# 3D Reconstruction of the Human Face using Multi-stereo Camera Network

Vikram Sandu

A thesis Submitted to

Indian Institute of Technology Hyderabad

In Partial Fulfillment of the Requirements for

The Degree of Master of Technology

भारतीय प्रौद्योगिकी संस्थान हैदराबाद

Indian Institute of Technology Hyderabad

Department of Electrical Engineering

June 2019

# Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

Vikram

(Signature)

(Vikram Sandu)

EE17MTECH11013

(Roll No.)

# Approval Sheet

This thesis entitled 3D Reconstruction of the Human Face using Multi-stereo Camera Network by Vikram Sandu is approved for the degree of Master of Technology from IIT Hyderabad.

(Dr. C. S. Sastry) Examiner
Dept. of Mathematics
IITH

(Dr. Asudeb Dutta) Examiner
Dept. of Electrical Engineering
IITH

(Dr. Soumya Jana) Adviser
Dept. of Electrical Engineering
IITH

(Dr. C. S. Sastry) Chairman
Dept. of Mathematics
IITH

# Acknowledgements

I thank each and every person who supported me throughout this course work. I would like to thank my Supervisor, Dr. Soumya Jana for his meaningful assistance, guidance, and patience. I would also like to thank IMT lab members, especially Roopak Tamboli, Shanmukh Reddy, Sandeep Chandra, and Rahul. I would also like to thank my juniors, Sreenivasulu, and Ramu for their assistance and help during my M.tech Project.

# Dedication

To my family and to everyone who has been part of my learning experiences.

# Abstract

This thesis presents a method to reconstruct a human face in 3D from its multiple 2D images acquired using multiple cameras arranged in a network. The image acquisition setup consists of three stereo pairs of cameras and two off-the-shelf projectors. Since a human face lacks many trackable features, a single structured light pattern was projected on the face to artificially create sufficient number of feature points. The image acquisition was instantaneous, whereas the 3D reconstruction — a computation-heavy process, involving minimization of sum of squared reprojection error — was performed offline. Minimization of the underlying objective function, known to possess large number of local minima, was carried out using genetic algorithm. Furthermore, The cost function of the camera network was defined in a way that allows us to get the 3D information over a large part of the periphery of the face. Our method achieved true-to-scale reconstruction with satisfactory accuracy, as revealed by measurements of certain key facial features. Furthermore, inspection of the reconstructed point cloud from different perspectives provided visual confirmation that shape was also preserved.

# Contents

# Chapter 1

# Introduction

Patient-centric healthcare enables rapid recovery from as well as prevention of health disorders. Scarcity of medical and allied resources hinder achieving this goal [1]. In general, healthcare services in urban areas are often burdened due to low doctor-to-patient ratio, prohibitively high costs, large waiting times etc. On the other hand, underdeveloped remote areas face a different set of challenges such as poor infrastructure (e.g. road, electricity and medical facilities), unavailability of trained medical professionals and lack of awareness among the patients [2]. As a result, many life-threatening communicable and non-communicable disorders remain untreated, claiming lives.

## 1.1 Teleophthalmology: Engineering Perspectives

Telemedicine is a promising solution in both of the aforementioned scenarios [2]. In addition to curative care, preventive measures can also be taken via telemedicine. Pertaining to eyecare, the telemedicine solution — known as teleophthalmology — is often realized as a pyramid structure, depicted in Figure 1.1. In this figure, the map shows the eye care network of the L. V. Prasad Eye Institute which has been established in the Indian states of Andhra Pradesh, Karnataka, Orissa and Telangana. Each state is coloured with a separate color where hues are indicative of the per capita income of the respective state [3, 4, 5, 6]. Darker hues indicate higher income and vice versa. Since the data were not collected for the same year across the state, one can only compare the districts of the same state based on the hues.

Typically, at the base of the pyramid are vision guardians that represent community involvement. Vision centers form the next level and serve the primary eye health needs of the community. Secondary eye care centers provide care that can diagnose the complete range of ophthalmological diseases and offer high quality surgical care. Tertiary centres provide a comprehensive range of services and also serve as training centres to the secondary centres. Centers of excellence treat complex diseases, provide training in subspecialties and rehabilitation and engage in advocacy [7].

In this paradigm of teleophthalmology, the primary care centres at the bottom of the teleopthalmology pyramid usually involve basic imaging equipment at a remote establishment. Images acquired at such centers are sent to the secondary or tertiary centers where doctors make a judgment based on visual inspection and devise a treatment plan. At the remote primary center, the patients are treated according to the plan and follow-up visits are made. Such a setting is useful for basic di-

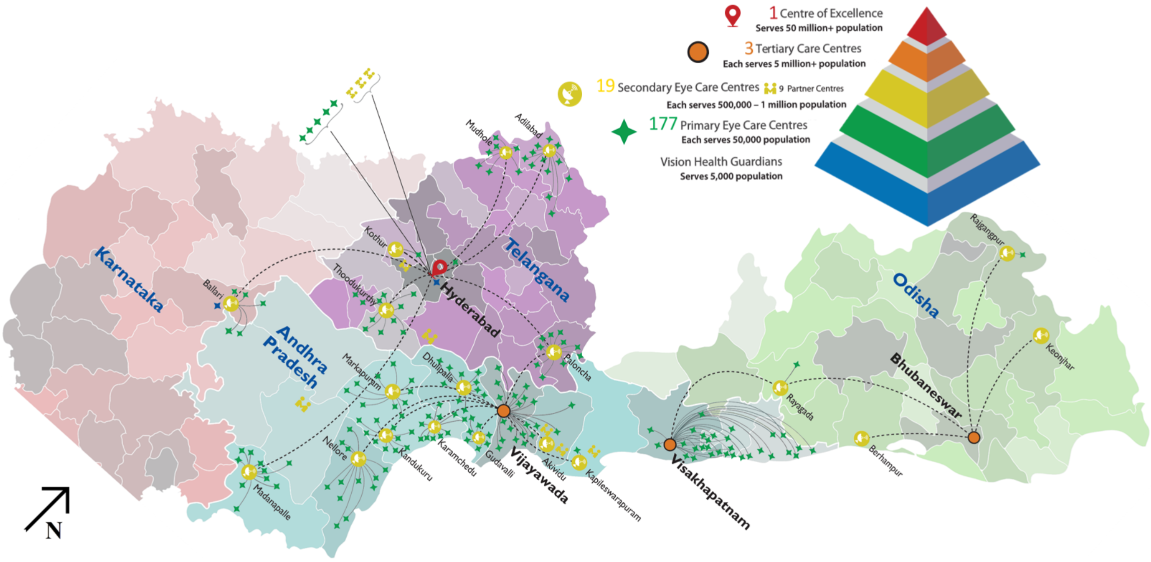
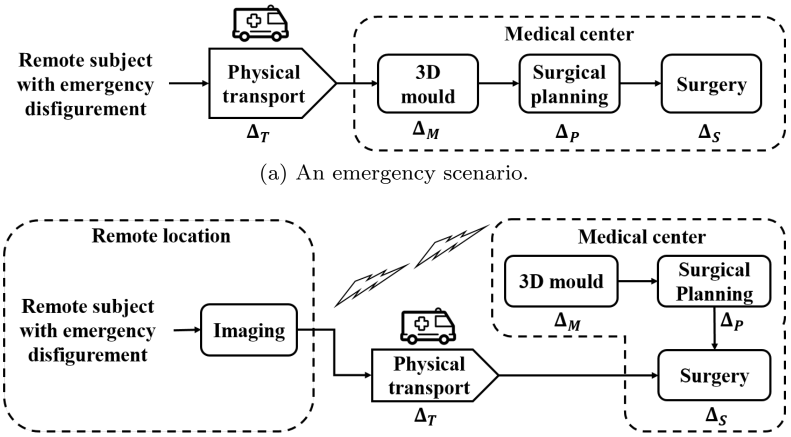


Figure 1.1: The L. V. Prasad Eye Care Network. Courtesy: L. V. Prasad Eye Institute, Hyderabad [7].

agnostics as well as pre-operative checkups for planned surgeries. This conventional paradigm is depicted in Figure 1.2(a).

In emergency situations such as facial trauma, the patient needs to travel to a nearby city where reconstructive facial surgery is performed. Usually — as depicted in Figure 1.2(b) — the subject needs to travel for a few hours, undergo some initial checkups, some of which help preparation of surgical material such as 3D moulds, wait until the required surgery is planned by experts, and then finally get operated. Furthermore, the patient has to stay in the hospital for post-operative assessment. Depending upon the financial situation of the subject, the time and money spent in the whole process could be devastating. Even if the surgery is paid for, the time lost in the process directly translates to lost income of the subject, a catastrophic situation if the subject's earning comes from daily or weekly wages. Further, it may be difficult to gather all the required experts multiple times during various stages of the proposed surgery. Clearly, a primary healthcare center cannot help in such a situation.

In this backdrop, we propose a hybrid teleophthalmology paradigm. In this proposal, the subject under emergency situation still has to travel to the secondary or tertiary center for surgery. However, certain amount of time in pre- and post-operative phases of surgery could be saved if the primary centers are equipped with an imaging system that can capture oculofacial information. As depicted in Figure 1.2(c), a subject under emergency situation is taken to primary center where images are acquired. While the patient travels to the city for surgery, the information acquired earlier can be transmitted to the hospital. At the hospital, key facial measurements can be made using those images. A 3D model reconstructed from this multiple views can be instrumental for planning the surgery. In this way, the planning phase of the surgery can start even before the subject reaches hospital, saving a considerable amount of time. Furthermore, during the post-operative phase and further examinations, the patient can just visit primary center capable of acquiring oculofacial information. In order to succeed, the imaging system to be deployed at the remote



(a) An emergency scenario.

(b) Proposed hybrid teleophthalmology paradigm.

Figure 1.2: Teleophthalmology paradigms.

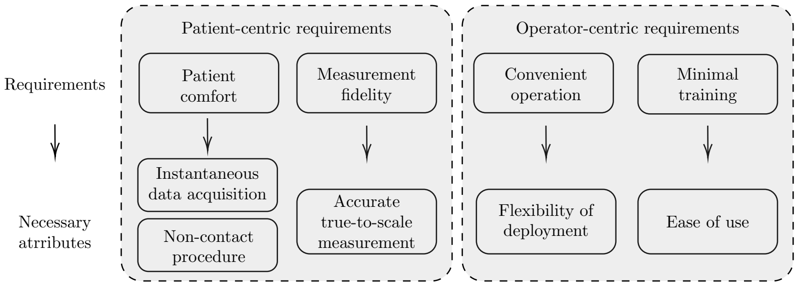

Requirements
Necessary atrributes
Patient-centric requirements
Patient comfort
Measurement fidelity
Instantaneous data acquisition
Non-contact procedure
Accurate true-to-scale measurement
Operator-centric requirements
Convenient operation
Minimal training
Flexibility of deployment
Ease of use

Figure 1.3: Key requirements and necessary attributes of a facial data acquisition system [8].

primary centers needs to meet certain requirements. From the clinical perspective, the subject under trauma needs to be relocated to hospital as soon as possible. Even in absence of trauma, a human being cannot remain stable for an extended period of time required by conventional face scanning systems. Therefore, the image acquisition operation should be instantaneous. Furthermore, due to the trauma and/or personal preferences, taking measurements of facial features using an instrument that requires physical contact with the skin is not suitable. From the perspective of regenerative surgery, the measurements should be accurate and retain the scale of the human face. Finally, the imaging system — unlike the legacy surface imaging equipment — should be easy to deploy and operate after a minimal training, owing to lack of trained professionals and other resources at a primary healthcare center. In summary, the following attributes — depicted in Figure 1.3 — are desirable in an oculofacial measurement platform.

## 1.2 Scope of the Thesis

In this thesis, we present a multicamera-based facial data acquisition system that meets the aforementioned requirements, and discuss its suitability for the proposed hybrid teleopthalmology paradigm. Our work builds upon the earlier works of Vupparaboina *et al.* on unambiguous Euclidean calibration of a camera network [9, 10], dense 3D reconstruction of a 3D head model [8, 11] and a preliminary sparse 3D reconstruction of a real human face [12] from multiple 2D views. The rest of this paper is organized as follows: Section 2 discusses the preliminaries to understand the rest of the thesis. Next, Section 3 discusses the Euclidean auto-calibration of a camera network consisting of a stereo pair and four monocular cameras. Further, Section 3 elaborates the theoretical underpinning, image acquisition setup, and 3D reconstruction. Euclidean auto calibration of the multi-stereo camera network is discussed in Section 4. The $3D$ point cloud obtained in Section 4 is denser and provides more information about the face. Section 5 discusses the broader impact of this work and concludes the thesis.

## 1.3 Literature Survey

In recent years, various 3D surface imaging technologies have been suggested for oculofacial surgical planning. Such systems are broadly classified as laser-based and optics-based systems [13]. Typical laser-based systems acquire 3D facial surface data on a band-by-band basis by projecting a line laser onto the subject's face and recording the time of flight of the laser. Albeit providing highly accurate 3D data, such systems require the subject to remain still for a considerably long period, and therefore remain inapt for conscious subjects, especially children. Movement increases the likelihood of distortion, noise, and voids in the scanned image.

The optics-based systems enable faster scanning. They primarily involve stereo photogrammetric techniques and/or structured light. The former ones use calibrated stereo camera pair(s). 3D information is obtained through triangulation of features across stereo images. Although instantaneous, such systems face two limitations (i) the pre-calibrated stereo cameras make the setup non-adaptive to ambient changes, and (ii) detecting and matching features across extra-ocular facial images is difficult and results sparse 3D reconstruction. To this end, a sequence of structured light patterns such as grids and/or dots are projected onto the face as features. The deviation in patterns from the reference patterns (precalibrated) provides required 3D information. However, such sequential image capturing unduly burdens the subjects by requiring them to remain still for several seconds.

# Chapter 2

# Preliminaries

In this chapter, we provide an understanding of the basic topics required to understand the rest of the work.

## 2.1 Image Formation

Let's consider a 3D point $A = [X;Y;Z]^T$ is imaged as point $D = [x;y]^T$ on image plane as shown in fig. 2.1. Note that ∆ABC and ∆DBE are similar triangle since they only differ in shape. Applying similar triangle rule in ∆ABC and ∆DBE gives

$$\frac{X}{x} = \frac{Z}{(f+v)}, \tag{2.1}$$

Similarly applying similar triangle rule in ∆GFB and ∆DFE gives

$$\frac{X}{x} = \frac{f}{v}, \tag{2.2}$$

substituting the value of $v$ from (2.2) to (2.1) gives

$$x = \frac{fX}{Z-f}, \tag{2.3}$$

Similarly applying similar triangle rule in orthogonal direction (Y direction) gives

$$y = \frac{fY}{Z-f}, \tag{2.4}$$

Considering thin lens approximation ($Z >> f$) in (2.3) and (2.4) grants

$$x = \frac{fX}{Z}, \tag{2.5}$$

$$y = \frac{fY}{Z}. \tag{2.6}$$

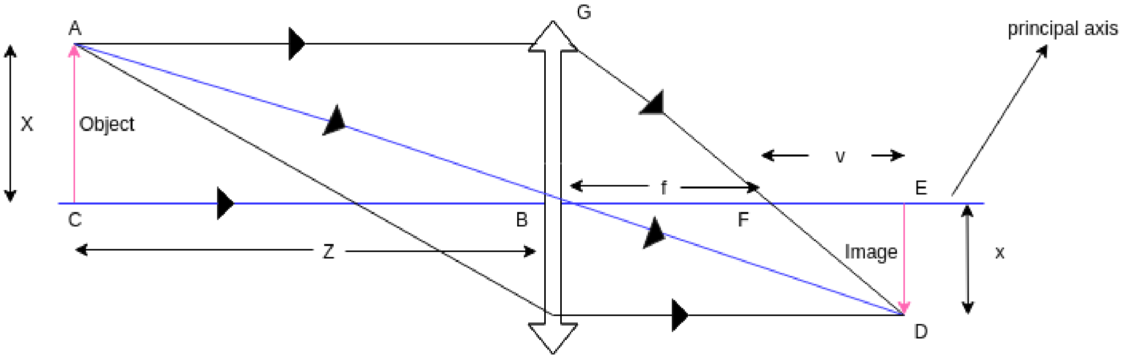


Figure 2.1: Image Formation.

Note that (2.5) and (2.6) can be written in matrix form as follows:

$$s \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & |0 \\ 0 & 1 & 0 & |0 \\ 0 & 0 & 1 & |0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2.7}$$

where $s$ is the scaling factor. So far all points in the 3D world have been represented in a camera centered coordinate system, that is, a coordinate system which has its origin at the camera center (point B in fig. 2.1). In practice however, the 3D points may be represented in terms of coordinates relative to an arbitrary coordinate system $(X', Y', Z')$. Assuming that the camera coordinate axes $(X, Y, Z)$ and the axes $(X', Y', Z')$ are of Euclidean type (orthogonal and isotropic), there is a unique Euclidean 3D transformation (rotation and translation) between the two coordinate systems.

The line, perpendicular to the image plane and goes through the camera center is called principal axis. The intersection of the principal axis with image plane is called principal point. (2.7) assumed that the origin of coordinates in the image plane is at the principal point. In practice, it may not be, so that in general there is a mapping

$$(X, Y, Z)^T \mapsto (fX/Z + u_x, fY/Z + u_y) \tag{2.8}$$

where $(u_x, u_y)$ are the coordinates of the principal point.

The generalized form of (2.7) is as follows :

$$s \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & u_x \\ 0 & f & u_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_1 & r_2 & r_3 & |t_x \\ r_4 & r_5 & r_6 & |t_y \\ r_7 & r_8 & r_9 & |t_z \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2.9}$$

where a 3D point $[X; Y; Z]^T$ is in any arbitrary coordinate system and it is brought to camera centered coordinate system by rotating and translating as shown in (2.9).

## 2.2 Triangulation

The method of finding corresponding 3D location of the point from two images of the point is called triangulation.

Let's consider a 3D point $\bar{X} = [X; Y; Z]$ is imaged as $\bar{x_1} = [x_1; y_1]$ in image 1 and $\bar{x_2} = [x_2; y_2]$ in image 2 as shown in fig 2.2. The distance between two camera centers ($C_1$ and $C_2$) is L. Considering the center of global coordinate system is at camera center 1 ($C_1$) and Z-axis as principal axis. Thus, the location of camera center 2 is at $[L; 0; 0]$. The image coordinates of the 3D point ($\bar{X}$) is given by (2.9) as shown below.

$$s_1 \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = \begin{bmatrix} f_1 & 0 & u_x^1 \\ 0 & f_1 & u_y^1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & |0 \\ 0 & 1 & 0 & |0 \\ 0 & 0 & 1 & |0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{2.10}$$

$$s_2 \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} f_2 & 0 & u_x^2 \\ 0 & f_2 & u_y^2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & |-L \\ 0 & 1 & 0 & |0 \\ 0 & 0 & 1 & |0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{2.11}$$

solving (2.10) and (2.11) yields

$$X = \frac{(x_1 - u_x^1)(y_2 - u_y^2)L}{(x_1 - u_x^1)(y_2 - u_y^2) - (x_2 - u_x^2)(y_1 - u_y^1)} \tag{2.12}$$

$$Y = \frac{(y_1 - u_y^1)(y_2 - u_y^2)L}{(x_1 - u_x^1)(y_2 - u_y^2) - (x_2 - u_x^2)(y_1 - u_y^1)} \tag{2.13}$$

$$f_2 = \frac{(y_2 - u_y^2)f_1}{(y_1 - u_y^1)} \tag{2.14}$$

$$Z = \frac{f_1 X}{(x_1 - u_x^1)} \tag{2.15}$$

where $s_1$ and $s_2$ denotes the scaling factors of camera 1 and camera 2. $(u_x^1, u_y^1)$ and $(u_x^2, u_y^2)$ denotes the principal points of camera 1 and camera 2.

Note that if $L$ and intrinsic parameters of the camera (focal length, principal points) are known then 3D location of the point can be found using triangulation described in (2.12)- (2.15). The method of finding these quantities (intrinsic parameters of the camera) is known as camera calibration.

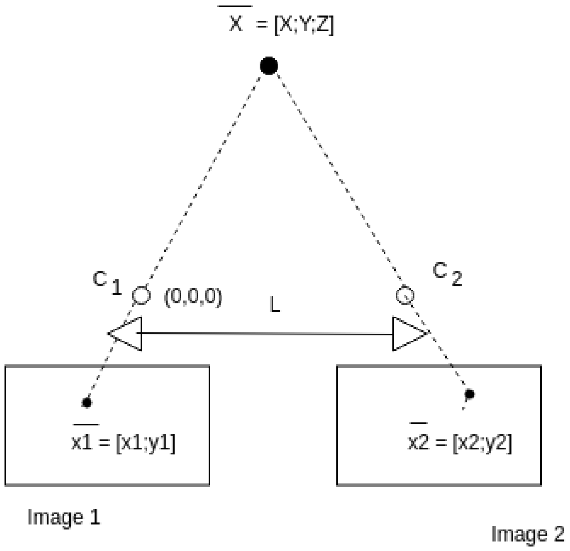


Figure 2.2: Triangulation.

# Chapter 3

# Euclidean Auto Calibration

Vuppparaboina *et al.* presented a method for auto calibration of the camera based on point correspondences between multiple images to recover the scale and shape of the object [10]. The author proposed a camera network configuration consisting of a stereo pair with known baseline separation and performed euclidean auto calibration for four camera network (one stereo pair and two monocular cameras). In which, first common point correspondences across all the four images are established and then cost function (reprojection error) is defined for all monocular cameras. The optimization problem is set up with the objective of minimizing the cost function (reprojection error). The solution to this optimization grants internal (focal length and principal points) and external parameters (rotations and translations) of the cameras, from which 3D location can be known with the help of triangulation between stereo pair.

Intuitively, the robustness of the euclidean auto calibration depends on the number of different views of the object. Increasing in number of cameras/views improves the quality of the auto calibration. However, this gives rise to the decrease in common matches across all the images, resulting in poor auto calibration. In this backdrop, we propose a six camera network configuration. We define a weighted cost function which considers not only common correspondences across all six cameras but also any subset of cameras (including stereo). The results shows that shape and scale of the object is well preserved from which we conclude the quality of auto calibration.

## 3.1 Theoretical Background

In this section theoretical principal enabling true-to-scale 3D reconstruction is explained.

We consider six camera network consisting of one stereo camera pair and four monocular cameras as shown in Fig. 3.1. In addition to this, we consider that the distance between camera centers of stereo pair is known and given by L. Let's say each image shares N common feature points. The homogeneous 3D coordinates of these feature points are $\bar{X}_k = [X_kY_kZ_k1]^T (k = 1, 2, ..., K)$. The corresponding image points of these feature points are given by $\bar{x_{ik}} = [x_{ik}y_{ik}1]^T$, where i =1, 2, 3, 4, 5, 6 is the indices of the Mono-1 (M1), Mono-2 (M2), Stereo left (SL), Stereo right (SR), Mono-3

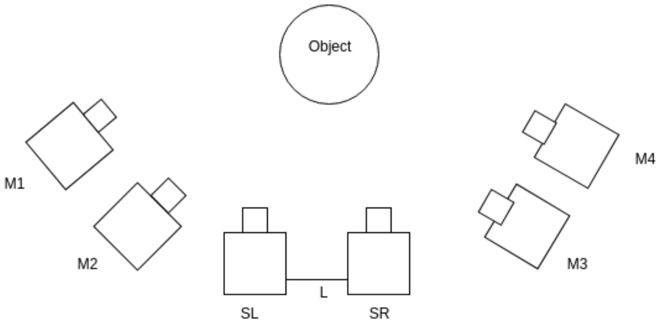


Figure 3.1: Schematic of camera network.

(M3) and Mono-4 (M4) cameras, respectively. Image point $\bar{x_{ik}}$ of $\bar{X_k}(k = 1, 2, ..., K)$ is given by

$$s_{ik}\begin{bmatrix}x_{ik}\\y_{ik}\\1\end{bmatrix} = K_i\begin{bmatrix}R_i|T_i\end{bmatrix}\begin{bmatrix}X_k\\Y_k\\Z_k\\1\end{bmatrix} \tag{3.1}$$

where $s_{ik}$ represents the scaling factor. $K_i$ represents the intrinsic camera matrix and consists of principal points and focal length of the cameras. Furthermore, $[R_i|T_i]$ represents the extrinsic camera matrix and consists of 3D rotation matrix and translation vector.

$$K_i = \begin{bmatrix}f_i & 0 & u_x^i\\0 & f_i & u_y^i\\0 & 0 & 1\end{bmatrix}, \tag{3.2}$$

$$R_i = R_i^Z \times R_i^Y \times R_i^X = \begin{bmatrix}\cos{\theta_3}^i & -\sin{\theta_3}^i & 0\\\sin{\theta_3}^i & \cos{\theta_3}^i & 0\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos{\theta_2}^i & 0 & \sin{\theta_2}^i\\0 & 1 & 0\\-\sin{\theta_2}^i & 0 & \cos{\theta_2}^i\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\0 & \cos{\theta_1}^i & -\sin{\theta_1}^i\\0 & \sin{\theta_1}^i & \cos{\theta_1}^i\end{bmatrix}, \tag{3.3}$$

$$T_i = [T_i^X\ T_i^Y\ T_i^Z]^T. \tag{3.4}$$

Since Camera 3 and 4 forms a stereo pair and the distance between two camera centers are known (L). Assuming global coordinate system is aligned with the camera centered coordinate system of the camera 3 and principal axis as Z-axis. Thus, the location of camera center 4 is $(L, 0, 0)^T$. Furthermore, we consider coordinate systems of camera 3 and 4 are shifted with respect to each other,i.e, rotation matrix is identity for both the cameras. Now solving for i = 3, 4 yields

$$X_k = \frac{(x_{3k} - u_x^3)(y_{4k} - u_y^4)L}{(x_{3k} - u_x^3)(y_{4k} - u_y^4) - (x_{4k} - u_x^4)(y_{3k} - u_y^3)}, \tag{3.5}$$

$$Y_k = \frac{(y_{3k} - u_y^3)(y_{4k} - u_y^4)L}{(x_{3k} - u_x^3)(y_{4k} - u_y^4) - (x_{4k} - u_x^4)(y_{3k} - u_y^3)}, \tag{3.6}$$

$$f_4 = \frac{(y_{4k} - u_y^4)f_3}{(y_{3k} - u_y^3)}, \tag{3.7}$$

$$Z_k = \frac{f_3 X_k}{(x_{3k} - u_x^3)}. \tag{3.8}$$

Note that camera centers and focal length are not known a priori since cameras are uncalibrated. The 3D location of the feature point can be found from (2.5)-(2.8) only if $f_3$, $(u_x^3, u_y^3)$ and $(u_x^4, u_y^4)$ are known. In order to find these quantities, monocular cameras are used. to see this, note the following. Using (2.5)-(2.8) in (2.1) will introduce the $f_3$, $(u_x^3, u_y^3)$ and $(u_x^4, u_y^4)$ in (2.1). Now one can obtain all the parameters including $f_3$, $(u_x^3, u_y^3)$ and $(u_x^4, u_y^4)$ by minimizing the squared error between left hand side and right hand side of (2.1) for all monocular cameras.

The objective function (reprojection error) is given by

$$\min \sum_{i=1,2,5,6} \sum_{k=1}^{K} ||(\bar{x_{ik}} - K_i[R_i|T_i]\bar{X_k})||_2 \tag{3.9}$$

Note that (3.9) shows the error between actual image coordinates and estimated image coordinates for all the feature points in all monocular cameras. $\bar{X_j}$ is calculated in terms of $f_3$, $(u_x^3, u_y^3)$ and $(u_x^4, u_y^4)$ using triangulation between stereo pair (SL and SR).

The solution to this optimization provides internal and external parameters of the cameras from which 3D location of the points can be found using triangulation between stereo pair. The common correspondences across six cameras are bound to be sparse due to large distance between the leftmost and rightmost camera. Intuitively the solution to this optimization is not robust since the number of points (common correspondences) used in optimization is very few.

### 3.1.1 Proposed Weighted Objective Function

We propose a weighted objective function which not only considers common correspondences across the six cameras but also uses the common correspondences across any subset of the six cameras. In particular, common correspondences across six cameras, left four cameras (M1, M2, SL, SR) and right four cameras (SL, SR, M3, M4) are used. Notice that each subset must include stereo pair since triangulation is used in the above formulation. Thus, one can define the reprojection error for each subset as described in (3.9). Let's consider the number of non-repetitive common correspondences across six cameras, left four cameras and right four cameras are respectively N, M1 and M2 and the corresponding homogeneous 3D coordinates of the feature points are given by $\bar{X_j} = [X_j Y_j Z_j 1]^T (j = 1, 2, ..., N)$, $\bar{X_p} = [X_p Y_p Z_p 1]^T (p = 1, 2, ..., M1)$ and $\bar{X_q} = [X_q Y_q Z_q 1]^T (q = 1, 2, ..., M2)$. The weighted reprojection error (f) is given by

$$f = \mu_1 \times Error_6 + \mu_2 \times Error_{L4} + \mu_3 \times Error_{R4} \tag{3.10}$$

where

$$Error_6 = \sum_{i=1,2,5,6} \sum_{j=1}^{N} ||(\bar{x_{ij}} - K_i[R_i|T_i]\bar{X_j})||_2 \tag{3.11}$$

$$Error_{L4} = \sum_{i=1,2} \sum_{p=1}^{M1} ||(\bar{x_{ip}} - K_i[R_i|T_i]\bar{X_p})||_2 \tag{3.12}$$

$$Error_{R4} = \sum_{i=5,6} \sum_{q=1}^{M2} ||(\bar{x_{iq}} - K_i[R_i|T_i]\bar{X_q})||_2 \tag{3.13}$$

and $\mu_1$, $\mu_2$ and $\mu_3$ are the relative weights given to the error terms and must add up to one. i.e $\mu_1 + \mu_2 + \mu_3 = 1$. Intuitively, $\mu_1$ should be greater than $\mu_2$ and $\mu_3$ since $Error_6$ contains more terms than $Error_{L4}$ and $Error_{R4}$.

Note that minimizing (3.10) yields internal (focal length and camera centers) and external (rotations, translations) parameters of the cameras. We expect that solution to this optimization is more robust than optimization proposed in (3.9) since more number of points are used in (3.10) compared to (3.9).

## 3.2 Methodology

In this section, end to end pipeline enabling true-to-scale 3D reconstruction is discussed in detail. The 3D reconstruction of the human face using euclidean auto calibration is demonstrated. We compare the proposed method with the earlier method by Vuppparaboina *et al.* The robustness of the auto calibration is measured in terms of the accuracy of the 3D reconstruction. We compare manually measured features of the face such as nose length, lips width, eye width, Z-span, etc with reconstructed 3D point cloud.

### 3.2.1 Image Acquisition

The image acquisition system comprises of a network of six Internet protocol (IP) cameras [14], two independent projectors [15] and a desktop computer which controls the cameras via an Ethernet switch [16]. Two of six cameras were arranged as a stereo pair with a baseline length of 45mm. The stereo pair was located between two monocular cameras on the left and two monocular cameras on the right. The monocular cameras to the left of the stereo, the two cameras in the stereo pair and the monocular cameras to the right of the stereo pair are identified as M1, M2, SL, SR, M3 and M4, respectively.

The projectors were used to project a fixed structured light pattern in order to create features to be tracked across the multiple images of the human face. The structured-light pattern largely determines the density of the reconstructed 3D point cloud. The density of the recovered 3D point cloud is resultant of the sequence of patterns projected. In contrast, we used a colored structured light pattern. The pattern contained square patches of selected colors. Each patch had a marker

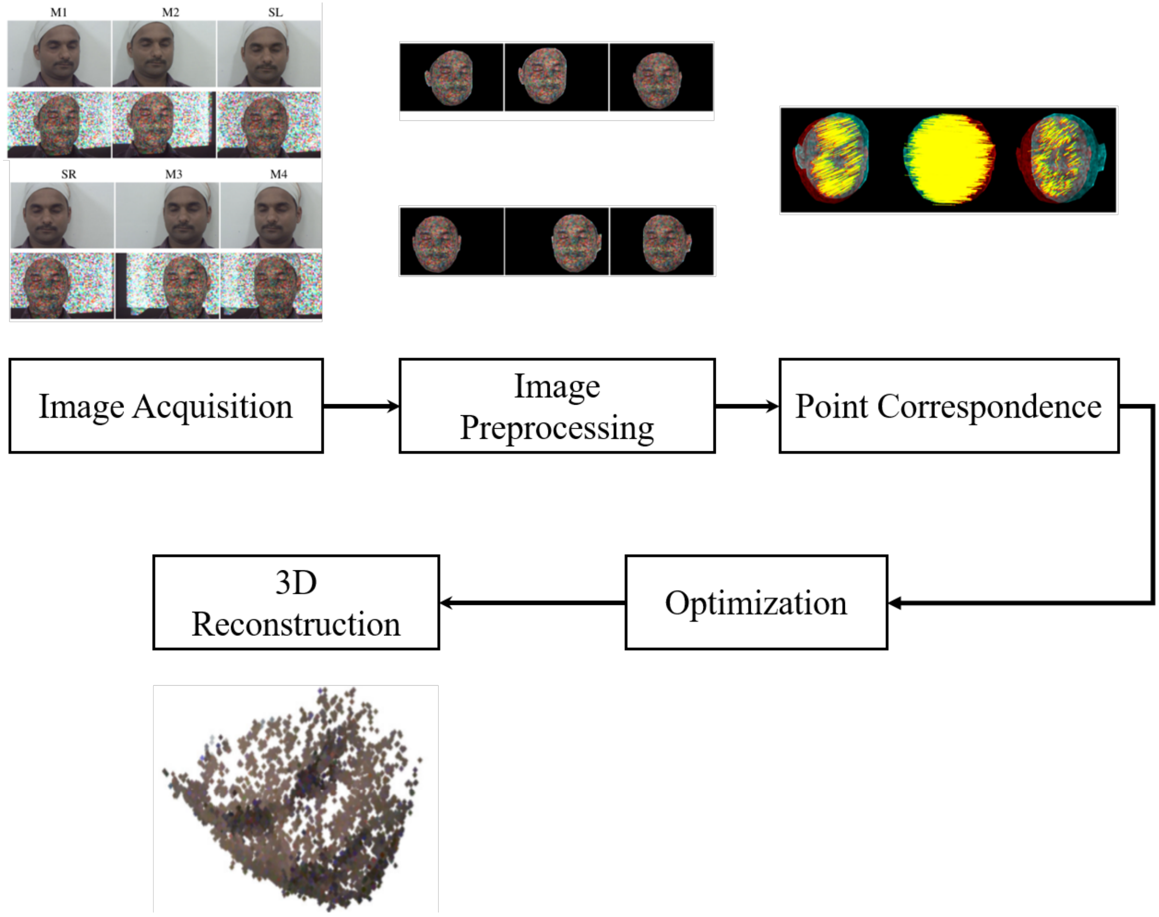


Figure 3.2: Flowchart of Methodology.

at its center to facilitate dense matching. The markers serve as feature points and corners of patches provide feature descriptors. In an earlier work, the structured light pattern projected from the front was deformed at the sides (cheeks) of the face, posing difficulties in establishing point correspondences [8]. Therefore, in this work, we projected the same structured light pattern from two projectors whose optical axes approximately converged at the centre of the face.

Then, one image per camera was captured in a sequential manner using software triggering. Although sequential, the image acquisition process was found to be nearly instantaneous. The process was repeated for two more patterns with increasing density of features. The camera settings, optimized to maximize the image quality given the indoor lighting conditions, were made identical across the cameras.

This image acquisition system was arranged in a laboratory whose dimensions were $11 \times 19.25$ sq. ft. The setup itself occupied the floor area of $7 \times 7$ sq. ft. The subject was seated on a stable plastic chair. The three stereo pairs were arranged on three separate tripods. The distance between middle stereo camera and the approximate centre point of the chair was 2 ft. The images were acquired with two lighting conditions for two specific purposes. First, the texture information was acquired with two fluorescent tube lights turned on. Next, with the tube lights turned off, the structured light patterns were projected and another set of 6 images were captured in order to track common features across the viewpoints. The entire operation took a few seconds to complete.

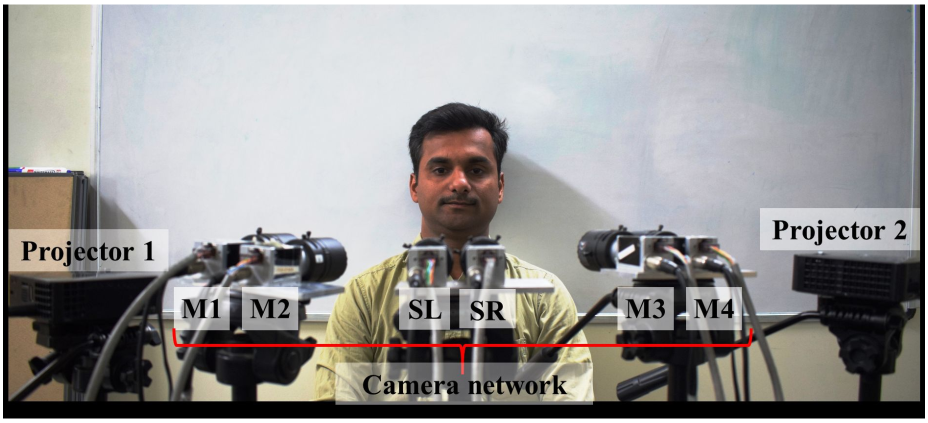


Figure 3.3: An instance of the oculofacial surface imaging platform used in this thesis. ‘SL’ and ‘SR’ cameras form a stereo pair. ‘Mi’ indicates i$^{\text{th}}$ monocular camera.

A set of six images of a human test subject captured in this manner is shown in Fig. 3.4. The top row depicts texture information, the middle row shows images acquired with structured light patterns projected, and the third row shows masked images in which only the regions of interest (ROI) are retained. The masking was achieved using morphological operations [17]. Extracting ROIs facilitates robust point correspondence across the images.

### 3.2.2 Point Correspondences

The point correspondences between multiple images are established using affine scale-invariant feature transform (ASIFT) [18]. ASIFT is employed because it gives many more feature descriptors than SIFT, SURF, etc. ASIFT provides common matches between two images. In order to find the common matches across multiple images, we consider any one of the images as a reference image and then matching is established for all the images w.r.t. the reference image. Afterwards, coordinates of the reference image which are common in all the matching sets are retained. Finally, respective correspondences in other images are traced through the retained common points in the reference image. In particular, SL image was considered as a reference image and feature correspondences were established w.r.t. SL image as shown in the figure below. Common correspondences across all six images, left four images and right four images are depicted in Figs. 3.5, **??** and **??**, respectively. Note that all correspondences are non repetitive. In our experiment, ASIFT provided 773, 1008 and 908 non repetitive common correspondences for six cameras, left four cameras and right four cameras respectively.

### 3.2.3 Optimization

In this section, we elaborate on the optimizer used in minimizing the cost function. A combination of `GA` (genetic algorithm) and `sqp` (Sequential quadratic programming) algorithm is used to find the global optima of the cost functions ( (3.9) and (3.10)). Note that $f_3$ and $fi, \theta_1^i, \theta_2^i, \theta_3^i, t_i^X, t_i^Y, t_i^Z, (u_x^i, u_y^i)$ $i = 1, 2, 5, 6$ constitute a set of 41 independent variables, from which the rest of the unknown quantities can all be derived [10], and it is enough to perform the aforementioned minimization with respect to only these 41 independent variables. The solver consists of `GA` and `sqp`. `GA` inputs the

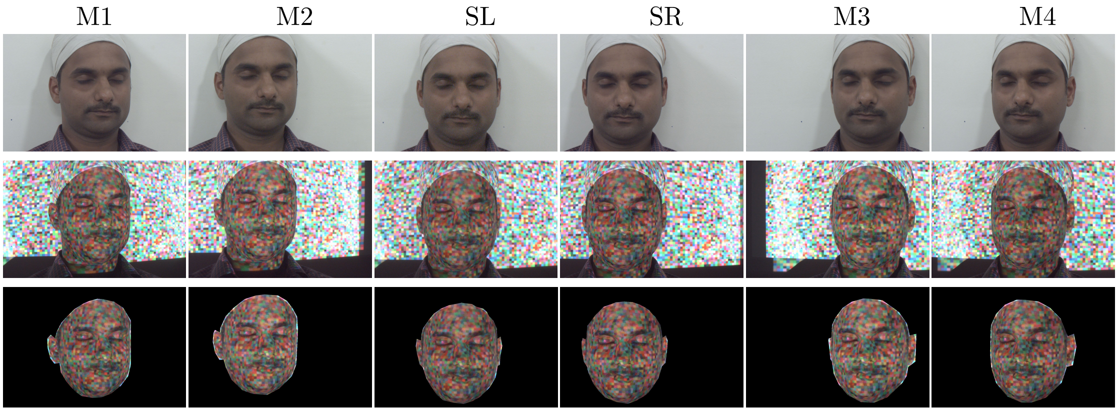
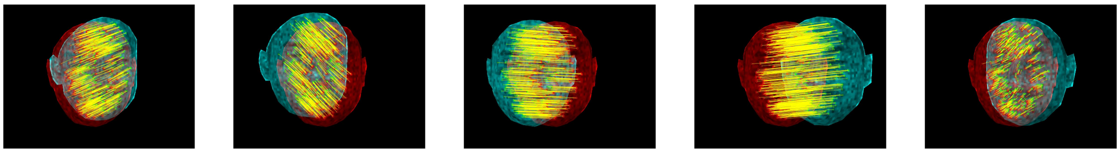
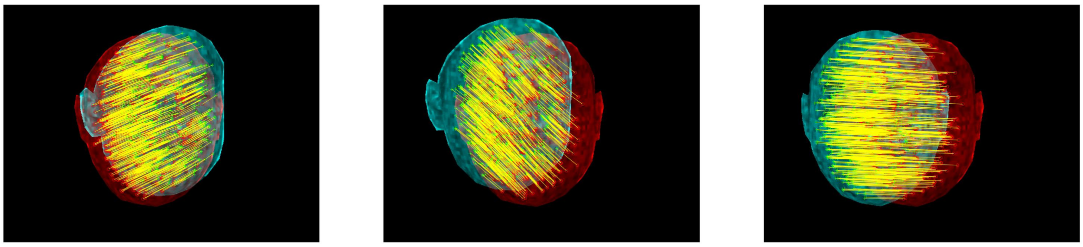


Figure 3.4: Multiple 2D views of the human face. Top row: Images captured for texture information. Middle row: Images with structured light pattern projected. Bottom row: Images with only the region-of-interest retained.



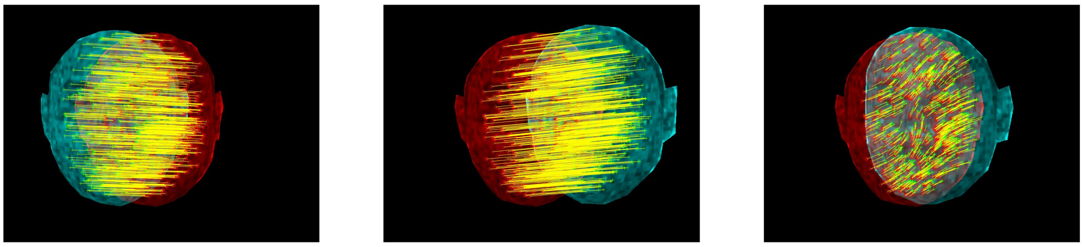
(a) SL-M1 (b) SL-M2 (c) SL-SR (d) SL-M3 (e) SL-M4

Figure 3.5: Common point correspondences across six cameras.



(a) SL-M1-L (b) SL-M2-L (c) SL-SR-L

Figure 3.6: Common point correspondences across left four cameras.



(a) SL-SR-R (b) SL-M3-R (c) SL-M4-R

Figure 3.7: Common point correspondences across right four cameras.

Table 3.1: Comparison between the proposed auto-calibration method and an earlier method by Vupparaboina *et al.* [10].

| Comparison | **Vupparaboina *et al.* [10]**<br>$(r_1, r_2, r_4)$<br>$(u_x, u_y)$<br>$f$<br>$(t_x, t_y, tz)$ | **Proposed Auto Calibration**<br>$(r_1, r_2, r_4)$<br>$(u_x, u_y)$<br>$f$<br>$(t_x, t_y, tz)$ |
|---|---|---|
| M1 | (0.9947,0.0029,-0.0015)<br>(448.8701,363.8923)<br>1272.6<br>(118.3267,-19.2712,-14.0177) | (0.9917,0.0038,-0.0018)<br>(499.7684,371.3564)<br>1652.3<br>(121.0157,-18.8506,-4.0237) |
| M2 | (0.9971,-0.0050,0.0060)<br>(348.7187,332.0304)<br>1302.5<br>(80.7601,-17.3661,-14.8395) | (1,0.0016,-0.0016)<br>(240.9126,372.7266)<br>1616.6<br>(78.4704,-13.1577,-32.7648) |
| SL | Rotation matrix is identity<br>(518.8079,369.6067)<br>1353<br>(0,0,0) | Rotation matrix is identity<br>(541.6839,379.9802)<br>1728.8<br>(0,0,0) |
| SR | Rotation matrix is identity,<br>(473.4096,361.1591)<br>1366.2<br>(45,0,0) | Rotation matrix is identity<br>(513.9444,371.7380)<br>1744<br>(45,0,0) |
| M3 | (0.9942,0.0057,-0.0077)<br>(874.9307,372.7389)<br>1419.10<br>(-104.4869,-20.6945,22.9805) | (0.9920,0.0049,-0.0047)<br>(861.9289,422.0566)<br>1711<br>(-103.4769,-21.0602,-9.1717) |
| M4 | (0.9929,0.0032,0.0053)<br>(762.0662,380.8103)<br>1338.3<br>(-144.0685,-21.4870,-1.3291) | (0.9885,0.0018,-0.0016)<br>(737.1850,426.6488)<br>1665.3<br>(-143.2529,-21.6132,-16.0516) |

number of iterations, population, and bounds. The number of iterations and population given to the `GA` solver are 700 and 500. The search space of the `GA` is constrained by providing loose bounds to the `GA` solver as mentioned in the Table 3.2. The search space of the `GA` is further reduced by providing integer indices to the camera centers, focal lengths, and translation parameters, i.e. `GA` solver provides only integer values to these variables. The solution of the GA serves the starting point to the `sqp`. `sqp` searches the Optima around the starting point and provides a more precise solution. The number of iterations given to the `sqp` solver is 25000.

The optimization is performed for the proposed cost function mentioned in (2.10). Note that, we weighted the $Error_6$, $Error_{4L}$ and $Error_{4R}$ based on number of cameras. i.e $\mu_1 = 6/14$ and $\mu_2 = \mu_3 = 4/14$. A comparison between the solution of the optimization is provided for the earlier method by Vupparaboina *et al.* and the proposed method in Table 3.1. Table 3.1 shows the variables (rotation, camera center, focal length and translation) obtained from optimization for each of the cameras.

Table 3.2: Bounds on optimization variables.

| Variables | f (all focal lengths) | r (all rotation parameters for mono) | Translations ($T_x$, $T_y$, $T_z$) | Camera centre (u, v) |
|---|---|---|---|---|
| Lower bound | 800 | $-\pi$ | (-200, -100, -100) | (300, 300) |
| Upper bound | 2000 | $\pi$ | (200, 100, 100) | (900, 700) |

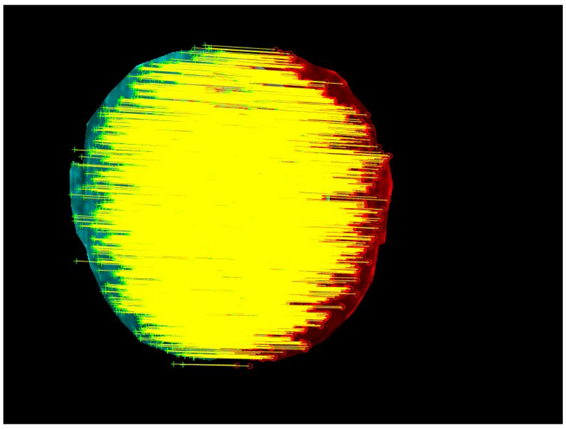


Figure 3.8: Dense correspondence between the two cameras of the stereo pair.

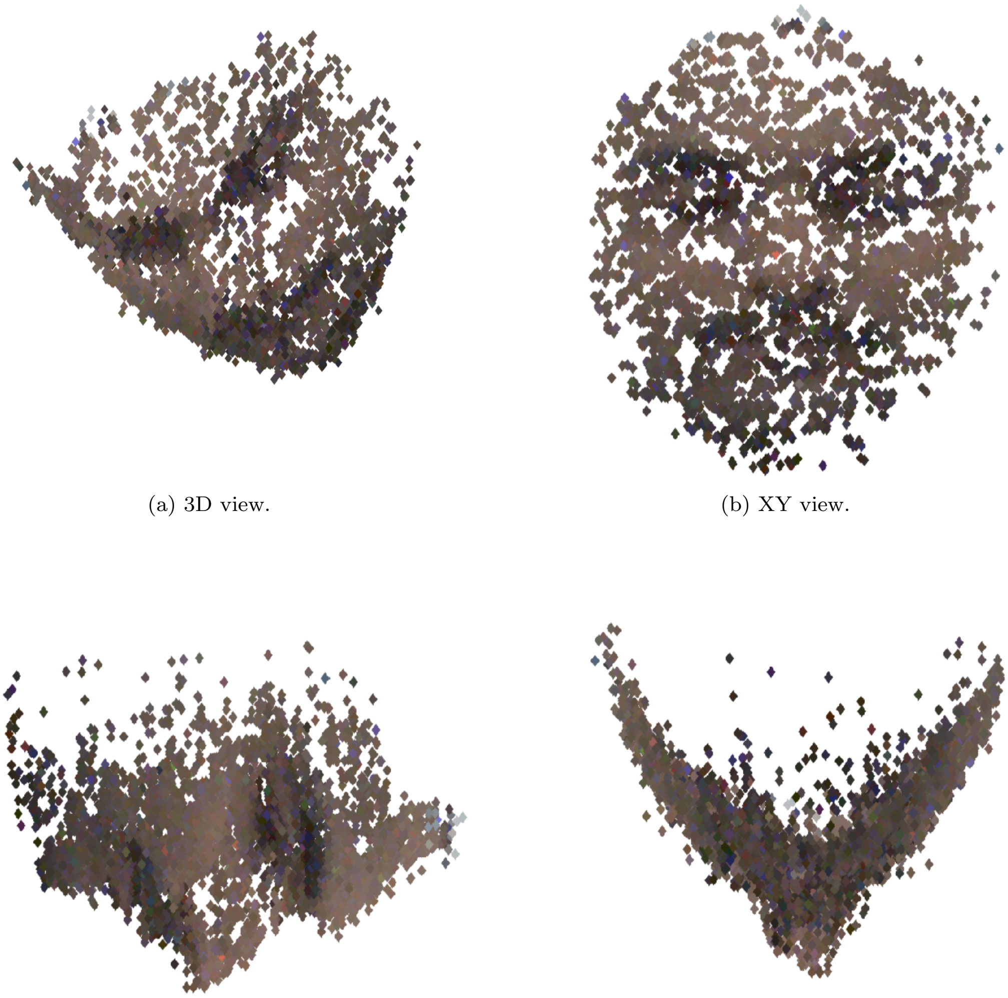
## 3.3 3D Reconstruction

We now provide the results of the 3D reconstruction. Note that, the robustness of the auto calibration is measured in terms of the accuracy of 3D reconstruction. Table 3.3 compares the few key facial measurements with the reconstructed point cloud by both of the methods. The measurements of the 3D point cloud are distances — measured in MATLAB — between two extreme points of the selected facial feature. Since the reconstructed point cloud was sparse and the points were selected manually, the measurements could have an error. Similarly, for measurements on the actual face, we attempted the best possible selection of the two extreme points between which the distances were measured. Clearly, the proposed method provides more accurate measurements compared to the earlier method by Vupparaboina *et al.*

The 3D point cloud with texture information is visualized in MATLAB. Note that, correspondences between stereo pair, $f_3$, $(u_x^3, u_y^3)$ and $(u_x^4, u_y^4)$ are enough to find the 3D location of the points, i.e. triangulation. ASIFT provides dense correspondence between stereo pair as shown in Fig. 3.8. We compute 3D location of these points using triangulation as described in (2.5)-(2.8). Further, outliers are removed by denoising the point cloud. The multiple views of the denoised 3D point cloud are shown in the Fig. 3.9. In addition to this, a surface fitted on this point cloud is also shown in Fig. 3.10. The screened Poisson surface reconstruction method [19] — available in Meshlab software [20] — was used to generate the surface.

Table 3.3: Key facial measurements (all measurements are in mm).

| Facial measurement | Original face | 3D point cloud by Vupparaboina *et al.* [10] | 3D point cloud by proposed method |
|---|---|---|---|
| Lip width | 55 | 63.64 | 52.76 |
| Nose length | 50 | 61.54 | 49.15 |
| eye socket length | 38 | 41.6 | 38.8 |
| Nose to upper lips | 30 | 38.57 | 29.24 |
| Z-span | 68 | 91 | 71 |



(a) 3D view.

(b) XY view.

(c) YZ view.

(d) XZ view.

Figure 3.9: 3D point cloud with texture information.

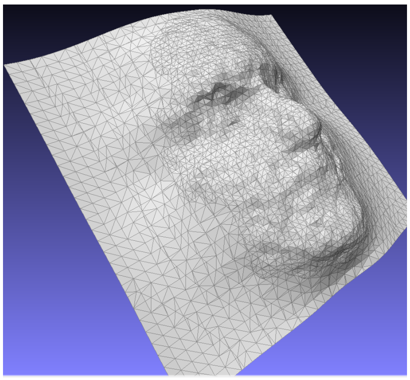


Figure 3.10: 3D mesh visualized in Meshlab.

# Chapter 4

# Euclidean Auto Calibration of Multi-stereo camera Network

In the previous chapter, we demonstrated the euclidean auto calibration of the camera network consists of one stereo pair and four monocular cameras. Triangulating the common points between this stereo pair provides the 3D information of only one part of the face. However, our primary goal is to provide a tool for the physician for facial surgical planning which requires the 3D information over the whole periphery of the face. For this, we propose the euclidean auto calibration of the camera network consisting of three stereo pair which can be further generalized to $N$ stereo pair. The advantage of auto-calibrating such network is it allows us to get the 3D information over the whole periphery of the face. i.e The cost function of the camera network is defined in a way that each stereo pair can be triangulated. The proposed multi-stereo camera network enables true-to-scale 3D reconstruction of the human face.

In this chapter, we demonstrate the euclidean auto calibration of the multi-stereo camera network. The auto-calibration grants the parameters from which 3D point cloud is reconstructed. The manually measured face features are compared with the reconstructed 3D point cloud from which we conclude that shape and scale of the object (face) are preserved.

## 4.1 Theoretical Background

In this section, we explore the geometry of the multi-stereo camera system based on which the cost function (reprojection error) is defined.

Let's consider a multi-stereo camera network consisting of three stereo pair with known baselines (L) as shown in Fig. 4.1. The earlier formulation presented in chapter 2 considers the origin of the world coordinate system is at camera center 3. Thus camera 3 and 4 form a stereo pair and rest of the cameras were treated as monocular since they all have different rotations and translations parameters w.r.t reference camera (camera 3). Clearly, all the stereo pairs cannot be treated as a stereo pair in a single world coordinate system.

The idea is to consider multiple world coordinate systems each of them having the origin at camera centers from one of the cameras from every stereo pair. In particular, we consider three world coordinate systems and the origins of the world coordinate system 1, 2 and 3 are at camera centers of

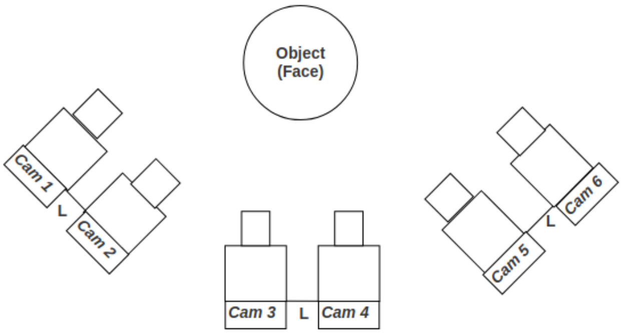


Figure 4.1: Schematic of camera network.

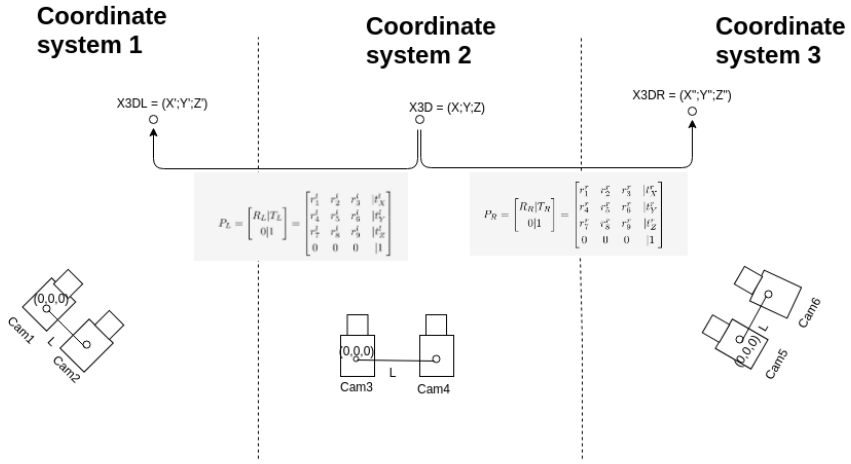


Figure 4.2: Multiple coordinate systems.

camera 1, 3 and 5. Now we define the relation between these three coordinate systems. If a 3D point $X3D$ has the homogeneous coordinates $[X;Y;Z;1]$ in the coordinate system 2 (middle coordinate system), then the same point will have the homogeneous coordinates $X3D_L = [X';Y';Z';1] = P_L \times X3D$ in the coordinate system 1 (left coordinate system) and $X3D_R = [X";Y";Z";1] = P_R \times X3D$ in the coordinate system 3 (right coordinate system), where $P_L$ and $P_R$ denotes the left and right coordinate transformation matrix respectively. The schematic diagram of the above explanation is shown in fig. 4.2. $P_L$ and $P_R$ consist of 3D rotations and translations matrices as shown below.

$$P_L = \begin{bmatrix} R_L|T_L \\ 0|1 \end{bmatrix} = \begin{bmatrix} r_1^l & r_2^l & r_3^l & |t_X^l \\ r_4^l & r_5^l & r_6^l & |t_Y^l \\ r_7^l & r_8^l & r_9^l & |t_Z^l \\ 0 & 0 & 0 & |1 \end{bmatrix} \tag{4.1}$$

$$P_R = \begin{bmatrix} R_R|T_R \\ 0|1 \end{bmatrix} = \begin{bmatrix} r_1^r & r_2^r & r_3^r & |t_X^r \\ r_4^r & r_5^r & r_6^r & |t_Y^r \\ r_7^r & r_8^r & r_9^r & |t_Z^r \\ 0 & 0 & 0 & |1 \end{bmatrix} \tag{4.2}$$

Note that $R_L$ and $R_R$ have 3 degree of freedom and it can be decomposed as shown.

$$R_L = R_L^Z \times R_L^Y \times R_L^X = \begin{bmatrix} \cos\theta_3{}^l & -\sin\theta_3{}^l & 0 \\ \sin\theta_3{}^l & \cos\theta_3{}^l & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta_2{}^l & 0 & \sin\theta_2{}^l \\ 0 & 1 & 0 \\ -\sin\theta_2{}^l & 0 & \cos\theta_2{}^l \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_1{}^l & -\sin\theta_1{}^l \\ 0 & \sin\theta_1{}^l & \cos\theta_1{}^l \end{bmatrix} \tag{4.3}$$

Now consider each image shares N common feature points. The homogeneous 3D coordinates of the feature points in **coordinate system 2** are given by $\bar{X_k} = [X_k Y_k Z_k 1]^T (k = 1, 2, ..., K)$. Each feature point $\bar{X_k}(k = 1, 2, ..., K)$ is imaged as $x_{ik} = [x_{ik} y_{ik} 1]^T$, where i =1, 2, 3, 4, 5, 6 represents the indices of the cameras as shown in Fig. 4.1. Image point $\bar{x_{ik}}$ of $\bar{X_k}(k = 1, 2, ..., K)$ in **coordinate system 1** is given by

$$s_{1k} \begin{bmatrix} x_{1k} \\ y_{1k} \\ 1 \end{bmatrix} = K_1 \begin{bmatrix} I|\bar{0} \end{bmatrix} \begin{bmatrix} R_L|T_L \\ 0|1 \end{bmatrix} \begin{bmatrix} X_k \\ Y_k \\ Z_k \\ 1 \end{bmatrix} \tag{4.4}$$

$$s_{2k} \begin{bmatrix} x_{2k} \\ y_{2k} \\ 1 \end{bmatrix} = K_2 \begin{bmatrix} I|\bar{L} \end{bmatrix} \begin{bmatrix} R_L|T_L \\ 0|1 \end{bmatrix} \begin{bmatrix} X_k \\ Y_k \\ Z_k \\ 1 \end{bmatrix} \tag{4.5}$$

where $s_{ik}$ represents the scaling factor. $I$ is the 3×3 identity matrix and $\bar{0}$ is the 3×1 zero vector. $K_i$ is the intrinsic matrix of the camera and given by

$$K_i = \begin{bmatrix} f_i & 0 & u_x^i \\ 0 & f_i & u_y^i \\ 0 & 0 & 1 \end{bmatrix} \tag{4.6}$$

Note that the center of coordinate system 1 is at camera center 1 and z-axis is considered as the optical axis. Thus center of camera 2 has the location $\bar{L} = [L; 0; 0]^T$. In the right-hand sides of 4.4 and 4.5, we transformed the feature points $\bar{X_j}$ in the coordinate system 1 using $P_L$. Thus the coordinates of feature points $\bar{X_j}$ in coordinate system 1 is $P_L \times \bar{X_j}$. Now we use the fact that camera 1 and 2 is stereo and gets the left-hand sides.

Similarly image of the feature point j in camera 5, 6 in **coordinate system 3** is given by

$$s_{5k} \begin{bmatrix} x_{5k} \\ y_{5k} \\ 1 \end{bmatrix} = K_5 \begin{bmatrix} I|\bar{0} \end{bmatrix} \begin{bmatrix} R_R|T_R \\ 0|1 \end{bmatrix} \begin{bmatrix} X_k \\ Y_k \\ Z_k \\ 1 \end{bmatrix} \tag{4.7}$$

$$s_{6k}\begin{bmatrix}x_{6k}\\y_{6k}\\1\end{bmatrix}=K_6\begin{bmatrix}I|\bar{L}\end{bmatrix}\begin{bmatrix}R_R|T_R\\0|1\end{bmatrix}\begin{bmatrix}X_k\\Y_k\\Z_k\\1\end{bmatrix}\tag{4.8}$$

where $\bar{X_k}=[X_kY_kZ_k1]^T(k=1,2,...,K)$ is calculated using triangulation in camera 3 and 4 in **coordinate system 2** as shown below.

$$X_k=\frac{(x_{3k}-u_x^3)(y_{4k}-u_y^4)L}{(x_{3k}-u_x^3)(y_{4k}-u_y^4)-(x_{4k}-u_x^4)(y_{3k}-u_y^3)},\tag{4.9}$$

$$Y_k=\frac{(y_{3k}-u_y^3)(y_{4k}-u_y^4)L}{(x_{3k}-u_x^3)(y_{4k}-u_y^4)-(x_{4k}-u_x^4)(y_{3k}-u_y^3)},\tag{4.10}$$

$$f_4=\frac{(y_{4k}-u_y^4)f_3}{(y_{3k}-u_y^3)},\tag{4.11}$$

$$Z_k=\frac{f_3X_k}{(x_{3k}-u_x^3)}.\tag{4.12}$$

Note that using (4.9)- (4.12) in (4.4), (4.5), (4.7) and (4.8) will introduce the $f_3$, $(u_x^3,u_y^3)$ and $(u_x^4,u_y^4)$ in (4.4), (4.5), (4.7) and (4.8). Now one can define the reprojection error as the composite sum of squared error between left-hand side (L.H.S) and right-hand side (R.H.S) of equations (4.4), (4.5), (4.7) and (4.8).

The overall reprojection error is given by

$$Error=Error_{Coordinate1}+Error_{Coordinate3},\tag{4.13}$$

where

$$Error_{Coordinate1}=\sum_{k=1}^{K}||(\bar{x_{1k}}-K_1\begin{bmatrix}I|\bar{0}\end{bmatrix}\begin{bmatrix}R_L|T_L\\0|1\end{bmatrix}\bar{X_k})||_2+||(\bar{x_{2k}}-K_2\begin{bmatrix}I|\bar{L}\end{bmatrix}\begin{bmatrix}R_L|T_L\\0|1\end{bmatrix}\bar{X_k})||_2,\tag{4.14}$$

and

$$Error_{Coordinate3}=\sum_{k=1}^{K}||(\bar{x_{5k}}-K_5\begin{bmatrix}I|\bar{0}\end{bmatrix}\begin{bmatrix}R_R|T_R\\0|1\end{bmatrix}\bar{X_k})||_2+||(\bar{x_{6k}}-K_6\begin{bmatrix}I|\bar{L}\end{bmatrix}\begin{bmatrix}R_R|T_R\\0|1\end{bmatrix}\bar{X_k})||_2.\tag{4.15}$$

Note that (4.13) consists of 29 variables including 12 parameters for left and right transformation matrix $(\theta_1,\theta_2,\theta_3,t_X,t_Y,t_Z)$ , 12 parameters for camera centers $(u_x^i,u_y^i)$ , i = 1, 2, 3, 4, 5, 6 and 5 parameters for focal lengths $f_i$, i = 1, 2, 3, 5, 6. Minimizing (4.13) yields the unknowns from which 3D point cloud can be reconstructed.

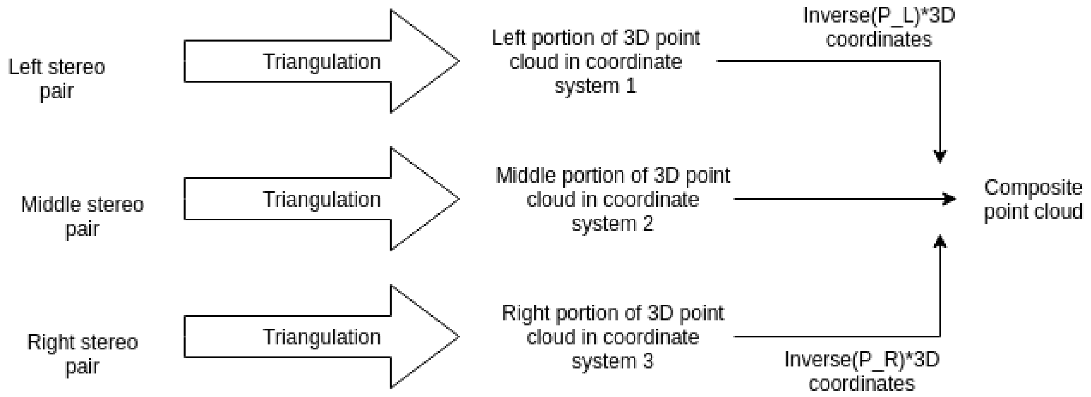

Left stereo pair
Triangulation
Left portion of 3D point cloud in coordinate system 1
Inverse(P_L)*3D coordinates
Middle stereo pair
Triangulation
Middle portion of 3D point cloud in coordinate system 2
Composite point cloud
Right stereo pair
Triangulation
Right portion of 3D point cloud in coordinate system 3
Inverse(P_R)*3D coordinates

Figure 4.3: Dense 3D reconstruction.

## 4.2 Dense 3D Reconstruction

Now we elaborate the advantage of minimizing the cost function defined in (4.13). To see this, note that minimizing equation (4.13) yields the left and right transformation matrix which defines the relation of coordinate system 2 with coordinate system 1 and coordinate system 3. One can get three different 3D point clouds using triangulation in coordinate systems 1, 2 and 3 for 3 different stereo pairs. These three point clouds provide the 3D information of the left, middle and right part of the face in 3 different coordinate systems respectively coordinate systems 1, 2 and 3. Now we bring together all three point clouds in a single coordinate system (coordinate system 2) using the left and right transformation matrix. To achieve this, we first perform the triangulation for each stereo pair then 3D points in coordinate systems 1 and 3 are transformed into coordinate systems 2 by multiplying them with the inverse of the left and right transformation matrices as shown in Fig. 4.3.

## 4.3 Experimental Results

In this section we present the result of dense 3D reconstruction. We use the earlier data set described in 3.4. The images were captured by three stereo pairs but earlier we treated 4 of them as mono. We get 773 common points across six images which were used in optimization. The solver consists of `GA` and `sqp`. ASIFT provides the dense correspondences between stereo pairs as shown in Fig. 4.4 . Triangulating the stereo correspondence between individual stereo pairs provides the information about 3 different parts of the face as shown in Figure 4.5. Finally, we bring all 3 point clouds together in a single coordinate system (coordinate system 2). Fig. 4.7 shows some of the different views of the composite 3D point cloud.

In addition to this, a surface fitted on this point cloud is also shown in Fig. 4.6. We now compare the few key facial measurements with the reconstructed point cloud as shown in Table 4.1. The reconstructed 3D point cloud is true-to-scale and also recovers the shape faithfully.

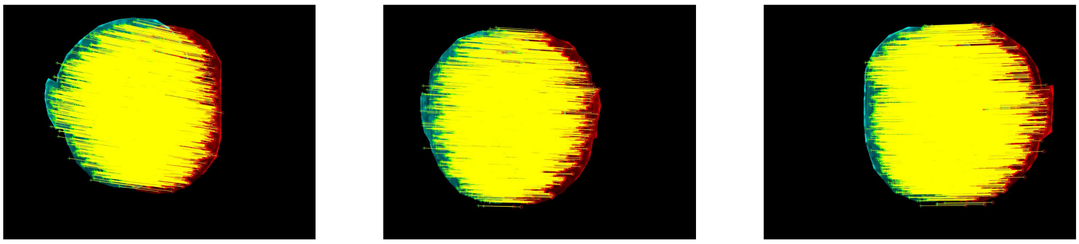

(a) Left stereo pair (b) Middle stereo pair (c) Right stereo pair

Figure 4.4: Dense point correspondences between stereo pairs.

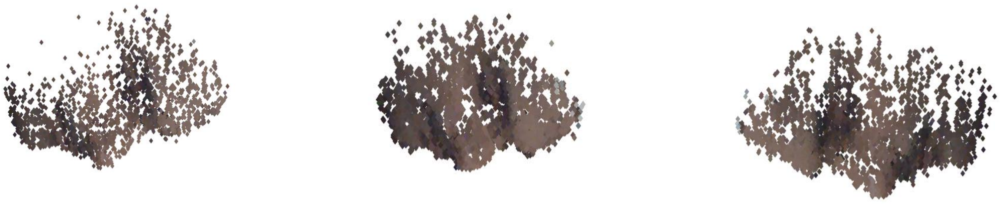

(a) Left stereo view. (b) Middle stereo view. (c) Right stereo view.

Figure 4.5: Stereo views of 3D point cloud with texture information.

Table 4.1: Key facial measurements

| Facial measurement | 3D point cloud (mm) | Original face (mm) | error | |
|---:|---:|---:|---:|---:|
| | | | mm | % |
| Lip width | 55.41 | 55 | 0.41 | 0.973 |
| Nose length | 51.50 | 50 | 1.5 | 2.91 |
| Nose height | 15.2 | 15 | 0.2 | 1.32 |
| Eye socket length | 39.71 | 38 | 1.71 | 4.3 |
| Nose to upper lips | 31.25 | 30 | 1.25 | 4 |
| Z-span | 100 | 98 | 2 | 2 |

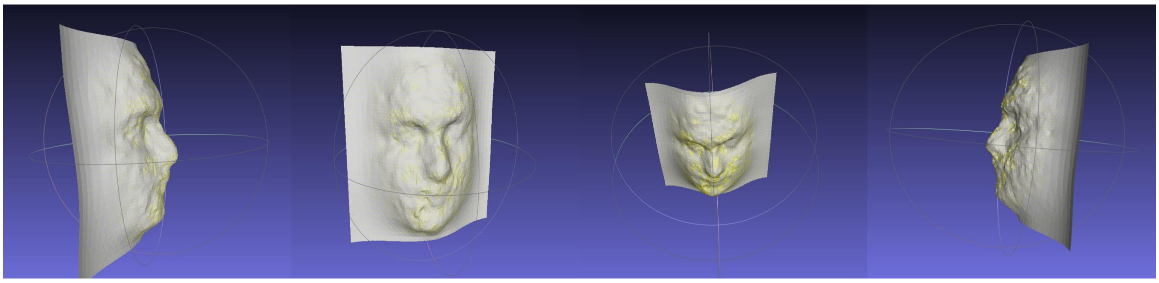
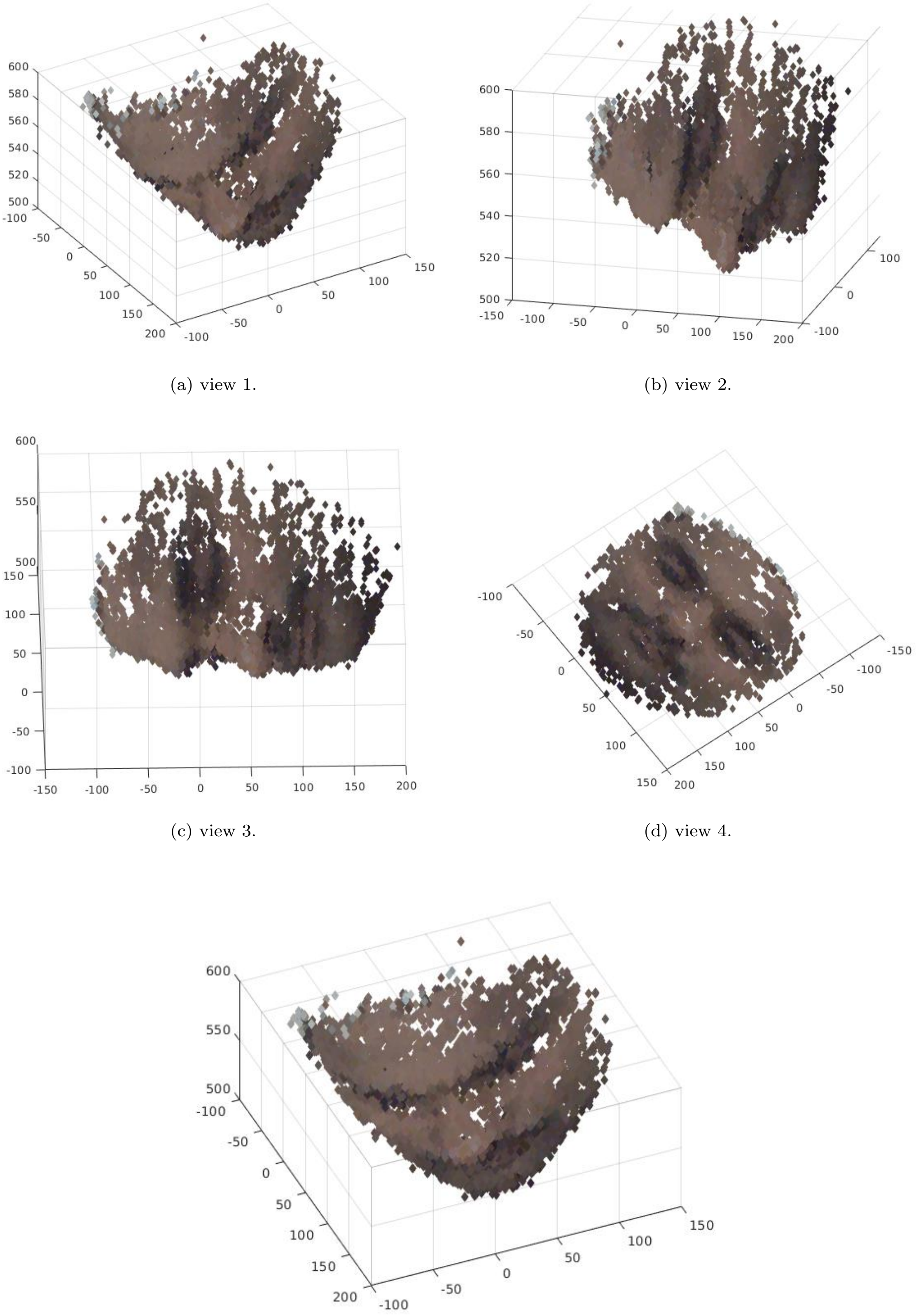

(a) view 1. (b) view 2. (c) view 3. (d) view 4.

Figure 4.6: Different views of 3D mesh visualized in Meshlab.



(a) view 1.

(b) view 2.

(c) view 3.

(d) view 4.

(e) view 5.

Figure 4.7: Different views of composite 3D point cloud with texture information.

# Chapter 5

# Summary and Discussion

In this thesis, we presented preliminary results on true-to-scale 3D reconstruction of the human face using its multiple 2D images. The images were acquired using a network six cameras consisting of three stereo pairs. An easy-to-use 3D face acquisition system is useful in a variety of situations beyond teleophthalmology. For example, many such inexpensive systems can be used to create a national-level database of 3D faces of citizens. Such database would help in law enforcement.

While we report our experiments and results for one human test subject in this thesis, the arrangement was found to be suitable for multiple test subjects of similar physical characteristics such as height and size of their head. Thus, one need not necessarily adjust the pose (extrinsic parameters) and the focal lengths (intrinsic parameters) of the cameras for every test subject. Furthermore, the depth-of-field of cameras used in this work was found to be sufficient so that the entire face was in focus. In this backdrop, we believe that the required array of cameras can also be built using inexpensive off-the-shelf USB cameras. If one manages to create reliable stereo pairs using such cameras and achieve near-instantaneous acquisition via a USB hub, the imaging setup can potentially be stored and carried in a backpack. The setup can be supplemented with a look-up table of recommended camera heights and camera-to-subject distances so that the face to be acquired remains focused and within the field-of-view of the cameras.

The total floor area occupied by our image acquisition system was 7 × 7 sq. ft. Interestingly, the required floor area can be easily managed inside a bus. Such a possibility complements the existing efforts towards mobile eyecare. In several resource-constrained situations where setting up primary healthcare centers is not possible, mobile eyecare solutions have been proposed. In particular, a few trained personnel can travel to remote villages using a vehicle such as a bus equipped with basic equipment to perform preliminary checkups. The proposed image acquisition system can potentially be accommodated in such a vehicle. This approach — at the least — solves the issue of unavailability of reliable electricity and poor lighting conditions in such remote areas.

Although promising, the success of the aforementioned proposals relies on the robustness of the 3D reconstruction algorithm against the imperfections in the acquisition setup and in the acquired images. First, the ad hoc stereo pair created from two discrete cameras deviates may not exhibit the ideal characteristics such as parallel optical axes and precisely known baseline length. Characterization of such imperfections has been discussed in several works such as Huang *et al.* [21] and Tamboli *et al.* [22]. Further, projection of bright structured light patterns — although for very short period — could be uncomfortable for some subjects. To this end, one can consider using infrared (IR) structured light patterns along with IR cameras. To the best of our knowledge, such a multicamera system containing both normal and IR cameras remains unexplored in terms of usability, cost and maintenance requirements. Several face scanning systems are available which contain at least one IR projector, one IR camera and one normal camera. However, such systems, as alluded earlier, put an undue burden on the test subjects.

Several 3D face reconstruction systems exist that offer practical results albeit requiring some prior knowledge [23], access to range-scanned 3D point clouds of faces [24], stereoscopic camera combined with infrared projection and sensing where either the subject or the camera needs to be moved [25] etc. It is important to note that the proposed system aims to minimize burden on the subject and does not use any prior information. In the future, we shall use some of the techniques used in aforementioned works, with the primary aim of convenient and instantaneous acquisition.

In the future, we plan to fit parametric models on the reconstructed 3D point clouds. For individual facial features such as nose, cheeks, forehead, etc., we expect the model parameters to be similar across a group of population. Knowledge of such parameters can aid research in anthropology.

# References

[1] B. S. Chandra, "Towards Next-Generation Cardiac Care in Resource-Constrained and High-Risk Scenarios," Ph.D. dissertation, Indian Institute of Technology Hyderabad, 2019.

[2] World Health Organization, *Telemedicine: opportunities and developments in member states. Report on the second global survey on eHealth.*, 2010.

[3] Directorate of Economics and Statistics, Planning Department, Government of Andhra Pradesh, *Performance Appraisal and District Economic Scenario, 2017-18*, 2018.

[4] Department of Planning, Programme Monitoring and Statistics, Government of Karnataka, *Economic Survey of Karnataka 2017-18*, 2018.

[5] Planning and Coordination Department, Directorate of Economics and Statistics, Government of Odisha, *Odisha Economic Survey 2014-15*, 2015.

[6] Planning Department, Government of Telangana, *Socio Economic Outlook 2018*, 2018.

[7] "The LVPEI Eye Care Network," http://www.lvpei.org/assets/images/services/icare/ICARE_Training_Brochure.pdf, accessed: Jun. 2019.

[8] K. K. Vupparaboina, R. R. Tamboli, S. Manne, A. Richhariya, M. Naik, and S. Jana, "Oculofacial surgical planning: true-to-scale 3D feature quantification using multicamera network," in *IEEE Annual India Conference*, 2016.

[9] K. K. Vupparaboina, K. Raghavan, and S. Jana, "Smart camera networks: an analytical framework for auto calibration without ambiguity," in *IEEE Recent Advances in Intelligent Computational Systems*, 2013, pp. 310–315.

[10] ——, "Euclidean auto calibration of camera networks: baseline constraint removes scale ambiguity," in *Twenty Second National Conference on Communication*, March 2016, pp. 1–6.

[11] K. K. Vupparaboina, "Perspectives on imaging and image analytics in ophthalmology," Ph.D. dissertation, Indian Institute of Technology Hyderabad, 2017.

[12] K. K. Vupparaboina, R. R. Tamboli, S. Manne, P. A. Kara, M. G. Martini, A. Barsi, A. Richhariya, and S. Jana, "Towards true-to-scale 3D reconstruction of the human face using structured light projection and off-the-shelf cameras," in *International conference on 3D immersion*, Dec. 2018, pp. 1–7.

[13] C. Boehnen and P. Flynn, "Accuracy of 3d scanning technologies in a face scanning scenario," in *Fifth International Conference on 3-D Digital Imaging and Modeling*, 2005, pp. 310–317.

[14] "acA1300-30gc - Basler ace," https://www.baslerweb.com/en/products/cameras/area-scan-cameras/ace/aca1300-30gc/, accessed: Jun. 2019.

[15] "Dell M110 Ultra-Mobile Projector," https://www.dell.com/is/business/p/dell-m110/pd, accessed: Jun. 2019.

[16] "DES1016D 16-Port Fast Ethernet Unmanaged Desktop Switch," https://eu.dlink.com/uk/en/products/des-1016d-16-port-10-100mbps-desktop-switch, accessed: Jun. 2019.

[17] A. C. Bovik, *The essential guide to image processing.* Academic Press, 2009.

[18] G. Yu and J.-M. Morel, "ASIFT: An Algorithm for Fully Affine Invariant Comparison," *Image Processing On Line*, vol. 1, pp. 11–38, 2011.

[19] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 3, p. 29, 2013.

[20] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia, "MeshLab: an Open-Source Mesh Processing Tool," in *Eurographics Italian Chapter Conference*, V. Scarano, R. D. Chiara, and U. Erra, Eds. The Eurographics Association, 2008.

[21] L. Huang, Q. Zhang, and A. Asundi, "Flexible Camera Calibration using Not-measured Imperfect Target," *Applied Optics*, vol. 52, no. 25, pp. 6278–6286, 2013.

[22] R. R. Tamboli, K. K. Vupparaboina, S. Manne, P. A. Kara, A. Cserkaszky, M. G. Martini, A. Richhariya, and S. Jana, "Towards Euclidean Auto-calibration of Stereo Camera Arrays," in *SPIE Optical System Alignment, Tolerancing, and Verification XII*, 2018.

[23] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1031–1039.

[24] A. Brunton, A. Salazar, T. Bolkart, and S. Wuhrer, "Review of Statistical Shape Spaces for 3D Data with Comparative Analysis for Human Faces," *Computer Vision and Image Understanding*, vol. 128, pp. 1–17, 2014.

[25] "Bellus3d: High-quality 3d face scanning," https://www.bellus3d.com/, accessed: Jun. 2019.