
Combining 2D and 3D models (AND operator)	NA	99.49%

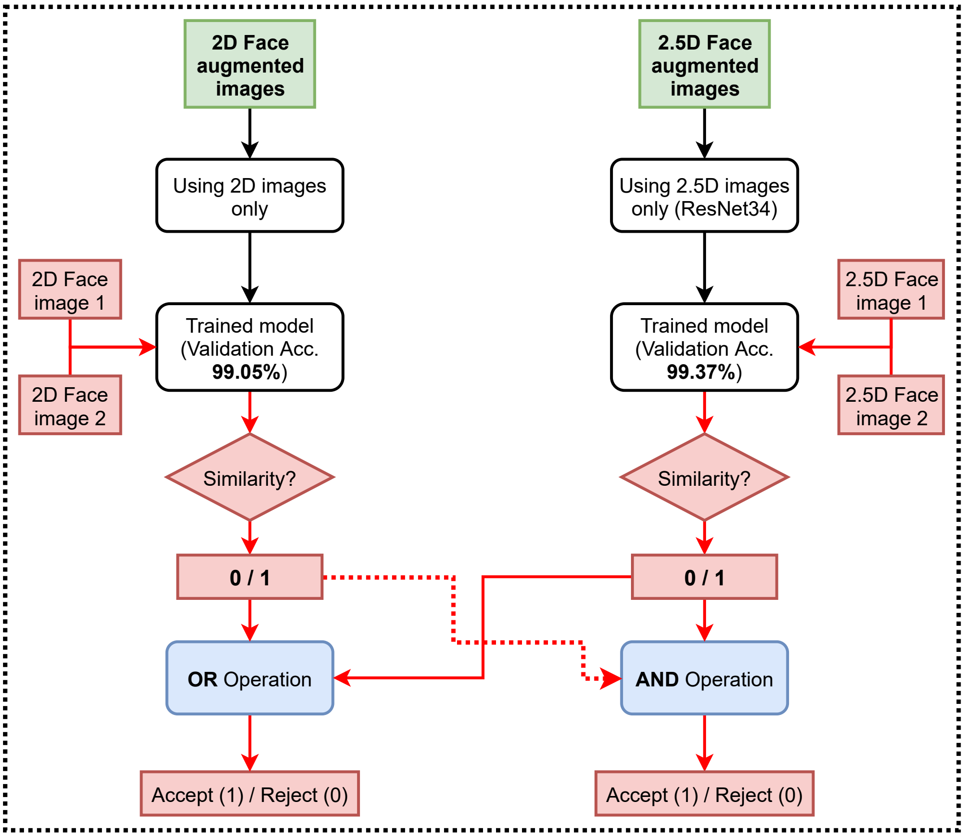
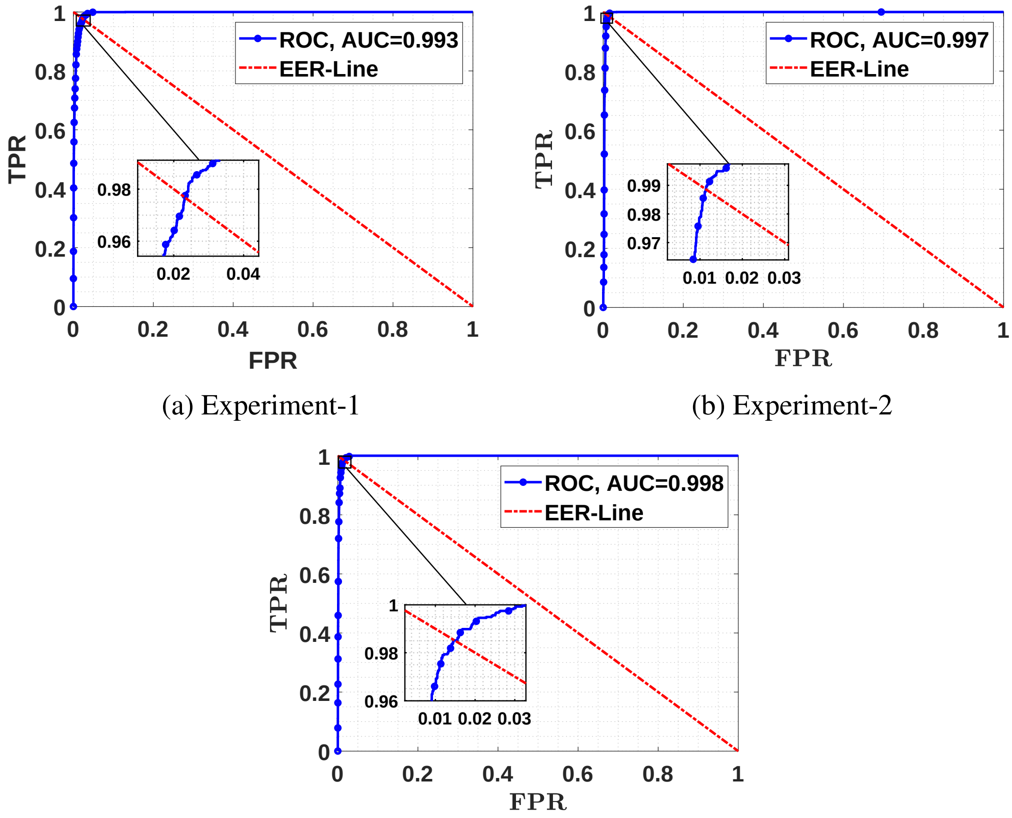


Figure 5.12: Block diagram of the procedure used for combining 2D and 2.5D models

5.2.5 Performance Evaluation in Terms of Some Additional Parameters

The proposed model is further evaluated for the above mentioned experiments on the basis of Rank-1 and Rank-2 accuracies, area under receiver operative characteristics (ROC) curve (AUC), and equal error rate (EER). Rank-*k* accuracy is used to analyse the identification performance of a biometric system. It shows the proportion of times in which the correct sample occurs with in the top-*k* matches. In order to judge the ranking capabilities of an identification system, the cumulative matching characteristic curve (CMC) is used.

Figure 5.13 shows ROC curves for the three experiments. Furthermore, the CMC curves where rank-1 accuracy for the three experiments have been shown are presented in Figure 5.14. In the training, the time taken per epoch is found to be different for all three exper-



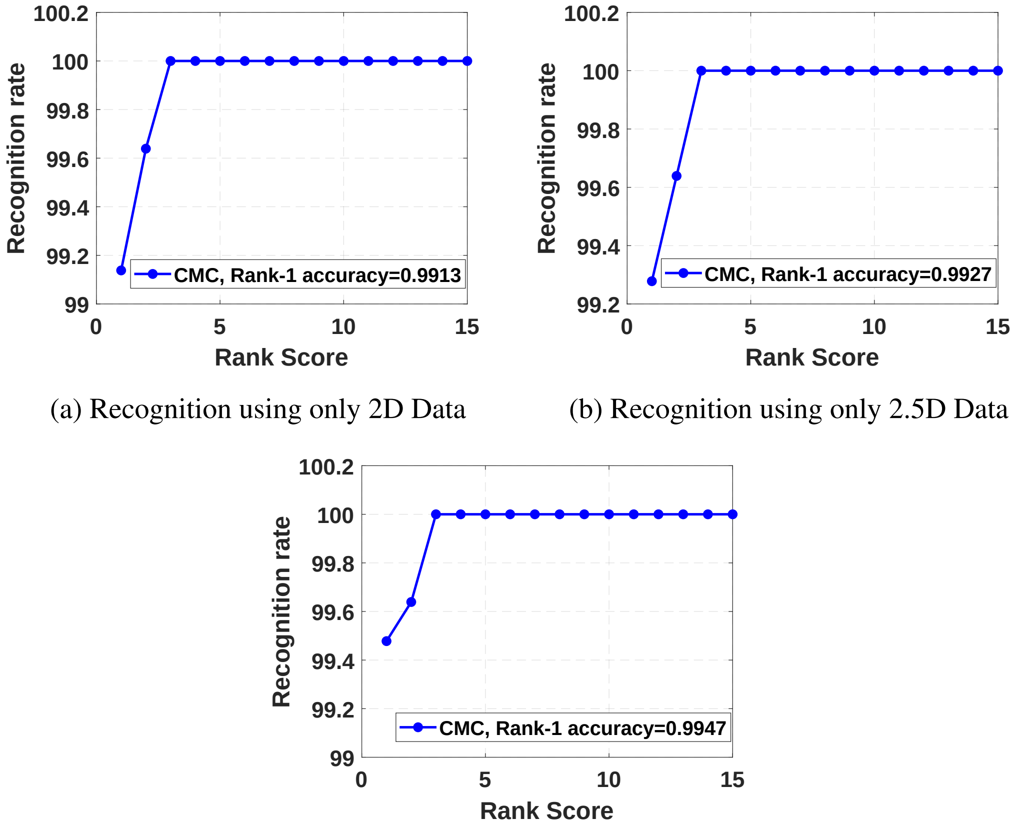
(a) Experiment-1

(b) Experiment-2

(c) Experiment-3

Figure 5.13: The ROC curves for three different experiments using 2D data only, 2.5D data only, and using transfer learning approach, respectively

iments. Table 5.3 shows the Rank-1 accuracy, AUC, EER, and the average training time per epoch for all three experiments. Here, the average training time per epoch with respect to experiment-4 has not been included in the table as this experiment is a fusion of trained 2D and 2.5D models. From the table, it is evident that the best results are obtained when transfer learning is employed. The training time per epoch of transfer learning approach is drastically reduced as compared to the training time per epoch of 2.5D data, and this leads to the reduction in overall computational time and resources. The reason behind this is that when transfer learning is used, only tuning of weights is required instead of performing training process from the scratch.



(a) Recognition using only 2D Data

(b) Recognition using only 2.5D Data

(c) Recognition using transfer learning

Figure 5.14: The CMC curves for three different experiments

Table 5.3: Performance of the proposed model in terms of different evaluation parameters

Experiment Name	Rank-1 Accuracy	Rank-2 Accuracy	AUC	EER	Average Training Time per Epoch
Using 2D Data only	99.13	99.63	99.3%	0.0228	110 sec.
Using 2.5D Data only	99.27	99.63	99.7%	0.0104	117 sec.
Transfer Learning	99.47	99.63	99.8%	0.0138	70 sec.

5.2.6 Comparative Analysis

We compare the performance of the proposed model with the existing 3D face recognition techniques on UND face database and present the comparative analysis in Table 5.4. The comparison has been performed in terms of EER and Rank-1 accuracy. It is clearly evident from the table that the performance of the proposed model is superior to that of the existing techniques. This concludes that the exploitation of the face features in the proposed network is capable of delivering results that are superior to those obtained by the conventional techniques.

Table 5.4: Performance comparison of the proposed network with the state-of-the-art techniques on UND face database

Techniques	EER	Identification rate (Rank-1 accuracy)
Chang et al. (7)	-	87.10%
Haar et. al. (87)	-	98.0%
Berretti et al. (5)	-	82.1%
Srivastava et. al. (83)	0.0080	97.0%
Proposed Model	0.0138	**99.49%**

Note: “-” indicates the non-availability of the data.

5.3 Conclusions

We observe that ResNet-34 acts as a good feature extractor for both 2D and 2.5D images. We also observe that the testing accuracy for 2.5D data is higher when compared to 2D data which proves that the 2.5D face recognition is better than 2D face recognition. In some cases where 3D data is not available, this network provides comparable accuracy for 2D data as well. We observe that the 2D data augmentation to increase the samples in the dataset helps in increasing the training and testing accuracy. Thus, we conclude that face recognition with 3D data helps in eliminating many issues such as poor illumination, shadows, etc. that we face with 2D data.

Chapter 6

Conclusions

Face recognition has attained a lot of relevance in human recognition due to several advantages. It is one of the most non-intrusive biometric techniques that is available currently and is considered as one of the most attractive biometrics. It is the most widely used identifier or token for representing people. Face images can be found in practically all personal documents; this alone demonstrates the importance of the face as a biometric for person recognition. Recognition scenario with face biometrics is broadly classified into two types, $i.e.$, 2D face recognition and 3D face recognition. The 2D face recognition uses 2D grayscale or color images whereas in 3D face recognition, 3D representation of the face is used where surface geometry of the face and geometrical depth information is highlighted.

In 2D face recognition, several challenges are there due to the large intra-class variations and small inter-class variations, which are caused by pose and lighting variations, expression, occlusion, aging, and non-robust representation of face image data. Unlike 2D face recognition, 3D face recognition is robust against pose and lighting variations due to the invariance of the 3D shape capturing procedures against these challenges. While 3D face recognition systems have been developed to overcome pose and illumination problems, many factors such as computational requirement for processing of 3D data, non-availability of large 3D face databases have hindered the practical deployment of 3D face recognition systems. Presently, new directions in 3D face recognition research and applications are being created due to advanced data capture technology and improved processing capacity. The contributions of the thesis in the domain of 3D face recognition are listed in the following section.

6.1 Thesis Contributions

The work presented in the thesis addresses the issues related to the 3D face recognition. It has dealt with the problem of limited 3D data, recognition of similar class objects (such as 3D faces), and reduction of computational requirements in 3D face recognition by making use of 2.5 D data and transfer learning.

In Chapter 3, the main work is related to data augmentation which solves the major issue that is the scarcity of 3D data required for training. The work has proposed efficient data augmentation technique for 3D face data that solves the problem of overfitting in deep neural network models. We have used IIT Indore Phase-3 (IITI) dataset for evaluation purposes in this chapter. The dataset contains challenging samples where many of them are noisy and are not aligned properly. We have augmented these 3D facial samples using three different sampling techniques, *viz.*, random, systematic, and stratified samplings and have compared the results. Regarding this, three experiments have been performed, *viz.*, Intra-sample Registration Error, Inter-sample Registration Error, and Inter-subject Registration Error. The results obtained from all three experiments show three common conclusions. First, the registration error between the original sample and its sub-samples is very similar for all the sub-samples. Second, the registration errors are very close to zero, verifying that all the sub-samples of a particular sample carry the same information. Third, stratified sampling is the best one out of three samplings because of the use of localization in selecting the points.

The first two results prove that the augmented sub-samples carry same information as their parent samples. We have also performed the experiment to check whether the sub-samples have same discriminative power as of their original samples or not. For this, we have computed the average of mean square error (MSE) of CLT mean and CLT standard deviation for different sampling techniques when all 170 subjects of the database are used. The results from this experiment have proved that the sub-samples created from the original sample have the same discriminative power as that in the original sample.

Chapter 4 has presented a generic 3D object recognition technique that is useful for matching of highly similar class of 3D objects. In this work three datasets, *viz.*, IIT Indore (IITI) Phase-3 dataset, Bosphorus dataset, and University of Notre Dame (UND) dataset have been used. In these datasets, number of samples for each subjects are very less and are on an average 3 to 4 per subject, which are not enough for robust training. In the availability of limited data, the model learns the details and noise of these few samples so well that it negatively impacts the testing of the selected model on new data. To avoid over-fitting, we have increased the variability of the 3D data by enlarging the dataset through new data augmentation technique, *i.e.*, type-I augmentation and type-II augmentation, where random points are selected from the point clouds of the available 3D image samples. We have constructed a model for 3D face recognition that improves over the existing PointNet architecture by combining it with the Siamese Network with minimal preprocessing. Experimental results have shown that the proposed model has achieved high accuracy in type-I augmentation.

In chapter 5, a solution to the problem of high computational power and resources required to process 3D data has been presented. As stated above, the 2D face recognition suffers from the problems like variation in pose, illumination, lightning conditions, and occlusion. To mitigate these problems, 3D face recognition is used as it is invariant from aforementioned variations. However, the cost of resources and computational power required to collect and preprocess the 3D face data is exceptionally high. One solution of the problem is that instead of dealing with 3D data directly, convert 3D data into 2.5D (depth images) data which saves the cost in terms of computational power and time required for processing. In the work proposed in Chapter 5, we convert 3D data to 2.5D data to reduce the resource consumption and time requirements, and use it along with 2D data for face recognition. Further, we use data augmentation to increase the size of the dataset to make the training robust. The work has used a pre-trained ResNet-34 architecture for feature extraction and Siamese Network for face matching. The UND 2D and 3D face datasets have been taken for experimentation. To evaluate the proposed technique, four experiments have been performed. In experiment-1, face recognition is performed using ResNet-34 that is trained on 2D face images only. In experiment-2, 2.5D face images are only used to train ResNet-34 for face recognition. In experiment-3, face recognition is performed using transfer learning where pre-trained model of 2D face images is utilized and then fine tune using 2.5D face images. In experiment 4, we used the fusion of the results of 2D and 2.5D models. Experimental results obtained in all three experiments have shown that ResNet-34 acts as a good feature extractor for both 2D images and 2.5D converted depth images, and the the testing accuracy for 2.5D data is higher when compared to 2D data. This proves that 3D face recognition is better than 2D face recognition. In the Experiment-3, where transfer learning with 2.5D data is used on the pre-trained model of 2D data, testing accuracy is almost similar to the one obtained in the Experiment-2; however, the training converges faster when compared to the case when 2.5D data is only used , thus saving time during training. It is also observed that the validation accuracy starts at a higher value when transfer learning is being used. This concludes that the face recognition with 3D data helps in eliminating many issues that are faced working with 2D data.

6.2 Further Research Directions

A few techniques related to 3D face recognition have been proposed in this thesis. There are several interesting and important future directions which can be taken up further.

- **Exploration of other Data augmentation Techniques** The data augmentation techniques proposed in this thesis utilize point cloud representation that use sub-sampling from the existing point clouds to increase the size and variability of the available data. The other data augmentation techniques including the generic and face specific transformations can also be utilized for producing the augmented samples. Hairstyle transfer, facial makeup transfer, and, accessory removal and wearing are some other transformations that can be used in data augmentation. Moreover, the techniques like neural style transfer (109) have potential to be used as a data augmentation method to add more variation to the training dataset in the new domain for face recognition.

- **Fusion of Networks** By combining networks used for 2D and 2.5D data, we can further increase the accuracy of the model. In this work, we have converted the 3D data into 2.5D data (depth images) due to which some information is lost. So, implementing a model which can accept 3D point cloud data directly can further improve the results.

- **Conjunction of Face with other Biometrics Modalities** There are many possibilities of fusing other modalities with face. In order to enhance the recognition performance, more studies need to be carried out on utilising the face recognition in conjunction with other biometrics such as iris, fingerprint, speech, and ear recognition. In addition, the proposed techniques can be further extended for a multi-modal authentication scenario.

Bibliography

[1] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2D and 3D face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885–1906, 2007.

[2] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 581–588 vol. 1, 2005.

[3] U. Asif, M. Bennamoun, and F. A. Sohel. A multi-modal, discriminative and spatially invariant CNN for RGB-D object labeling. *IEEE Trans Pattern Anal Mach Intell*, 40(9):2051–2065, 2018.

[4] M. Aubry, U. Schlickewei, and D. Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1626–1633, 2011.

[5] S. Berretti, A. del Bimbo, and P. Pala. Sparse matching of salient facial curves for recognition of 3-d faces with missing parts. *IEEE Transactions on Information Forensics and Security*, 8(2):374–389, 2013.

[6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH '99*, 1999.

[7] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D

and multi-modal 3D+ 2D face recognition. *Computer vision and image understanding*, 101(1):1–15, 2006.

[8] S. Braeger and H. Foroosh. Curvature Augmented Deep Learning for 3D Object Recognition. In *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3648–3652, 2018.

[9] A. M. Bronstein, M. M. Bronstein, L. J. Guibas, and M. Ovsjanikov. Shape Google: Geometric Words and Expressions for Invariant Shape Retrieval. *ACM Trans. Graph.*, 30(1), feb 2011.

[10] M. M. Bronstein and I. Kokkinos. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1704–1711, 2010.

[11] A. Caglayan and A. B. Can. 3D convolutional object recognition using volumetric representations of depth data. In *Proceedings of the 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 125–128, 2017.

[12] Y. Cai, Y. Lei, M. Yang, Z. You, and S. Shan. A fast and robust 3D face recognition approach based on deeply learned face representation. *Neurocomputing*, 363:375–397, 2019.

[13] E. Cengıl and A. Çinar. Multiple classification of flower images using transfer learning. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–6, 2019.

[14] K. Chang, K. Bowyer, , and P. Flynn. Face recognition using 2D and 3D facial data. In *In: Proceedings of the ACM workshop on multimodal user authentication.*, page 1–6. 2003.

[15] D. Chetverikov, D. Stepanov, and P. Krsek. Robust euclidean alignment of 3d point sets: The trimmed iterative closest point algorithm. *Image and Vision Computing;*,

23(3):299–309, 2005.

[16] C. S. Chua, F. Han, and Y. K. Ho. 3D human face recognition using point signature. In *Proc. of 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 233–238, 2000.

[17] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 1–8, 2017.

[18] X. Deng, F. Da, and H. Shao. Efficient 3D face recognition using local covariance descriptor and Riemannian kernel sparse coding. *Computers Electrical Engineering*, 62:81–91, 2017.

[19] F. G. Donoso, A. G. Garcia, J. G. Rodriguez, S. O. Escolano, and M. Cazorla. Lonchanet: a sliced-based CNN architecture for real-time 3D object recognition. In *Proceedings of the 2017 international joint conference on neural networks (IJCNN)*, page 412–418. 2017.

[20] A. F. Elaraby, A. Hamdy, and M. Rehan. A Kinect-Based 3D Object Detection and Recognition System with Enhanced Depth Estimation Algorithm. In *Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 247–252, 2018.

[21] P. J. Flynn, K. W. Bowyer, and P. J. Phillips. Assessment of time dependency in face recognition. In *In: Proceedings of the audio- and video-based biometric person authentication.*, page 44–51. 2003.

[22] T. Funkhouser and P. Shilane. Partial Matching of 3D Shapes with Priority-Driven Search. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, page 131–142, 2006.

[23] I. I. Ganapathi, S. Prakash, I. R. Dave, P. Joshi, S. S. Ali, and A. M. Shrivastava.

Ear recognition in 3D using 2D curvilinear features. *IET Biometrics;*, 7(6):519–529, 2018.

[24] L. Ge, Y. Cai, J. Weng, and J. Yuan. Hand PointNet: 3D Hand Pose Estimation Using Point Sets. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8417–8426, 2018.

[25] S. Z. Gilani and A. Mian. Learning from Millions of 3D Scans for Large-Scale 3D Face Recognition. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1896–1905, 2018.

[26] G. G. Gordon. Face recognition based on depth maps and surface curvature. In B. C. Vemuri, editor, *Geometric Methods in Computer Vision*, volume 1570, pages 234 – 247. International Society for Optics and Photonics, SPIE, 1991.

[27] G. G. Gordon. Face Recognition from Frontal and Profile Views. In *Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition, IWAFGR 95*, pages 47–52, 1995.

[28] B. Graham. Sparse arrays of signatures for online character recognition. *ArXiv;*, abs/1308.0371, 2013.

[29] K. Guo, D. Zou, and X. Chen. 3D Mesh Labeling via Deep Convolutional Neural Networks. *ACM Transactions on Graphics;*, 35(1):3:1–3:12, 2015.

[30] W. Hariri, H. Tabia, N. Farah, A. Benouareth, and D. Declercq. 3D facial expression recognition using kernel methods on Riemannian manifold. *Engineering Applications of Artificial Intelligence*, 64:25–32, 2017.

[31] W. Hayale, P. Negi, and M. Mahoor. Facial expression recognition using deep siamese neural networks with a supervised loss function. In *Proceedings of the 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–7, 2019.

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

[33] M. He, B. Li, and H. Chen. Multi-scale 3D deep convolutional neural network for hyperspectral image classification. In *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*, pages 3904–3908, 2017.

[34] C. Hesher, A. Srivastava, and G. Erlebacher. A novel technique for face recognition using range imaging. In *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.*, volume 2, pages 201–204 vol.2, 2003.

[35] C. Heyde. *Wiley StatsRef: Statistics Reference Online*, chapter Central Limit Theorem. 2014.

[36] S. Hinterstoißer, V. Lepetit, P. Wohlhart, and K. Konolige. On Pre-Trained Image Features and Synthetic Images for Deep Learning. *ArXiv*, abs/1710.10710, 2018.

[37] H. Hu, S. A. A. Shah, M. Bennamoun, and M. Molton. 2D and 3D face recognition using convolutional neural network. In *Proceedings of the TENCON 2017 - 2017 IEEE Region 10 Conference*, pages 133–132, 2017.

[38] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang. Automatic 3D reconstruction for face recognition. *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 843–848, 2004.

[39] D. Huang, G. Zhang, M. Ardabilian, Y. Wang, and L. Chen. 3D Face recognition using distinctiveness enhanced facial representations and local feature hybrid matching. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–7, 2010.

[40] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Pro-*

ceedings of Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, 2008.

[41] A. Ioannidou, E. Chatzilari, S. Nikolopoulos, and I. Kompatsiaris. Deep Learning Advances in Computer Vision with 3D Data: A Survey. *ACM Computing Surveys;*, 50(2):20:1–20:38, 2017.

[42] M. Iwasaki and R. Yoshioka. Data Augmentation Based on 3D Model Data for Machine Learning. In *Proceedings of the 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pages 1–4, 2019.

[43] A. K. Jain, A. A. Ross, and K. Nandakumar. *Introduction to Biometrics.* Springer New York Dordrecht Heidelberg London, 2011.

[44] J. C. Joshi, A. Gupta, K. Tiwari, and K. K. Gupta. Scanned to Digital Face Images Matching With Siamese Network. In *Proceedings of the 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–6, 2018.

[45] D. Kim, M. Hernandez, J. Choi, and G. Medioni. Deep 3D face identification. In *Proceedings of the 2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 133–142, 2017.

[46] R. S. Kute, V. Vyas, and A. Anuse. Component-based face recognition under transfer learning for forensic applications. *Information Sciences*, 476:176–191, 2019.

[47] S. Lawrence, C. Giles, A. C. Tsoi, and A. Back. Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, 1997.

[48] K. Lee, J. Ho, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:I/313–I/320, 2003.

[49] Y. Lei, Y. Guo, M. Hayat, M. Bennamoun, and X. Zhou. A Two-Phase Weighted Collaborative Representation for 3D partial face recognition with single sample. *Pattern Recognition;*, 52:218–237, 2016.

[50] Y. Lei, Y. Guo, M. Hayat, M. Bennamoun, and X. Zhou. A Two-Phase Weighted Collaborative Representation for 3D partial face recognition with single sample. *Pattern Recognition*, 52:218–237, 2016.

[51] B. Leng, K. Yu, and J. QIN. Data augmentation for unbalanced face recognition training sets. *Neurocomputing*, 235:10–14, 2017.

[52] H. Li, D. Huang, P. Lemaire, J.-M. Morvan, and L. Chen. Expression robust 3D face recognition via mesh-based histograms of multiple order surface differential quantities. In *2011 18th IEEE International Conference on Image Processing*, pages 3053–3056. IEEE, 2011.

[53] H. Li, D. Huang, J.-M. Morvan, Y. Wang, and L. Chen. Towards 3D Face Recognition in the Real: A Registration-Free Approach Using Fine-Grained Matching of 3D Keypoint Descriptors. *International Journal of Computer Vision;*, 113(2):128–142, 2015.

[54] S. Li and H. Feng. EEG Signal Classification Method Based on Feature Priority Analysis and CNN. In *Proceedings of the 2019 International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 403–406, 2019.

[55] Z. Li, H. Zou, X. Sun, T. Zhu, and C. Ni. 3D Expression-Invariant Face Verification Based on Transfer Learning and Siamese Network for Small Sample Size. *Electronics*, 10(17), 2021.

[56] J. Luo, F. Hu, and R. Wang. 3D Face Recognition Based on Deep Learning. In *Proceedings of the 2019 IEEE International Conference on Mechatronics and Au-*

tomation (ICMA), pages 1576–1581, 2019.

[57] J. Luttrell, Z. Zhou, C. Zhang, P. Gong, and Y. Zhang. Facial Recognition via Transfer Learning: Fine-Tuning Keras_vggface. In *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 576–579, 2017.

[58] J. J. Lv, X. H. Shao, J. S. Huang, X. D. Zhou, and X. Zhou. Data augmentation for face recognition. *Neurocomputing*, 230:184–196, 2017.

[59] A. M. Martinez. Recognition of partially occluded and/or imprecisely localized faces using a probabilistic approach. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, 1:712–717 vol.1, 2000.

[60] D. Maturana and S. A. Scherer. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928, 2015.

[61] A. B. Moreno, A. Sánchez, J. F. Vélez, and F. J. Díaz. Face recognition using 3D surface-extracted descriptors. In *Irish Machine Vision and Image Processing Conference*, volume 2. Citeseer, 2003.

[62] J. Neves and H. Proença. “a leopard cannot change its spots”: Improving face recognition using 3d-based caricatures. *IEEE Transactions on Information Forensics and Security*, 14(1):151–161, 2019.

[63] M. Oberweger and V. Lepetit. DeepPrior++: Improving Fast and Accurate 3D Hand Pose Estimation. In *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 585–594, 2017.

[64] G. Pan, Z. Wu, and Y. Pan. Automatic 3D face verification from range data. In *2003 International Conference on Multimedia and Expo. ICME’03. Proceedings (Cat. No. 03TH8698)*, volume 3, pages III–133. IEEE, 2003.

[65] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowl-*

edge and Data Engineering, 22(10):1345–1359, 2010.

[66] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12, 2015.

[67] H. Patil, A. Kothari, and K. Bhurchandi. 3-D face recognition: Features, databases, algorithms and challenges. *Artificial Intelligence Review;*, 44:393–441, 2015.

[68] S. Porcu, A. Floris, and L. Atzori. Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems. *Electronics*, 9(11), 2020.

[69] J. Procházková and D. Martišek. Notes on iterative closest point algorithm. In *Proceedings of the 17th Conference on Applied Mathematics (APLIMAT 2018)*, 2018.

[70] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2016.

[71] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas. Volumetric and Multi-view CNNs for Object Classification on 3D Data. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5648–5656, 2016.

[72] R.Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Matching 3D models with shape distributions. In *Proceedings International Conference on Shape Modeling and Applications*, pages 154–166, 2001.

[73] A. Rosebrock. What is face recognition? `https://www.pyimagesearch.com/2021/05/01/what-is-face-recognition`, 2021.

[74] R. B. Rusu, N. Blodow, and M. Beetz. Fast Point Feature Histograms (FPFH) for 3D registration. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217, 2009.

[75] D. Sales, J. Amaro, and F. Osorio. 3D shape descriptor for objects recognition.

In *Proceedings of the ´2017 Latin American robotics symposium (LARS) and 2017 Brazilian symposium on robotics (SBR).*, page 1–6. 2017.

[76] C. Samir, A. Srivastava, and M. Daoudi. Three-dimensional face recognition using shapes of facial curves. *IEEE transactions on pattern analysis and machine intelligence*, 28(11):1858–1863, 2006.

[77] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[78] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *Proceedings of the workshop on biometrics and identity management.*, page 47–56. 2008.

[79] L. Shi, X. Wang, and Y. Shen. Research on 3D face recognition method based on LBP and SVM. *Optik*, 220:165157, 2020.

[80] D. Smeets, J. Keustermans, D. Vandermeulen, and P. Suetens. meshSIFT: Local surface features for 3D face recognition under expression variations and partial data. *Computer Vision and Image Understanding*, 117(2):158–169, 2013.

[81] S. Soltanpour and Q. J. Wu. Multimodal 2D–3D face recognition using local descriptors: pyramidal shape map and structural context. *IET Biometrics*, 6(1):27–35, 2016.

[82] Song, Hwanjong, U. Yang, Sohn, and Kwanghoon. 3D face recognition under pose varying environments. In *International Workshop on Information Security Applications*, pages 333–347. Springer, 2003.

[83] A. M. Srivastava, A. Jain, P. Rotte, S. Prakash, and U. Jayaraman. A technique to match highly similar 3D objects with an application to biomedical security. *Multimedia Tools and Applications*, 2021.

[84] S. Taertulakarn, C. Pintavirooj, P. Tosranon, and K. Hamamoto. The preliminary investigation of ear recognition using hybrid technique. In *Proceedings of the 2016 9th Biomedical Engineering International Conference (BMEiCON)*, pages 1–4, 2016.

[85] H. Taherdoost. Sampling methods in research methodology; how to choose a sampling technique for research. *International Journal of Academic Research in Management (IJARM);*, 5(2):18–27, 2016.

[86] H. Tanaka, M. Ikeda, and H. Chiaki. Curvature-based face surface recognition using spherical correlation: Principal directions for curved object recognition. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 372–377, 1998.

[87] F. B. ter Haar and R. C. Veltkamp. Expression modeling for expression-invariant face recognition. *Computers Graphics*, 34:231–241, 2010.

[88] T. Terada, Y. Chen, and R. Kimura. 3D Facial Landmark Detection Using Deep Convolutional Neural Networks. In *Proceedings of the 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 390–393, 2018.

[89] V. Uchôa, K. Aires, R. Veras, A. Paiva, and L. Britto. Data Augmentation for Face Recognition with CNN Transfer Learning. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 143–148, 2020.

[90] G. Vishnuvardhan and V. Ravi. *Face Recognition Using Transfer Learning on Facenet: Application to Banking Operations*, pages 301–309. Springer International Publishing, Cham, 2021.

[91] F. Wang and Z. Zhao. A survey of iterative closest point algorithm. In *Proceedings of the 2017 Chinese Automation Congress (CAC)*, pages 4395–4399, 2017.

[92] K.-C. Wong, W.-Y. Lin, Y. H. Hu, N. Boston, and X. Zhang. Optimal linear combi-

nation of facial regions for improving identification performance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(5):1138–1148, 2007.

[93] Wu, Yijun, Pan, Gang, and Zhaohui. Face Authentication Based on Multiple Profiles Extracted from Range Data. In J. Kittler and M. S. Nixon, editors, *Audio- and Video-Based Biometric Person Authentication*, pages 515–522, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[94] H. Wu, Z. Xu, J. Zhang, W. Yan, and X. Ma. Face recognition based on convolution Siamese networks. In *Proceedings of the 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5, 2017.

[95] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015.

[96] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, and J. Yuan. A2J: Anchor-to-Joint Regression Network for 3D Articulated Pose Estimation From a Single Depth Image. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 793–802, 2019.

[97] C. Xu, Y. Wang, T. Tan, and L. Quan. Depth vs. intensity: Which is more important for face recognition? In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 342–345. IEEE, 2004.

[98] K. Xu, X. Wang, Z. Hu, and Z. Zhang. 3D Face Recognition Based on Twin Neural Network Combining Deep Map and Texture. In *Proceedings of the 2019 IEEE 19th International Conference on Communication Technology (ICCT)*, pages 1665–1668, 2019.

[99] L. Yang, S. Li, D. Lee, and A. Yao. Aligning Latent Spaces for 3D Hand Pose Estima-

tion. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2335–2343, 2019.

[100] Y. Yu, F. Da, and Y. Guo. Sparse ICP With Resampling and Denoising for 3D Face Verification. *IEEE Transactions on Information Forensics and Security*, 14(7):1917–1927, 2019.

[101] W. Yun, J. Lee, J. Lee, and J. Kim. Object recognition and pose estimation for modular manipulation system: Overview and initial results. In *Proceedings of the 2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 198–201, 2017.

[102] D. Zeng, L. Spreeuwers, R. Veldhuis, and Q. Zhao. Combined training strategy for low-resolution face recognition with limited application-specific data. *IET Image Processing*, 13:1790–1796(6), 2019.

[103] X. Zhang and Y. Gao. Face recognition across pose: A review. *Pattern recognition*, 42(11):2876–2896, 2009.

[104] Y. Zhang, Z. Lu, J. Xue, and Q. Liao. A New Rotation-Invariant Deep Network for 3D Object Recognition. In *Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1606–1611, 2019.

[105] Y. Zhang, K. Shang, J. Wang, N. Li, and M. M. Zhang. Patch strategy for deep face recognition. *IET Image Processing*, 12:819–825(6), 2018.

[106] Z. Zhang, S.-P. Xie, M. Chen, and H. Zhu. HandAugment: A Simple Data Augmentation Method for Depth-Based 3D Hand Pose Estimation. *ArXiv*, abs/2001.00702, 2020.

[107] H. Zhao, Q. Liu, and Y. Yang. Transfer learning with ensemble of multiple feature representations. In *2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 54–61, 2018.

[108] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face Recognition: A Literature Survey. *ACM Comput. Surv.*, 35(4):399–458, dec 2003.

[109] X. Zheng, T. Chalasani, K. Ghosal, S. Lutz, and A. Smolic. Stada: Style transfer as data augmentation. *CoRR*, abs/1909.01056, 2019.

[110] H. Zhou, A. Mian, L. Wei, D. Creighton, M. Hossny, and S. Nahavandi. Recent advances on singlemodal and multimodal face recognition: A survey. *IEEE Transactions on Human-Machine Systems*, 44(6):701–716, 2014.