3D RECONSTRUCTION FROM STEREO IMAGES OF MOBILE PHONES

Thesis submitted to

Visvesvaraya National Institute of Technology, Nagpur

In partial fulfilment of requirement for the award of degree of

Bachelor of Technology

In

Electronics and Communication Engineering

By

Krishna SumanthGundluru **BT12ECE020**

Tushar H. Jagdale **BT12ECE024**

Guide

Dr.Saugata Sinha

Department of Electronics and Communication Engineering

Visvesvaraya National Institute of Technology

Nagpur-440010(India)

April 2016

VISVESVARAYA NATIONAL INSTITUTE OF TECHNOLOGY, Nagpur-440010

(INDIA)

DECLARATION

I, hereby declare that the thesis titled "3D RECONSTRUCTION FROM STEREO IMAGES OF MOBILE PHONES" submitted herein has been carried out by me in the Department of Electronics Engineering of Visvesvaraya National Institute of Technology, Nagpur. The work is original and has not been submitted earlier as a whole or in part for the award of any degree / diploma at this or any other Institution / University.

Krishna SumanthGundluru
BT12ECE020

Tushar H. Jagdale
BT12ECE024

Electronics and Communication Engineering
V.N.I.T., NAGPUR

VISVESVARAYA NATIONAL INSTITUTE OF TECHNOLOGY, Nagpur-440010

(INDIA)

CERTIFICATE

This is to certify that the thesis

"3D RECONSTRUCTION FROM STEREO IMAGES OF MOBILE PHONES"

has been successfully completed in the session 2015-2016 in a

satisfactory manner as a fulfilment of the requirements of the degree of

"**Bachelor of Technology in Electronics and Communication Engineering**"

Dr.Saugata Sinha

Project guide

Dr.A.S Gandhi

Head
Electronics and Communication Engineering,
V.N.I.T., NAGPUR

ACKNOWLEDGEMENT

This project is the product of the collective effort put in by many people and we take this opportunity to acknowledge their contributions.

First and foremost, we would like to thank our guide Dr.Saugata Sinha, who has been of immense help and without whose valuable guidance this project would not have been possible. His persistent encouragement and motivation along with his profound insight inspired us to achieve our best. Our gratitude for his timely solutions for the problems we faced during the course of this work is boundless. We are also highly grateful to him for providing us with all the possible facilities required for the successful completion of the project.

We would also like to thank Dr.A.S.Gandhi, Head of Electronics and Communication

Engineering Department for providing us with all the required facilities as and when needed to successfully complete the project. We would also like to thank the teaching and non-teaching staff of the department for aiding us in this endeavour.

ABSTRACT

This study deals with understanding and analysis of 3D Reconstruction from stereo images using mobile phones. The ability to perform fast and accurate 3 dimensional reconstruction of an environment or object is central to many areas of computer vision processing. 3D face recognition, parts inspection, autonomous drivers and a near infinite number of other applications have driven research in 3D reconstruction forward for the last decade.

But as everybody carries a camera nowadays by virtue of owning a cell phone, but few of these devices capture the three-dimensional contours of objects like a depth camera can.

Depth cameras are quickly gaining prominence for their potential in pocket-sized devices, where the idea is that if our phones capture the contours of everything from street corners to the arrangement of your living room, developers can create applications ranging from better interactive games to helpful guides for the visually impaired. A 3D Selfie and photograph are few where this could be applied in the entertainment field.

As reconstruction through mobile phones is gaining importance and ways to reduce computational complexity supported by processors are being developed, our aim of this project is to check if the existing mobile phones when placed as a stereo system produce appropriate reconstruction.

Checking the accuracy of the reconstructed images is one part and its effect on reconstructed objects by varying baseline is the second aim of the project. For this several setups were taken into consideration. The experiments were conducted indoors as well as outdoors. For analysis, Scene Reconstruction Example from MATLAB is employed.

Results showed that for a fixed baseline, accurate reconstruction took place between a particular range. As baseline is increased, objects at a greater distance are reconstructed appropriately. But this is true only for a limited range of baseline. As baseline gets further large, reconstruction isn't possible since stereo matching between images reduces and for smaller baselines the objects get distorted.

Self Calibration and Auto-focus are the main challenges that have to be worked upon in this reconstruction process which would enable the system to work in a dynamic environment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 3D RECONSTRUCTION?

In computer vision and computer graphics, 3D reconstruction is the process of capturing the shape and appearance of real objects.

1.2 IMPORTANCE

The possibility of obtaining high accuracy 3D information from 2D images has been a highly active and fruitful area of research over the past decade. The rising attention has been fuelled by promising application development in areas such as architectural conservation, scene of crime analysis, architectural design, movie post processing, and face recognition and in a multitude of additional research domains. A second motivational factor has without a doubt been the exponential rise in computing power and graphics processing ability of recent years. Whilst many of the geometric and theoretical elements essential to 3D reconstruction have been widely known for many years it is only recently that available computing power has made practical implementations of fast stereo matching and projection a possibility. Many researchers have therefore been working towards the development of robust and efficient methods of performing 3D reconstruction from 2D imagery.

The ability to extract information about the world in which a computer is immersed is without a doubt of fundamental importance to a wide range of commercial and academic interests. Much previous research has focussed on extracting information from 2D images and this approach has led to a wide range of successes, however, the wealth of information to be found in the third dimension is proving an ever more seductive lure for the computer vision researcher. As more development effort is focused on the 3D reconstruction problem it is becoming increasingly more important to efficiently organise approaches to reconstruction both to aid the development process and to allow a greater degree of comparison between varying reconstruction strategies.

In situations where 2D data insufficiently represents an object requiring analysis the ability to construct a 3D model is essential. Architectural and artistic preservation and cataloguing are a good example of fields where such 3D data has found many useful applications. Statues and buildings which may have significant artistic value require preservation but are often exposed to the elements. In order to preserve such buildings and sculptures for future generations, or to provide a catalogue of work for people without direct physical access, a photograph is

clearly insufficient. Highly accurate 3D models are of massive importance in such a scenario. Theoretically a sufficiently accurate 3D model could be used to reconstruct a building or work from scratch or, more likely, to carry out repairs when they become a necessity. Currently it is more likely the data will simply be used for cataloguing and reference purposes, however, it is clear that a 3D model is of significantly greater use than a simple photograph in such cases.

(a) Without Pre-Processing

FIGURE 1: 3D RECONSTRUCTION FOR ARCHITECTURAL PRESERVATION

A second example where 3D data proves superior to its 2D counterpart can be observed during the part inspection process used in many manufacturing plants. Parts inspection involves verifying the accuracy of a manufactured part to a given set of tolerances. This is usually achieved by constructing a model (using either a multi-view or laser based approach) and then registering directly to the CAD model of a given part and measuring the resultant error. Manufacturing, until recently, used touch sensors for parts verification, however, the faster speed and efficiency acquired through the use of visual sensors has led to such methods becoming more dominant within the industry. Thus vision based 3D reconstruction allows a cheaper and more passive approach to parts inspection requiring only the original designs for a given part to allow its manufacturing error to be determined. Furthermore the vision based system is far more adaptable than other approaches; allowing the measurements of different part types without significant reconfiguration.

A near infinite number of other applications have driven research in 3D reconstruction forward.

1.3 COMPARISION OF METHODS USED FOR RECONSTRUCTION

1.3.1 ACTIVE METHODS

Active methods, i.e. range data methods, given the depth map, reconstruct the 3D profile by numerical approximation approach and build the object in scenario based on model. These methods actively interfere with the reconstructed object, either mechanically or radio-metrically using rangefinders, in order to acquire the depth map, e.g. structured light, laser range finder and other active sensing techniques.

An insight into different depth Cameras has been illustrated below.

1. The T-O-F Cameras (RF-modulated light sources with phase detectors/Range gated imagers/Direct Time-of-Flight imagers)-Scanner less Depth Cameras

They illuminate the scene with infrared light and measure the Time-of-Flight. There are two operation principles: pulsed light and continuous wave amplitude modulation. The earlier comes with the problem of measuring very short time intervals in order to achieve a resolution which corresponds to a few centimeters in depth (e.g. ZCam of 3DV Systems). The continuous wave modulation approach avoids this by measuring the phase shift between emitted and received modulated light which directly corresponds to the Time-of-Flight and in turn to the depth, where ambiguities in form of multiples of the modulation wavelength may occur. For a long time the ToF imaging sensors suffered two major problems: a low resolution and a low sensitivity resulting in high noise levels. Additionally, background light caused problems, when used outdoors.

2. Laser Scanner Depth Cameras

Other depth cameras or measuring devices, such as laser scanners or structured light approaches, were not able to provide (affordably) high frame rates for full images with a reasonable resolution and not e.g. lines.

In addition to low accuracy capture devices, sensors which take excessive time during the capture phase are more suited to reconstructions of static objects rather than the decidedly dynamic human face. The former of these constraints rules out SONAR and similar capture

systems, the latter rules out laser range finders and other such slow speed reconstruction techniques.

3. The KINECT

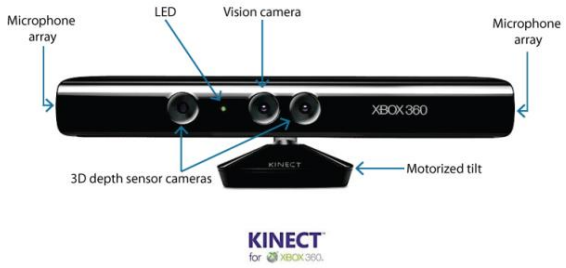


FIGURE 2: KINECT DEPTH SENSOR

This was true until in 2010 Microsoft (PrimeSense) released the Kinect. Instead of using a time varying pattern as was widely applied previously, it uses a fixed irregular pattern consisting of a great number of dots produced by an infrared laser led and a diffractive optical element. The Kinect determines the disparities between the emitted light beam and the observed position of the light dot with a two megapixel grayscale imaging chip. The identity of a dot is determined by utilizing the irregular pattern. Here it seems that the depth of a local group of dots is calculated simultaneously, but the actual method remains a secret up until today. Once the identity of a dot is known the distance to the reflecting object can be easily triangulated. In addition to the depth measurements, the Kinect includes a color imaging chip as well as microphones.

1.3.2 PASSIVE METHODS

Passive methods of 3D reconstruction do not interfere with the reconstructed object; they only use a sensor to measure the radiance reflected or emitted by the object's surface to infer its 3D structure through image understanding.In this case we talk about image-based reconstruction and the output is a 3D model.

1. Monocular cues methods

Monocular cues methods refer to use image (one, two or more) from one viewpoint (camera) to proceed 3D construction. It makes use of 2D characteristics(e.g. Silhouettes, shading and texture) to measure 3D shape, and that’s why it is also named Shape-From-X, where X can be silhouettes, shading, texture etc. 3D reconstruction through monocular cues is simple and

quick, and only one appropriate digital image is needed thus only one camera is adequate. Technically, it avoids stereo correspondence, which is fairly complex.

Shape-from-shading: Due to the analysis of the shade information in the image, by using Lambertian reflectance, the depth of normal information of the object surface is restored to reconstruct.

Photometric Stereo:This approach is an update of Shape-of-shading. Images taken in different lighting conditions are used to solve the depth information. It deserves to be mentioned that more than one images are required by this mean.

Shape-from-texture: Suppose such an object with smooth surface covered by replicated texture units, and its projection from 3D to 2D cause's distortion and perspective. Distortion and perspective measured in 2D images provide the hint for inversely solving depth of normal information of the object surface.

2. Binocular stereo vision

FIGURE 3: STEREO CAMERAS

Binocular Stereo Vision obtains the 3-dimensional geometric information of an object from multiple images based on the research of human visual system. The results are presented in form of depth maps. Images of an object acquired by two cameras simultaneously in different viewing angles, or by one single camera at different time in different viewing angles, are used to restore its 3D geometric information and reconstruct its 3D profile and location. This is more direct than Monocular methods such as shape-from-shading.

Binocular stereo vision method requires two identical cameras with parallel optical axis to observe one same object, acquiring two images from different points of view. In terms of trigonometry relations, depth information can be calculated from disparity. Binocular stereo vision method is well developed and stably contributes to favorable 3D reconstruction,

leading to a better performance when compared to other 3D construction. Unfortunately, it is computationally intensive, besides it performs rather poorly when baseline distance is large.

1.4 SHORT-COMINGS AND DIFFICULTIES

Biology solves the 3D reconstruction problem through a combination of sophisticated sensors (the eyes) and the advanced pattern recognition powers of the brain. Depth perception is important to the survival of many species of animal, particularly predators, and thus the biological model may be of use if parallels can be drawn between natural and computer vision approaches to depth perception. The most obvious of nature's solutions to depth perception is the evolution of two eyes to facilitate stereo vision. Both the differing data reaching each eye and the relative focus between eyes is used to create a sense of depth. A second technique for determining depth utilised by many animals involves using parallax motion data as a depth cue. Again a variety of research has been carried out to enable computers to obtain depth data via parallax motion. In addition to visual data animals have a variety of auxiliary senses from which depth data can be extracted. Even humans with their relatively poor hearing can infer limited information about the size and nature of their environment without the use of their eyes. A combination of complex factors which lie outside the scope of this thesis allow this to occur, however, it is clear that a range of senses combine to give animals and humans a 3D representation of their environment.

The final biological method for building an internal model of the world is also perhaps the most difficult to translate into a structured computer vision approach. Through years of collected experience a human builds up large amounts of a posteriori knowledge about their environment and as such can use inductive reasoning to estimate 3D data. For example a human looking at a house, partially occluded by a tree in the foreground, could provide a reasonable estimate of the structure of the occluded section of the house. Except in highly constrained situations, such a feat is difficult to reproduce in the world of computers and would likely require significant progress in strong AI before computer vision could achieve significant results in this area.

The final difficulty in attempting to reproduce the biological model for depth perception in software involves the manner in which the brain combines and utilises all its available senses simultaneously in order to produce a consistent model of the environment, in contrast such sensor integration is difficult in software. Current research tends to focus on a single method

for reconstruction at a time although some work does attempt to combine multiple data sources it is nowhere near the complexity achieved in nature.

1.5 OBJECTIVES

1. To describe an approach to representing the different conceptual approaches to reconstruction in a consistent and practical manner.
2. Mainly focussed on individual components of a system such as the stereo correlation process and thus would benefit dramatically from their expansion.
3. Aims to indicate a link between current research supported by high end processors and a future, where computations would be reduced drastically thus making it possible to reconstruction algorithms with pocket sized devices
4. To check if the existing mobile phones when placed as a stereo system produce appropriate reconstruction in terms of accuracy and to exhibit the effect of baseline on such systems.

CHAPTER 2 LITERATURE REVIEW

The majority of existing methods for live 3D surface reconstruction require active depth sensors and/or high end GPUs. This limits object scanning scenarios to PCs or high end laptops and tablets, with custom hardware.

But as everybody carries a camera nowadays by virtue of owning a cell phone, but few of these devices capture the three-dimensional contours of objects like a depth camera can.

Depth cameras are quickly gaining prominence for their potential in pocket-sized devices, where the idea is that if our phones capture the contours of everything from street corners to the arrangement of your living room, developers can create applications ranging from better interactive games to helpful guides for the visually impaired.

The literature review provides a study of approaches to 3D reconstruction on mobile phones. The primary concern of this thesis is to develop a comprehensive understanding of full reconstruction process. As such the focus of the literature review will be toconsider research which closely matches this main theme. Several frameworks existdescribing various aspects of the reconstruction process and these are given primaryconsideration and an in-depth analysis. Perceived shortcomings within the current researchare addressed and form the basis for the work carried out .

A secondary concern of the thesis is to extensively study the requirements of a complete 3D Reconstruction system in the mobile phones. This topic covers a great deal of ground within the computer vision field of research. Literature is reviewed concerning the research in order to develop a complete understanding of the factors affecting and attributing to the accuracy of both the reconstruction and recognition process. The literature review is in accordance with the most influential and relevant papers critically analyzed and their applicability to this thesis defined.

A number of the required processes to achieve 3D reconstruction contain critical problems which are not yet completely solved by the proposed approaches.

Our results stresses on to what constraints one should take care while using mobile phone for 3D reconstruction using stereo vision such as the baseline distances to be maintained, up

towhat amount scene could be reconstructed with only stereo ,how far from the cameras should the objects lie for better results.

Two recent approaches are considered which include Google’s Project Tango and Microsoft's Mobile Fusion

1) Project Tango-Google

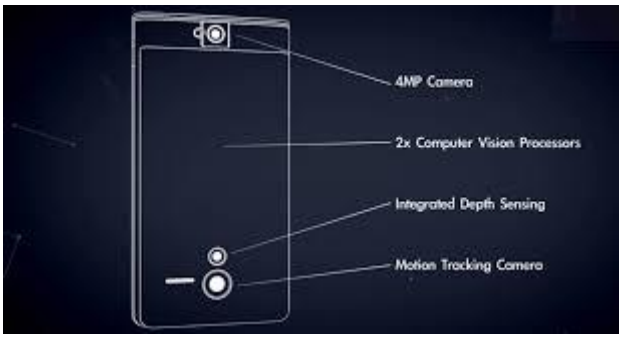


FIGURE 4: GOOGLE PHONE WITH DEPTH CAMERA

Project Tango is a Google technology platform that uses computer vision to enable mobile devices, such as smartphones and tablets, to detect their position relative to the world around them without using GPS or other external signals. This allows application developers to create user experiences that include indoor navigation, 3D mapping, measurement of physical spaces, and recognition of known environments, augmented reality, and windows into virtual 3D worlds. The first product to emerge from Google's ATAP skunkworks group, Project Tango was developed by a team led by computer scientist Johnny Lee, a core contributor to Microsoft's Kinect. Whilst work, such as Google’s Tango project have begun to explore more integrated solutions for mobile depth camerahardwarei.e. they are adding depth cameras into mobile gadgets,, the cost, form-factor, and power considerations have yet to make such devices ubiquitous.

2) Mobile Fusion-Microsoft

Microsoft shows that with some simple modifications and machine-learning techniques an ordinary smartphone camera or webcam can be used as a 3-D depth camera.

FIGURE 5: WITH A FEW HARDWARE CHANGES, SUCH AS A RING OF NEAR-INFRARED LEDS, A MICROSOFT LIFECAM IS ADAPTED TO WORK AS A DEPTH CAMERA.

In addition, they present a new system for making real-time scanning of 3D surface models even cheaper and ubiquitous, using mobile phones we already have in our pockets, without any hardware modification. Existing state of the art mobile methods approximate surfaces using simple visual hull constraints, Delaunay triangulation, or purely point based representations or use cloud-based processing to avoid on-device computation. Instead their system reconstructs real-time 3D surface models directly on the device, allowing lightweight capture of detailed 3D objects, at speeds that have yet to be demonstrated byother systems. They describe their pipeline in full, emphasizing components for uncertainty-aware dense model tracking, robust key-frame selection, dense stereo, volumetric fusion and surface extraction, and their efficient mobile phone implementation, which allows for 25Hz performance and also compared the accuracy of our method quantitatively against a consumer depth camera baseline using a new dataset of 3D reconstructions which we make public.

As reconstruction through mobile phones is gaining importance and ways to reduce computational complexity supported by processors are being developed, our aim of this project is to check if the existing mobile phones when placed as a stereo system produce appropriate reconstruction.

Our approach has been processing the images from mobile in MATLAB, Checking its accuracy, effect on reconstructed objects for varying baseline, indoor vs outdoor reconstruction and disparity map comparison.

This approach could be applied as follows:

- With few hardware changes as to add a camera in the existing phones with an appropriate baseline
- Without any hardware change aligning 2 mobile phones and connecting them with an appropriate software.

CHAPTER 3 MATHEMATICS OF 3D RECONSTRUCTION

This chapter summarises the most important mathematical concepts and algorithms contained within the thesis. Discussing in general terms the vast amounts of mathematical material related to calibration, stereo matching, deformable models and 3D projection would merely reproduce research which can easily be obtained elsewhere. Therefore this chapter considers mathematical approaches to reconstruction that are directly applicable to the reconstruction system implementation described in chapter 5 and topics required to fully understand the later chapters of the thesis. The scope of this chapter covers methods relating to 3D reconstruction and specifically to camera models, the geometry of multiple views and camera calibration algorithms. Thus this chapter describes the essential mathematical components of 3D reconstruction. Particular consideration is given to the geometric and algebraic constraints that hold throughout multiple views, as represented by the fundamental matrix, as well as the mapping from world space to image space as defined by the camera projection matrix. Finally the chapter considers how the combination of calibrated cameras and a set of points correlated across multiple views can be used to compute 3D information from multiple 2D images.

Section 3.1 provides an introduction to the concepts behind the various strata of geometries important to reconstruction. Depending on the specific reconstruction requirements certain assumptions about calibration and reconstruction can be relaxed. This section considers the differences between the various geometries as well as the requirements for achieving reconstruction to a particular geometric level. Following this introduction to the fundamental geometric principles required for 3D reconstruction section 3.2 considers the mathematical approach to camera calibration. The algorithms presented here outline methods for determining the linear mapping between coordinates in 3-space and their 3D projection on the camera image plane. Particular attention is given to the Direct Linear Transform (DLT) since this technique forms the basis for calculating the desired projective mappings. This section also shows a step by step description for the gold standard method of performing camera calibration which is utilised later in the thesis in the form of a practical implementation.

The geometry of multiple views and the constraints this places on the imaging process are central to the multi-view reconstruction methods presented in this thesis. Section 3.3

describes some of the constraints and how they can be utilised to aid both the stereo correlation process and reconstruction. This section also considers the fundamental matrix and demonstrates how it describes the geometry of two views and the relation between image mappings in multiple view planes. The final section of this chapter describes the manner in which the calibration data and stereo correlation data can be combined in order to compute the depth of points within a scene. This section also makes use of the DLT in order to compute mappings from two dimensions into three.

Whilst this section does not aim to be a comprehensive guide to the mathematical processes behind 3D reconstruction it should provide sufficient information to appreciate the algorithms described elsewhere in the thesis, as well as to provide understanding of some of the difficulties faced by 3D reconstruction systems. Much of the information in this chapter is readily available from other sources however by replicating some of that information here it is hoped that the thesis provides a more thorough analysis of the complete reconstruction process in addition to providing interested researchers a starting point for the mathematical concepts involved.

3.1 CAMERA MODELS AND THE IMAGING PROCESS

3D reconstruction is heavily based on the geometric properties of the imaging process. The imaging process has close ties with the four differing types of geometry and the relationship between the geometries. As show below, starting with projective geometry each subsequent type is a subset of the previous level.

Projective → Affine → Metric → Euclidean

Projective geometry is the least structured and the simplest whereas Euclidean is the most structured but also the most complicated. The world is usually perceived as a Euclidean 3D space however the difference between perception and the actual image arriving at the retina is very real. Despite this the brain is able to formulate an internal Euclidean representation of the world from the projective vision process. The process by which a camera views the world is equivalent to the biological model with respect to the geometric properties involved. The remainder of this section deals with the concepts of varying geometry types and how such concepts affect image formation and in turn the reconstruction process.

In Euclidean geometry the sides of objects have lengths, intersecting lines have defined

angles between them and two lines are parallel if they lie in the same plane but never intersect. Furthermore these properties are all invariant to transformations for translation, rotation and scale. Initially the need for additional geometries may not be clear, however, considering the imaging process of a camera it is obvious that Euclidean geometry is insufficient as now lengths and angles are no longer preserved and parallel lines may intersect.

Euclidean geometry is actually a subset of projective geometry, with affine and metric geometries in between the two as shown above. As the number of invariants for a particular geometry increase the number of possible transforms at a given geometric level decreases. Thus whilst Euclidean geometry only allows for transformations in rotation and translation in order to allow the invariants to remain invariant, projective geometry has a much wider range of available transforms at the expense of fewer invariants. The allowance of additional transforms is what enables projective geometry to better represent the camera imaging process. Specifically the application of the projective transform is what enables the modelling of the imaging process in this respect. Projective transforms preserve type (points remain points and lines remain lines), incidence (whether a point lies on a line) and the cross ratio (ratios of distance). In keeping with Euclidean geometry, projective geometry can exist in any number of dimensions. P2 is equivalent to a plane in Euclidean space and P3 is related to 3D Euclidean space. The imaging process is a projection from P3 to P2; a projection from three dimensional space to the two dimensional image plane.

A projective reconstruction of a cube may differ from its Euclidean representation since the concepts of parallelism, angles and length are not preserved. As a result any projective reconstruction which preserves the edges of the cube is equally valid. Thus in the projective case a number of representations are equivalent to the Euclidean representation with no way to determine the correct Euclidean case. The additional constraints imposed by calibrating cameras prior to reconstruction are what allow the upgrade from projective reconstruction to the more accurate and Euclidean representation of the world.

To move upwards through the geometry layers additional concepts are introduced which reduce the number of available transforms but increases the number of invariants. Affine geometry introduces the concept of a plane at infinity. The plane at infinity is defined by the intersection of parallel lines in the scene and therefore affine geometry expands the concepts of projective geometry by introducing the property of parallelism which is invariant to affine

transformations. Thus now the Euclidean cube reconstructed to affine geometry contains all the correct edges with the appropriate edges parallel to each other. Angles are still not invariant however so the reconstruction may still differ significantly from the Euclidean representation.

The introduction of metric geometry provides a representation for the cube which preserves angles in addition to the invariants provided by the earlier levels of geometry. Metric geometry adds to the infinity plane the concept of the absolute conic. The absolute conic is invariant under Euclidean transformations and as such represents a calibration object naturally present in all scenes. The absolute conic is a particular conic which lies on the infinity plane. Metric geometry can therefore be derived from projective geometry by selecting a particular plane as the plane at infinity and specifying a particular conic to be the absolute conic.

The single aspect differentiating Euclidean and metric geometries is the measure of absolute lengths. Euclidean geometry can define the absolute length and size of an object whereas metric geometry has only the concept of ratios between lengths. In order to produce the final upgrade the absolute length of some object in the real world must be known from which it becomes possible to compute the lengths of the remaining objects in the scene.

Many of the geometric concepts discussed here are important to several areas of computer vision. Most notably the concepts of projective geometry are important when considering the imaging process discussed next. Whilst a brief summary of the various geometries is presented here more detailed explanations are widely available [5, 7] which more comprehensively describe the appropriate techniques.

As discussed extensively, one of the principle concepts of 3D reconstruction is the image formation process. This covers the manner in which a 3D scene is mapped onto a 2D view plane and the information that can be determined once the initial 3D scene has been imaged and reduced to two dimensions. The standard method for modelling this drop from 3 to 2 dimensions is performed by central projection whereby a 3D world point is imaged by drawing a ray through the world point to a fixed location in space called the centre of projection. This ray will intersect a specific plane known as the image plane. The point of intersection with the image plane represents the image of the world point. This model is roughly equivalent with the simple camera model where rays of light pass through a lens and causes a reaction on a film (or CCD in digital cameras) at the back of the camera, thus

producing an image of the original point.

As discussed the process of imaging a scene is essentially a mapping from 3D projective world space to 2D projective space. Central projection encompasses this mapping from differing projective spaces and may be represented by a 3X4 matrix, P. This matrix is known as the camera matrix whose action may be expressed in terms of a linear mapping of homogenous coordinates as follows:

$$\begin{pmatrix} x \\ y \\ w \end{pmatrix} = P_{3\times4} \begin{pmatrix} X \\ Y \\ Z \\ T \end{pmatrix} \qquad 3.1$$

Equation 3.1 represents one of the most fundamental processes of this thesis. Obtaining the projective mapping from the real world to the image plane is the goal of the camera calibration process. In addition much of reconstruction process involves attempting to reverse the mapping from 3 to 2 dimensions with the additional information provided by multiple views.

3.1.1 A NOTE ON HOMOGENOUS COORDINATE SYSTEMS

An important factor when considering differing geometries is how coordinates are represented in a given system. A point in Euclidean 2-space is represented by the ordered pair of real numbers (x,y). A homogeneous coordinate of a point introduces a third entity into the pair to form the triple (x,y,1). Conceptually the points (x,y,1) and (2x,2y,2) represent the same point with the extension that all points (kx,ky,k) define the same coordinate. This leads to the idea of equivalence classes where coordinate triples are considered equivalent when they differ by a common multiplier. Given a coordinate triple (kx,ky,k) we may recover the original coordinates simple by dividing by k. The importance of these concepts becomes apparent when we consider the nature of triples with a final coordinate of 0. Obviously this leads to attempting to divide the first two coordinates of the triple by 0, thus leading to an infinite solution, which in turn yields the mathematical concept of points lying at infinity therefore representing projective geometries “ideal points”. Obviously the concept of homogenous coordinates can be expanded to 3 dimensions which give rise to a plane at infinity in much the same way as ideal points are defined. It should now be apparent that the

set of all ideal points on the projective plane constitutes a line, unsurprisingly called the ideal line, in the same way points of a projective 3-space combine to form the ideal plane. This concept can be expanded through higher dimensions in exactly the same manner as desired.

The equations for perspective projection to the image plane are non-linear when expressed in non-homogeneous coordinates, but change to linear problems when represented in homogeneous form. This characteristic is shared by all perspective transformations, not just projection and provides one of the many motivations for the use of homogenous coordinates since, in most situations linear systems are numerically easier to handle than their non-linear equivalent.

3.2 CALIBRATION

This section deals specifically with the mathematics required for the implementation of components discussed in section 5.3. The various available approaches to camera calibration are discussed in more detail in section 4.1. Commonly the full projection matrix is calculated by re-sectioning using corresponding 3D points (Xi) and their image entities (xi). Given a sufficient amount of Xi <xi correspondences the camera matrix P may be calculated. Typically, it is possible to generate a set of known 3D world coordinates and their corresponding image plane entities through the use of a calibration pattern. This process is examined in more detail in section 4.1 with this sections primary concern being to express the mathematical requirements and techniques involved in estimating the camera calibration matrix. The most commonly used camera calibration technique is perhaps the Direct Linear Transform (DLT) method originally reported by Abdel-Aziz and Karara. The DLT method uses a set of control points whose object space/plane coordinates are already known. The control points are normally fixed to a rigid frame, known as the calibration frame. The problem is essentially to calculate the mapping between the 2D image space coordinates (xi) and the 3D object space coordinates (Xi). For this 3D - 2D correspondence the mapping should take the form of a 3x4 projection matrix (P) such that xi = PXi for all i. The Direct Linear Transform and it application to the calibration problem is demonstrated in section 3.2.1.

3.2.1 THE DIRECT LINEAR TRANSFORM

Whilst the DLT algorithm has been extensively utilised for camera calibration it is also a suitable technique for finding linear mappings between any two data sets, given a certain number of corresponding data points between the two. The simplest form of the DLT algorithm is described below, however, it should be evident that the only difference between this method and the 3D case is the dimensionality of the problem. In the 2D case the solution matrix has dimension 3x3 where as the 3D result produces the desired 3x4 projection matrix. The algorithm for the 3D DLT case is described after the 2D case.

The most basic form of the 2D DLT algorithm requires a set of four 2D to 2D point correspondences: xi ÅÆ x'i . The transform is then given by the equation x'i = Hxi . The equation may then be expressed in terms of a vector cross-product: x'i × Hxi = 0. Expressing the transform in terms of a vector cross-product allows a simple linear solution to H to be calculated.

The j th row of the matrix H is denoted by HjT as shown in equation 3.2 shown below:

$$Hx_i = \begin{pmatrix} h^{1T} x_i \\ h^{2T} x_i \\ h^{3T} x_i \end{pmatrix} \qquad 3.2$$

Denoting X'i as (x'i ,y'i , w'i) T the cross-product may be given explicitly as

$$x'_i \times Hx_i = \begin{pmatrix} y'_i h^{3T} x_i - w'_i h^{2T} x_i \\ w'_i h^{1T} x_i - x'_i h^{3T} x_i \\ x'_i h^{2T} x_i - y'_i h^{1T} x_i \end{pmatrix} \qquad 3.3$$

Since hjTxi = Xi T hj for j = 1,2,3, this gives a set of three equations for H which may be written as in the following equation:

$$\begin{bmatrix} 0^T & -w'_i x_i^T & y'_i x_i^T \\ w'_i x_i^T & 0^T & -x'_i x_i^T \\ -y'_i x_i^T & x'_i x_i^T & 0^T \end{bmatrix} \begin{pmatrix} h^1 \\ h^2 \\ h^3 \end{pmatrix} = 0 \qquad 3.4$$

When each of the four coordinates being considered is presented in this form we have a set of equations: Ai h = 0, where A is a 3x9 matrix and h is a 9-vector made up of entries to the matrix H. This equation is linear in the unknown h.

It should be noted that whilst each set of coordinate matches leads to a set of three equations only two of them are linearly independent. Thus, it is standard practice whilst using the DLT algorithm to ignore the third equation whilst solving for H. The set of equations then becomes:

$$\begin{bmatrix} 0^T & -w'_i x_i^T & -y'_i x_i^T \\ w'_i x_i^T & 0^T & -x'_i x_i^T \end{bmatrix} \begin{pmatrix} h^1 \\ h^2 \\ h^3 \end{pmatrix} = 0$$

3.5

This gives the equation Ai h = 0, where Ai is now a 2x9 matrix. This equation holds true for any homogeneous coordinate representation of the coordinates involved.

Each point correspondence gives rise to two independent equations in the entries for H. Given four correspondences a set of equations Ah = 0 is obtained where A is formed from the equation coefficients built from the matrix rows Ai . Next, in order to solve for A, the singular value decomposition (SVD) of A is calculated and the smallest singular value is selected as the solution and thus the linear transform between xi and x'I is obtained.

If more than four corresponding points are known and the measurements contain noise (as is usual in computer vision processing) then an over-determined solution must be found for the equation Ah = 0. This is achieved simply by stacking the n 2x9 matrices Ai into a single 2nx9 matrix and using SVD to solve for A. The process described above therefore fully defines the classic DLT solution for estimating A.

In order to apply the basic 2D ÅÆ 2D DLT algorithm to the 2D ÅÆ 3D case the dimensionality of the problem is modified as described below, In the 3D case for each correspondence Xi ÅÆ xi the following equation is derived:

$$\begin{bmatrix} 0^T & -w_i X_i^T & y_i X_i^T \\ w_i X_i^T & 0^T & -x_i X_i^T \\ -y_i X_i^T & x_i X_i^T & 0^T \end{bmatrix} \begin{pmatrix} P^1 \\ P^2 \\ P^3 \end{pmatrix} = 0$$

3.6

As in the 2D case the third equation is dependant on the first two and as such can be discounted. This leaves the following

$$\begin{bmatrix} 0^T & -w_i X_i^T & y_i X_i^T \\ w_i X_i^T & 0^T & -x_i X_i^T \end{bmatrix} \begin{pmatrix} p^1 \\ p^2 \\ p^3 \end{pmatrix}$$

3.7

A set of n point correspondences now results in the 2nx12 matrix A formed by stacking each of the equations from their respective point correspondences. The projection matrix for a given camera can then be computed by solving the set of equations Ap = 0, where p is a 3x4 projection matrix.

The algorithm outlined in this section presents a basic approach to computing each of the camera calibration matrices. The next section defines a more complete and robust solution for solving P which is based on the work outlined in the section but through the use of normalisation and the minimisation of geometric error in order to produce a more accurate solution.

3.2.2 THE GOLD STANDARD FOR ESTIMATING P

As shown in the previous section the projection matrix is calculated by solving the set of equations Ap = 0. This solution can be further refined by assuming that the world points defined during calibration are accurately known and minimising the geometric error present within the initial estimate of P. The geometric error of a given calibration can be defined as in equation 3.8.

$$\sum d(x_i, \hat{x}_i)^2$$

3.8

Where xi is the re-projected point and is the exact projection of the world point. Thus the

solution to the following minimisation is the maximum likelihood estimate of P.

$$\min_P \sum_i d(x_i, PX_i)^2 \qquad 3.9$$

Minimising the geometric error requires the use of iterative techniques. This increases the computation time but as the calculation only occurs during calibration it is an acceptable loss in performance. The Levenberg-Marquardt minimisation technique is suitable for calculating the initial DLT estimate for P which can then be used as an initial parameterisation for calculating the maximum likelihood of the projection matrix. When used in conjunction with data normalisation and the DLT this calibration method is known as the Gold Standard algorithm for estimating P. The full details of this method are detailed by Hartley and Zisserman in Multiple View Geometry for Computer Vision with the complete algorithm reproduced on the following page.

It is important to apply normalisation to the data prior to the homography estimation. Before application of the DLT the corresponding point coordinates should be translated such that the 2D coordinates centroid is at the origin and scaled such that the root mean square (RMS) distance from the origin is $\sqrt{2}$. In a similar fashion the 3D coordinates should be centred about the world coordinate system origin however in this case the RMS distance to origin should be $\sqrt{3}$. This ensures that the average point has coordinates of magnitude $(1,1,1,1)^T$.

Normalisation is necessary since the result of the DLT algorithm is dependant on the coordinate frame in which the points are expressed thus affecting the accuracy of results. Secondly, data normalisation provides invariance to the effects of coordinate changes and scale selection. By using a canonical coordinate frame for the measurement data the DLT algorithm is in practice invariant to similarity transforms. As will be demonstrated later the normalisation stage proves to be significantly more important when handling less well conditioned problems such as computation of the fundamental matrix.

Using the normalised DLT algorithm a camera matrix P is obtained for each of the cameras in the stereo rig. Given that all the cameras are calibrated simultaneously each camera matrix will project stereo matches into the same world coordinate system. Assuming sufficient accuracy in the location of calibration points, there is no requirement to align range data taken from different stereo pairs during reconstruction.

3.3 MULTI-VIEW GEOMETRY AND THE FUNDAMENTAL MATRIX

Epipolar geometry is the intrinsic projective geometry between two views. It is dependant only on the cameras internal parameters and relative pose. The fundamental matrix encapsulates this geometric relationship. It is a 3X3 matrix which satisfies the relation x'TFx=0 where a 3D point is imaged as x in the first view and x' in the second. Multiple views may be acquired by multiple cameras simultaneously or by the motion of a single camera. These two situations are geometrically identical and are treated as such throughout this section.

Figure 3.1 shows the basic components of epipolar geometry and their relation. The diagram demonstrates the relationship between a 3D world point X and its projection x and x', on two differing image planes labelled I and I' respectively. The grey triangle shows the epipolar plane for the given world point and imaging planes. e and e' represent the epipoles which intersect each of the image planes with the baseline intersecting e and e'

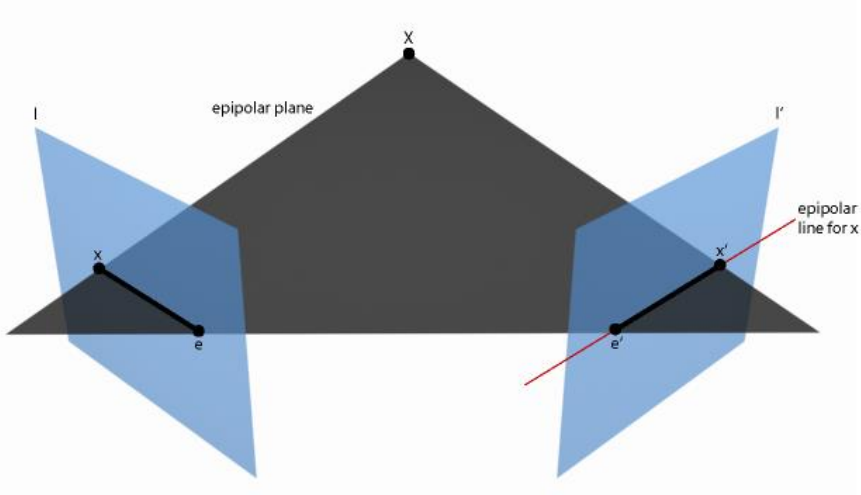


FIGURE 6: EPIPOLAR GEOMETRY IN A MULTI-VIEW SYSTEM SHOWING THE EPIPOLAR PLANE OF A SINGLE IMAGED POINT

The two image frames (I and I') are directly related via a translation vector T and a rotation matrix R. The algebraic relationship between the projection of the world point X in each of the image frames (x and x') is defined by the fundamental matrix which must satisfy equation 3.10. The solution to F should be a 3X3 rank 2 matrix, however, some of the estimation methods presented to do not conform to this rank 2 matrix constraint.

$$x'^{T} Fx = 0 \qquad 3.10$$

From this equation it follows that for any point x in the first image the corresponding epipolar line in the second image is defined by the equation l'=Fx. Obviously the converse is true for points in the second image such that l=FTx' is also true. The two equations represent a essential component of epipolar geometry since they imply the epipolar constraint which states that a point imaged in one view plane will lie somewhere on the epipolar line of the corresponding view plane. This allows the search space for corresponding image points in image pairs with known epipolar geometry to be reduced to a single dimension, thus increasing match accuracy and search speed. Furthermore conjugate points along corresponding epipolar lines have the same order in each image with the exception of corresponding points that lie on the same epipolar plane imaged from different sides. Figure 3.2 demonstrates this exceptional case. The ordering constraint can however be utilised to further reduce correlation search space and to aid propagation based correlation search strategies.

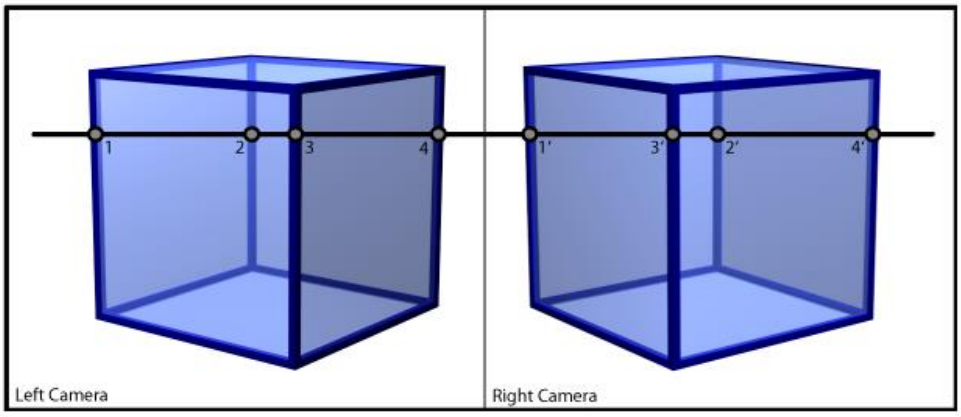


FIGURE 7: EXCEPTION TO THE EPIPOLAR ORDERING CONSTRAINT

The whole reconstruction process is heavily dependant on the ability to robustly estimate the fundamental matrix, primarily due to its ability to constrain the matching process. This section briefly describes both linear and non-linear estimation methods. These techniques are largely based on research first proposed by Zhang [100, 101] and Luong and Faugeras [102]. The first and simplest solution is a linear method requiring a minimum of 8 corresponding points between view planes to compute the fundamental matrix and is unimaginatively named the 8-point algorithm. This method finds F via a linear minimisation of a mapping between corresponding points. The second method outlined below operates by attempting to minimise the distance of a matched point between itself and the epipolar line on which it should lie. Unfortunately neither of these methods have much practical application in real world systems since neither is capable of handling outliers and erroneous matches in the correlated point set.

This means that incorrect correlations must be pruned prior to attempting fundamental matrix estimation. Therefore following the presentation of these less robust methods a number of solutions capable of dealing with errors in correlation and outliers are discussed.

The eight point algorithm is the simplest method for estimating the fundamental matrix. Unsurprisingly the algorithm requires a minimum of 8 corresponding points between view planes. Generally a calibration object is used to obtain accurate point correspondences between view planes by ensuring highly salient feature points are visible, although some implementations may use other known scene feature points. Equation 3.10 shows the required minimisation for the 8-point algorithm and is linear and homogeneous in the 9 unknown elements of F. Thus, given 8 matches it is possible to determine F up to a defined scale factor using linear methods. One method for solving equation 3.11 is to apply linear least squares in order to produce an estimate for F however any of the widely available linear minimisation methods may be utilised.

$$\min_F \sum_i (x_i'^T F x_i)^2 \qquad 3.11$$

The variables xi and xi' are vectors represent the corresponding image points where as F represents the 3X3 fundamental matrix. Using the methodology it is trivial to compute the estimate for F using non iterative techniques however this estimation is quite sensitive to noise even when using a large number of corresponding points [102]. This is, at least in part, due to the rank 2 constraint on F not being satisfied by linear approximation methods. A second approach to solving equation 3.11 can be found using Eigen analysis and singular value decomposition, however this is susceptible to the same instability issues as the linear least squares solution. Hartley [103] proposes improving the stability of 8 point algorithms by using normalised coordinates for the matched points and is a widely accepted approach, however depending on the application other, more robust, algorithms may be more suitable.

A non-linear method for estimation of F which properly satisfies the rank 2 constraint involves the minimisation of distances to epipolar lines. If we define l to represent the epipolar line of x then it should be obvious that Fl is the image of the epipolar line in the second image. If x' corresponds exactly to x then the distance between x' and l should be precisely zero, that is each matched point should fall exactly on the projection of the corresponding points epipolar line. Thus it is logical to try and use this property in order to

determine a solution to the fundamental matrix. Minimising equation 3.12 gives an estimate for F:

$$\sum d^2(x'_i, Fx_i) \tag{3.12}$$

Where d(x',Fx) is the Euclidean distance of the point x' to its corresponding epipolar line or more precisely the distance between x' and Fl. In order to produce epipolar geometry consistent across both images the distances from points to their corresponding epipolar line must take into account measurements from both images. This yields the following equation which seeks to minimise such distances in both images of a stereo pair:

$$\sum_i (d^2(x'_i, Fx_i) + d^2(x_i, F^T x'_i)) \tag{3.13}$$

Usually this method is supplemented with the use of a linear algorithm to obtain an initial estimate for F, which is then refined by minimising the distance to the relevant epipolar lines.

Unlike the methods discussed so far several methods implement fundamental matrix estimation in such a way as to detect and ignore the presence of outliers in the correlation data. Given that in most situations for estimating F it is desirable to automatically obtain correspondences from a calibration scene there is always the potential to produce erroneous matches. As such estimation techniques that are robust against such outliers are desirable. Both the Least Median of Squares (LMS) method and the RANSAC algorithm meet this criterion.

The LMS method was adapted from earlier work by Zhang [101] in order to approach the fundamental matrix estimation problem. Following corner point detection and correlation the algorithm progresses as follows. For n point correspondences (xi, x'i) a Monte Carlo technique is used to obtain m samples of 8 corner matches. For each sub-sample j an appropriate estimate for the fundamental matrix Fj is computed. The median of squared residuals (Mj) is determined for each of the sub-samples with respect to the superset containing all obtained corresponding corner points as demonstrated by equation 3.14.

$$M_j = med_{i=1,\ldots,n}[d^2(x'_i, F_j x_i) + d^2(x_i, F_j^T x')] \tag{3.14}$$

The number of sub-samples m is determined by equation 3.15 which calculates the probability that at least one of the sub-samples is good, assuming that the superset of correspondences contains no more than ε correspondences which are outliers.

$$P = 1 - [1 - (1 - \varepsilon)^{p}]^{m} \tag{3.15}$$

Rousseeuw and Leroy calculate a robust standard deviation estimate to compensate for Gaussian noise in the input correlations using the following equation:

$$\tilde{\sigma} = 1.4826[1 + 5/(n - p)]\sqrt{M_j}$$

Mj is the previously calculated minimal medial. Following calculation of the robust standard deviation weight is assigned for each of the matched corner correspondences as shown below:

$$w_i = \begin{cases} if\,(r_i^2 \le (2.5\tilde{\sigma})^2)\,then(1) \\ else(0) \end{cases}$$

Where:

$$r_i^2 = d^2(x'_i, F\,x_i) + d^2(x_i, F^T x')$$

After iterating the above weighting algorithm to each of the input point correspondences the result is a sub-set of points marked as outliers with wi=0. These outliers are eliminated from the correspondence set and not used further in the calculation of F. Given that outliers are removed from the set of correspondences the fundamental matrix may now be computed by solving the weighted least-squares problem as shown in equation 3.16:

$$\min \sum w_i r_i^2 \tag{3.16}$$

Any suitable minimisation technique can be used to solve equation 3.16 however the popular Levenberg-Marquardt algorithm is commonly used to arrive at a solution.

Another factor that must be considered is the method by which correspondences are initially divided into sets of 8 matches. In order to obtain accurate F estimation using any of the described methods, input point correspondences should be selected from an area covering a large amount of the image. If all matches are located in a small image region epipolar geometry estimates will be poor. Thus in order to ensure a good distribution of points within each subset of correlations Zhang et. al. [105] implemented a regularly random selection method based on bucketing techniques. Each subset of 8 matches used in the LMS estimation method are selected by first dividing the image into regions of a predefined size, 8 regions are randomly selected and one match from each region is added to the current subset. This process is repeated until all matches have been divided into subsets and F estimation via the minimisation method outlined above can progress.

The final algorithm for consideration in relation to estimating the epipolar geometry between two views is the Random Sample Consensus (RANSAC) method. First proposed by Torr [106] the method is similar to LMS, differing mainly in the method by which outliers are determined. Henrichsen [6] states that if the fundamental matrix needs to be specified for many images then the LMS method should be used on one image pair to determine an outlier threshold for use with RANSAC on the remainder of the image pairs. Additional details concerning the RANSAC estimation method and its variants are widely available.

3.4 3D PROJECTION USING LINEAR TRIANGULATION

3D projection involves computing the world space coordinate of a point imaged in one or more cameras. In order to perform this calculation the calibration matrices must be known for each of the image planes in which the point is visible. The 2D coordinates of the imaged point must also be known for both view planes. The 3D projection problem can thus be defined as follows:

Given two corresponding points m and m', compute the 3D coordinates of M in accordance with some global coordinate system.

Obviously, in order to produce a 3D model the correspondence problem must be solved prior to projecting the matches into 3 dimensions. Assuming that a set of reliable matches across multiple view planes have been computed and each planes associated projection matrix is known the following techniques allow the triangulation of the world coordinates from correlated image points. The most trivial solution to the triangulation problem involves

simply back-projecting rays from the measured image points to their intersection. However, outside of a purely mathematical application, errors in the estimation of P or in the correlated image points causes the back projected rays not to intersect, thus in general it is necessary to estimate the optimum point coordinates in world space.

The aim of 3D projection is to estimate a 3D point X which exactly satisfies the supplied camera geometry such that it projects as:

$$x = PX$$
$$x' = P'X \qquad 3.17$$

Assuming, however, that there are errors both in the set of correlated image points and the camera calibration the back projected rays will be skewed. As a consequences of this skew there will not be a point X which satisfies x=PX, x'=P'X nor will the epipolar constraint be fully satisfied such that X'TFx=0 is not true. The two previous statements are equivalent since matching pairs of points will only intersect if and only if the pair of points already satisfy the epipolar constraint. Many methods for estimating the intersection of the rays and the resultant 3D coordinate have been developed, the most popular of which are explained in the remainder of this section.

The maximum likelihood estimate, under Gaussian noise, is given by the point X which minimises the re-projection error of the measure image points. Re-projection error is the summed squared distances between the projections of X into the image planes and the position of the initial measurements which were used for projection. A number of the available approaches to obtaining a good estimate for X are now considered. The scope of this analysis is limited to linear triangulation methods, despite the existence of more robust approaches such as Samson approximation and the optimal polynomial approach described by Hartley [5].

Linear triangulation methods are the simplest approach to computing 3D structure from a set of corresponding matches. The estimated point does not exactly satisfy the geometric relations of the camera system, however, utilising robust estimation techniques a reasonable coordinate can be established. Linear triangulation is a direct analogue of the DLT method used for camera calibration earlier in this chapter and therefore many of the same concepts apply. Firstly the two definitions from equation 3.17 are combined to form a single equation

that is linear in X: AX=0. The homogenous scale factor is eliminated by calculating a cross product to yield a set of three equations for each correlated image point, two of which are linearly independent. Thus the equation resulting from the first image is written as xX(PX)=0 which expanded gives the following:

$$x(p^{3T}x) - (p^{1T}x) = 0$$
$$y(p^{3T}x) - (p^{2T}x) = 0$$
$$x(p^{2T}x) - y(p^{1T}x) = 0$$

3.18

PiT are the rows of the camera projection matrix calculated during the calibration phase of the reconstruction. These equations are linear in X and can now be used to produce an equation in the form AX=0 as shown in equation 3.19.

$$A = \begin{bmatrix} xp^{3T} - p^{1T} \\ yp^{3T} - p^{2T} \\ x' p'^{3T} - p'^{1T} \\ y' p'^{3T} - p'^{2T} \end{bmatrix}$$

3.19

In the above result two equations from each image have been included, since two of the equations in 3.18 are linearly independent the third can be excluded from the calculations. Obviously (x,y) is the correlated coordinate from the view plane with projection matrix p and (x',y') represents the image coordinates of the correlated point in the view plane described by p'. Solving 3.19 for X in this manner allows the computation of a linear estimate for the 3D point being reconstructed.

CHAPTER 4 EXPERIMENTATION AND ANALYSIS

4.1SET UP FOR THE EXPERIMENTS

Two mobile phones were kept on the table such that both the cameras were **coplanar**. The distance between the cameras is known as **baseline distance**.

FIGURE 8: STEREO CAMERA SETUP

4.2STEREO CAMERA CALIBRATION

This is done by finding checkerboard pattern (chess board) in the pair of stereo images. Block Size of the checkerboard is known (here we used 49mm). Checkerboard pattern must be asymmetric(normally chess board is 8x8, here we used 5x4). Set of images were captured by placing checkerboard at different positions. For better results, we need at least 10 test patterns.

So to find pattern in checkerboard board, we used the **Stereo Camera Calibrator App** in **Matlab**.

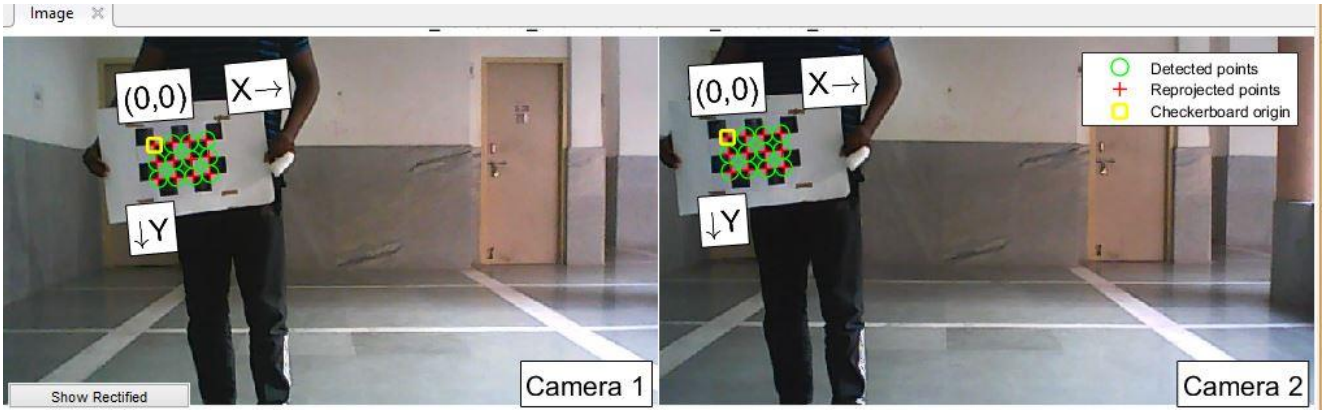
Workflow for calibration of stereo camera using the app:

1. Preparation of images, camera, and calibration pattern.
2. Loading image pairs.
3. Calibration of the stereo camera.

4. Evaluation calibration accuracy.
5. Adjustment of parameters to improve accuracy (if necessary).
6. Exporting the parameters object.

After capturing set of images with different locations of checkerboard pattern, we need to add them to the app for finding pattern in the images and calibrate it.

4.3 CALIBRATION RESULTS



Left Image Right Image

FIGURE 9: CHECKERBOARD PATTERN DETECTION

Examining Reprojection Errors

The reprojection errors are the distances in pixels between the detected and the reprojected points. The Stereo Camera Calibrator app calculates reprojection errors by projecting the checkerboard points from world coordinates, defined by the checkerboard, into image coordinates. The app then compares the reprojected points to the corresponding detected points. As a general rule, reprojection errors of less than one pixel are acceptable.

- **Reprojection Errors Bar Graph**

  The bar graph displays the mean reprojection error per image, along with the overall mean error. The bar labels correspond to the image pair IDs. The highlighted pair of bars corresponds to the selected image pair.

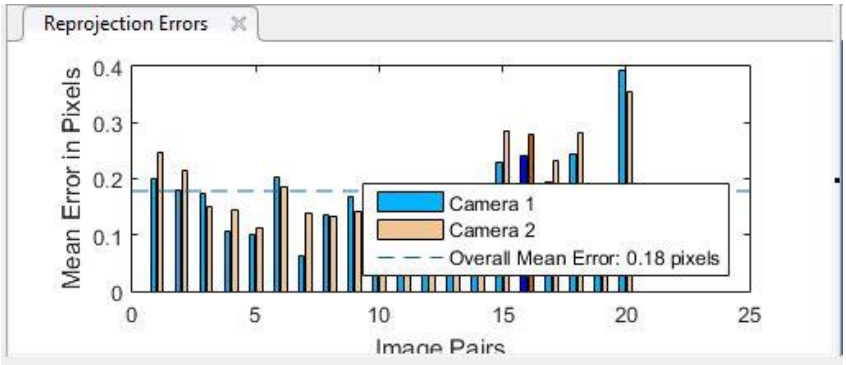


FIGURE 10: REPROJECTION ERROR BAR GRAPH

Here, we can **improve calibration** (accuracy) by removing the images having large reprojection error (here, 20th image) and recalibrate.

4.4 EXTRINSIC PARAMETER VISUALIZATION

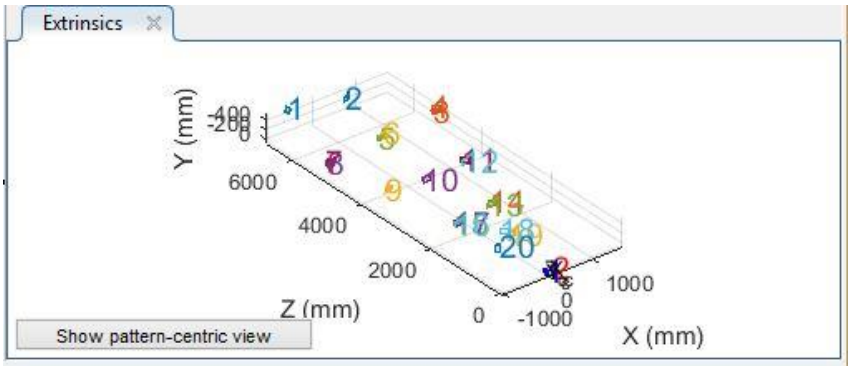


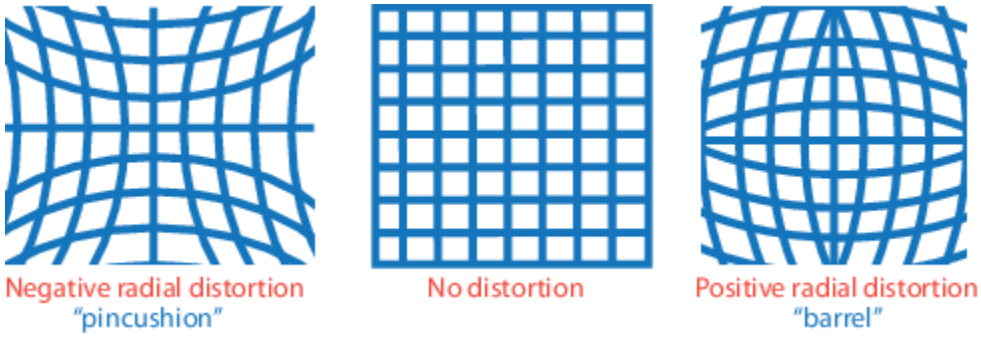
FIGURE 11: EXTRINSIC PARAMETER VISUALIZATION

The 3-D extrinsic parameters plot provides a camera-centric view of the patterns and a pattern-centric view of the camera. The camera-centric view is helpful if the camera was stationary when the images were captured. The pattern-centric view is helpful if the pattern was stationary. Click the button on the display to toggle between the two visuals. Click and drag a graph to rotate it. Click a checkerboard or a camera to select it. The highlighted data in the visualizations correspond to the selected image in the list. Examine the relative positions of the pattern and the camera to see if they match what you expect. For example, a pattern that appears behind the camera indicates a calibration error.

4.5 RADIAL, SKEW AND TANGENTIAL DISTORTION

Changing the Number of Radial Distortion Coefficients

You can specify 2 or 3 radial distortion coefficients by selecting the corresponding radio button from the Options section. **Radial distortion** occurs when light rays bend more near the edges of a lens than they do at its optical center. The smaller the lens, the greater the distortion.



The radial distortion coefficients model this type of distortion. The distorted points are denoted as ($x_{distorted}$, $y_{distorted}$):

$$x_{distorted} = x(1 + k_1*r^2 + k_2*r^4 + k_3*r^6)$$

$$y_{distorted} = y(1 + k_1*r^2 + k_2*r^4 + k_3*r^6)$$

- x, y — Undistorted pixel locations. x and y are in normalized image coordinates. Normalized image coordinates are calculated from pixel coordinates by translating to the optical center and dividing by the focal length in pixels. Thus, x and y are dimensionless.
- k_1, k_2, and k_3 — Radial distortion coefficients of the lens.
- r^2: $x^2 + y^2$

Typically, two coefficients are sufficient for calibration. For severe distortion, such as in wide-angle lenses, you can select 3 coefficients to include k_3.

Computing Skew

When you select the **Compute Skew** check box, the calibrator estimates the image axes skew. Some camera sensors contain imperfections that cause the *x*- and *y*-axis of the image to not be perpendicular. You can model this defect using a skew parameter. If you do not select the check box, the image axes are assumed to be perpendicular, which is the case for most modern cameras.

Computing Tangential Distortion

Tangential distortion occurs when the lens and the image plane are not parallel. The tangential distortion coefficients model this type of distortion.

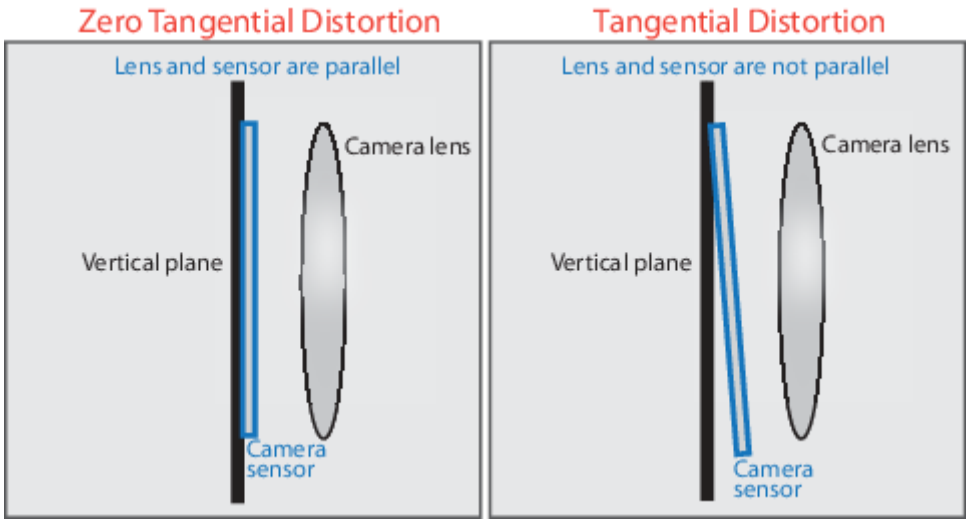


FIGURE 12: TANGENTIAL DISTORTION

The distorted points are denoted as ($x_{distorted}$, $y_{distorted}$):

$$x_{distorted} = x + [2 * p_1 * x * y + p_2 * (r^2 + 2 * x^2)]$$

$$y_{distorted} = y + [p_1 * (r^2 + 2 * y^2) + 2 * p_2 * x * y]$$

- x, y — Undistorted pixel locations. x and y are in normalized image coordinates. Normalized image coordinates are calculated from pixel coordinates by translating to the optical center and dividing by the focal length in pixels. Thus, x and y are dimensionless.
- p_1 and p_2 — Tangential distortion coefficients of the lens.

- $r^2 = x^2 + y^2$

When you select the **Compute Tangential Distortion** check box, the calibrator estimates the tangential distortion coefficients. Otherwise, the calibrator sets the tangential distortion coefficients to zero.

Exporting Camera Parameters

When you are satisfied with calibration accuracy, click **Export Camera Parameters**. You can save and export the camera parameters to an object or generate the camera parameters as a MATLAB script.

CHAPTER 5 RESULTS AND DISCUSSION

5.1 EXPERIMENT -SETUP 1

A pair of Stereo Images captured using two **webcams** with baseline distance of **14cm** are shown below. In this arrangement, four objects are kept at different distances from the camera plane.

a. Left Image b. Right Image

FIGURE 13: STEREO PAIR OF IMAGES TAKEN FROM TWO WEBCAMS

Before and after rectification images have been shown below.Any minute differences in the setup, in terms on alignment are adjusted.

a. Before Rectification b. After Rectification

FIGURE 14: RECTIFICATION

The disparity map is shown below. Objects are placed according to their depth. Nearby objects are in the red region. Far away objects are in the blue region.

FIGURE 15: DISPARITY MAP

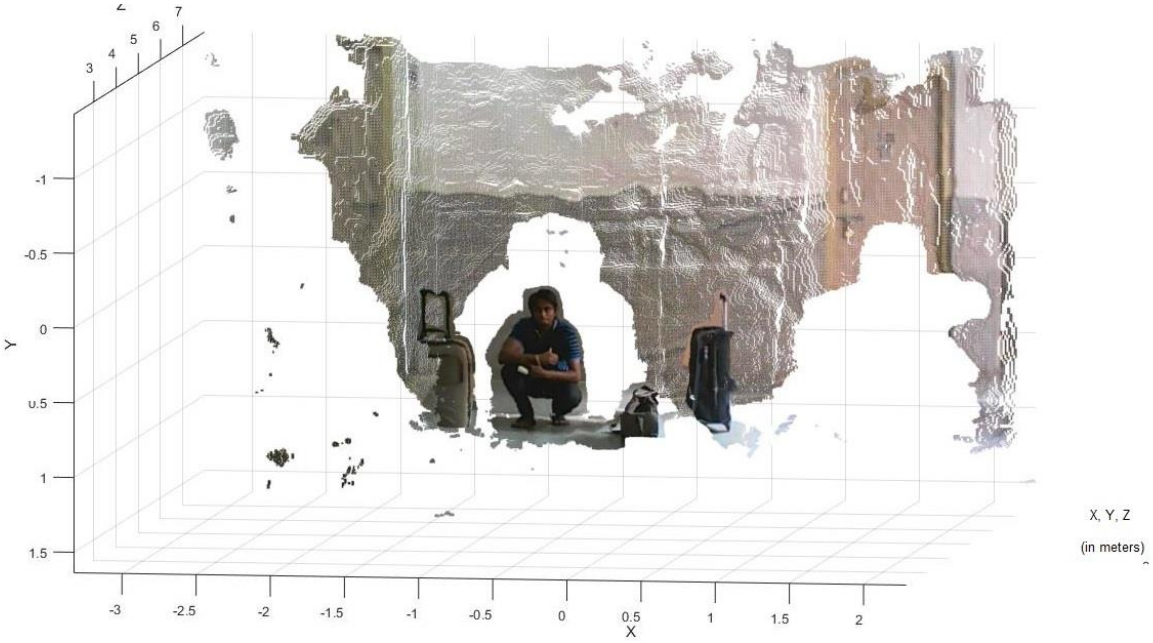


FIGURE 16: 3D RECONSTRUCTED SCENE (FRONT VIEW)

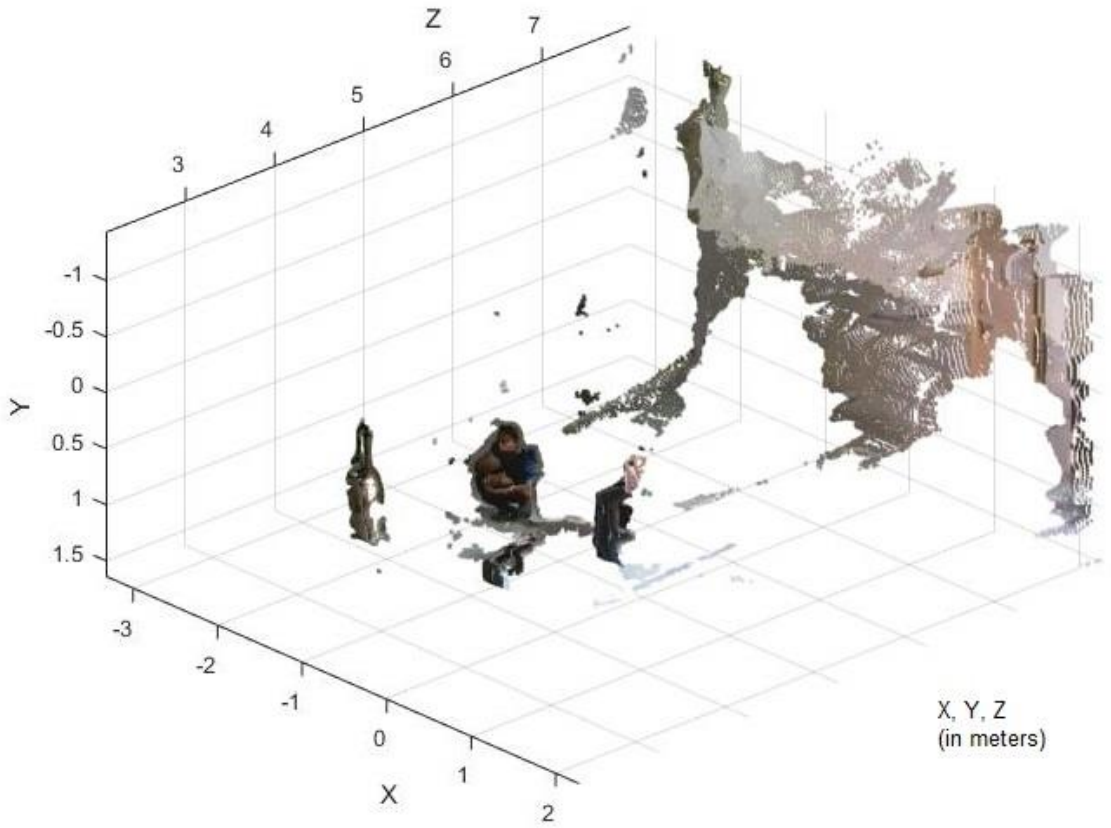


FIGURE 17: RECONSTRUCTED SCENE (ISOMETRIC VIEW)

The objects are reconstructed at actual depths from the camera and in actual size.

5.2 EXPERIMENT-SETUP 2

MOBILE PHONE-BASELINE 14CM

Stereo images captured from mobile phone with baseline distance 14cm. Here , the objects are kept nearby in the range of 3 to 4 m perpendicular distance from the camera plane.

LEFT IMAGE RIGHT IMAGE

FIGURE 18: STEREO IMAGES FROM MOBILE PHONE WITH BASELINE DISTANCE 14CM

After calibration and rectification , disparity map is calculated and 3D scene is shown below

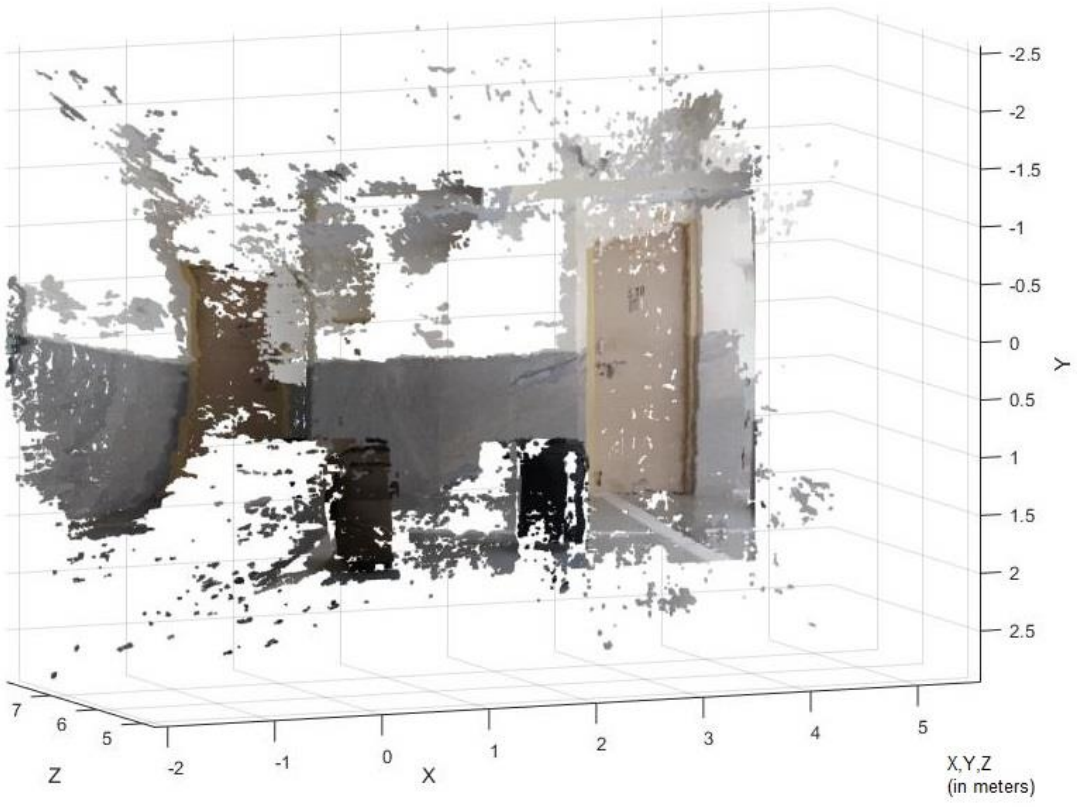


FIGURE 19: 3D SCENE RECONSTRUCTION (FRONT)

Another view(isometric) of 3D reconstructed scene is shown below.

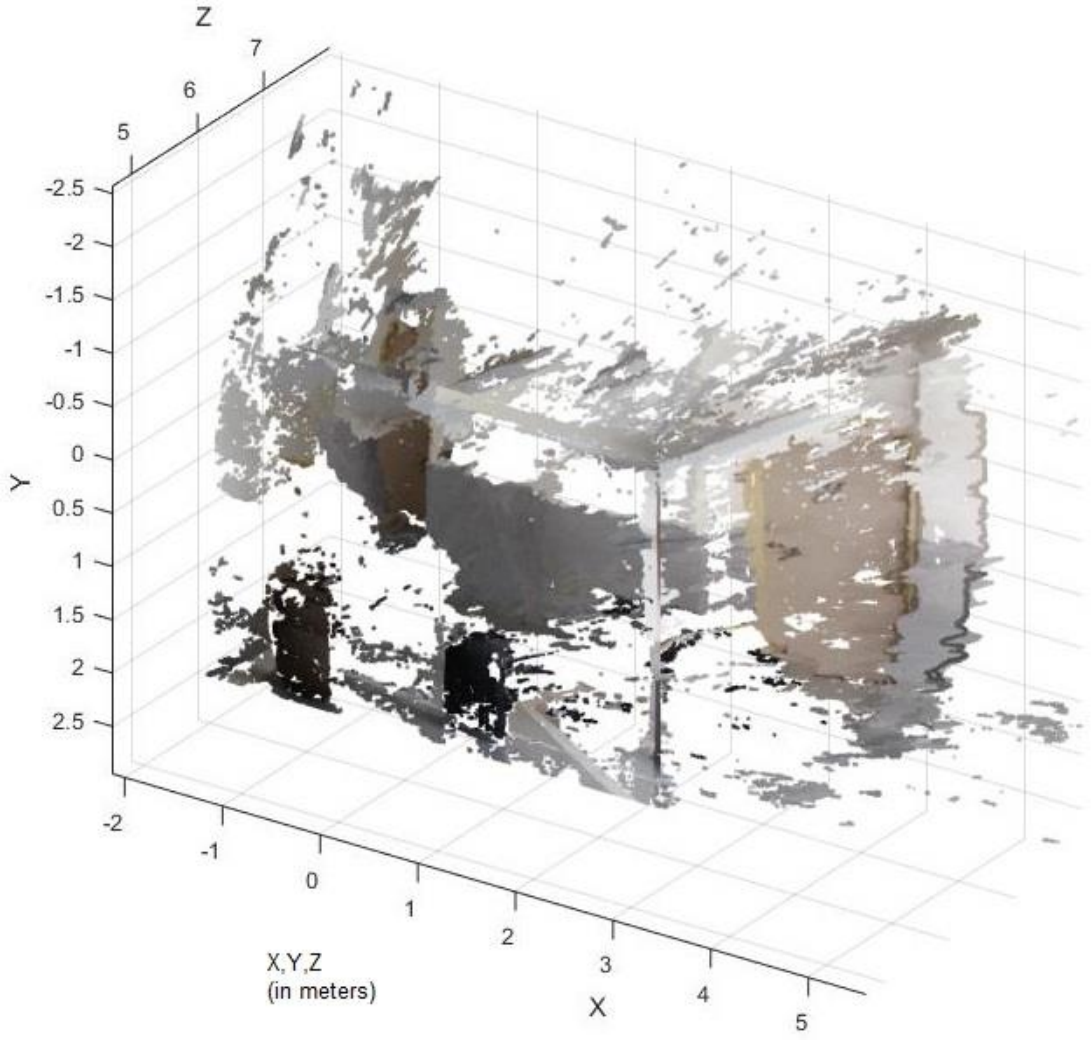


Figure 20: 3D scene reconstruction (isometric view).

Here, the depth of both the objects appears to be same. It is not able to reconstruct the depth of the objects kept in the range of 3-4 m. Depth of both the object is 4.719m (reconstructed) whereas the actual depths of left object and right object are 3.1m and 3.70m respectively.

Objects are distorted in size as well as in depth.

Objects	Actual distance (m)	Reconstructed distance (m)
Left Bag	3.1	4.179
Right Bag	3.7	4.179
Wall	7.2	6.32

TABLE 1: COMPARISION OF ACTUAL AND RECONSTRUCTED DISTANCES FOR SETUP 2A

Stereo images captured from mobile phone with baseline distance 14cm. Here , the objects are kept far in the range of 4.5 to 6 m perpendicular distance from the camera plane.

Left Image Right Image

FIGURE 21: STEREO IMAGES FROM MOBILE PHONE WITH BASELINE DISTANCE 14CM

After calibration and rectification , disparity map is calculated and 3D scene is shown below

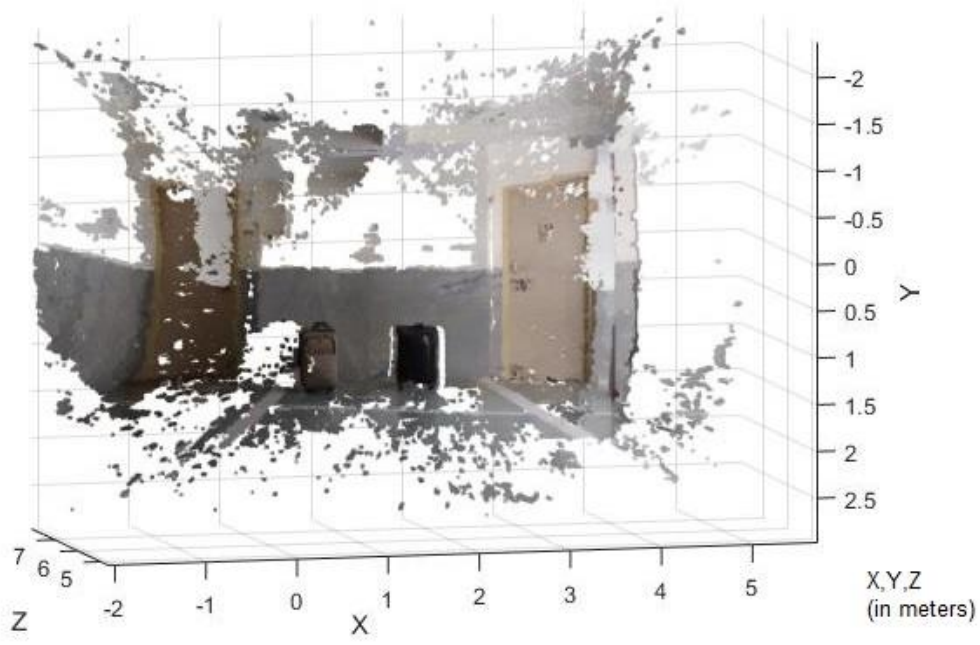


FIGURE 22: 3D SCENE RECONSTRUCTION (FRONT)

Another view(isometric) of 3D reconstructed scene is shown below.

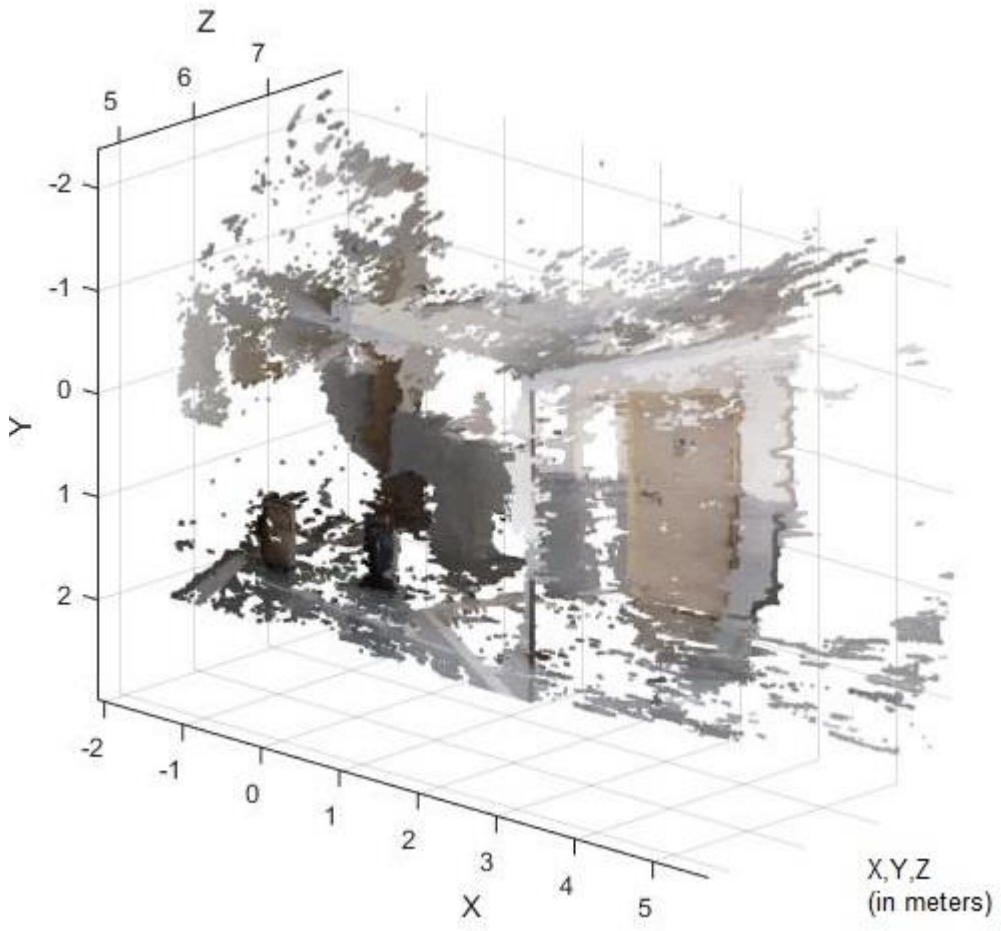


FIGURE 23: 3D SCENE RECONSTRUCTION (ISOMETRIC VIEW).

Here, actual depth of the objects has been recovered. It is able to reconstruct the depth of the objects kept in the range of 4.5-6 m. The actual size of the object with respect to the surrounding is recovered.

Objects	Actual distance (m)	Reconstructed distance (m)
Left Bag	4.94	4.812
Right Bag	5.52	5.42
Wall	7.2	6.8

TABLE 2: COMPARISION OF ACTUAL AND RECONSTRUCTED DISTANCES FOR SETUP 2B

5.3 EXPERIMENT-SETUP 3

MOBILE PHONE-BASELINE 7CM

Stereo images captured from mobile phone with baseline distance 14cm. Here , the objects are kept far in the range of 4.5 to 6 m perpendicular distance from the camera plane.

a. Left Image b. Right Image

FIGURE 24: STEREO IMAGES FROM MOBILE PHONE WITH BASELINE DISTANCE 7CM

After calibration and rectification , disparity map is calculated and 3D scene is shown below

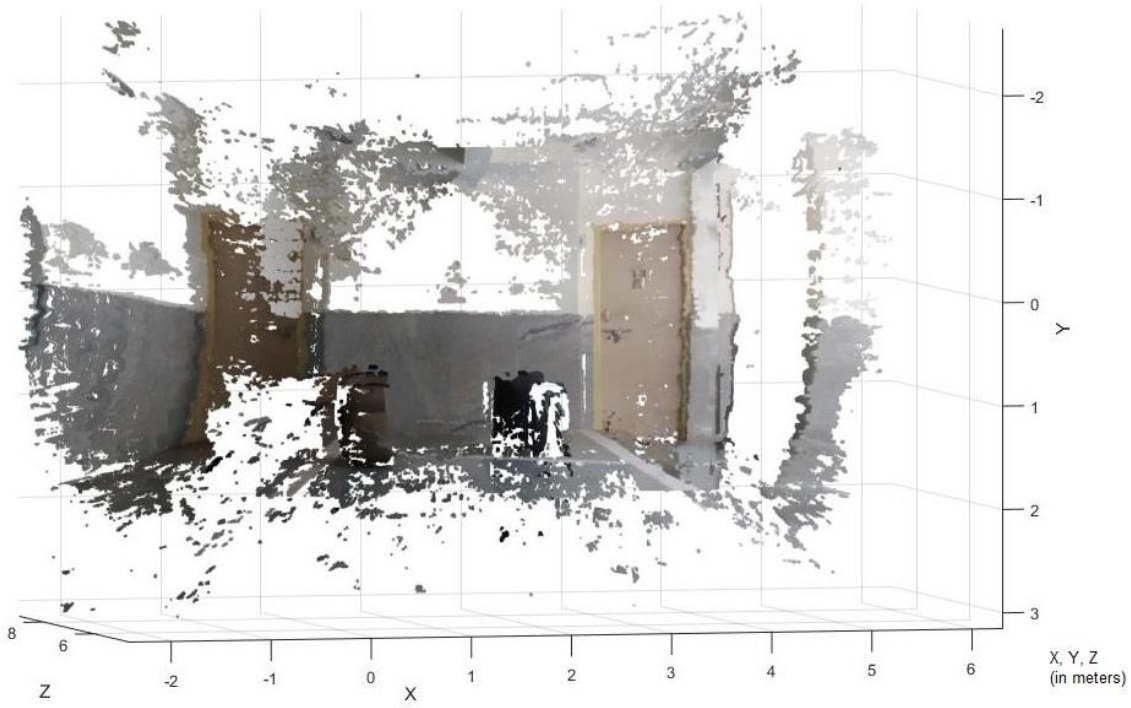


FIGURE 25: 3D SCENE RECONSTRUCTION (FRONT)

Another view(isometric) of 3D reconstructed scene is shown below.

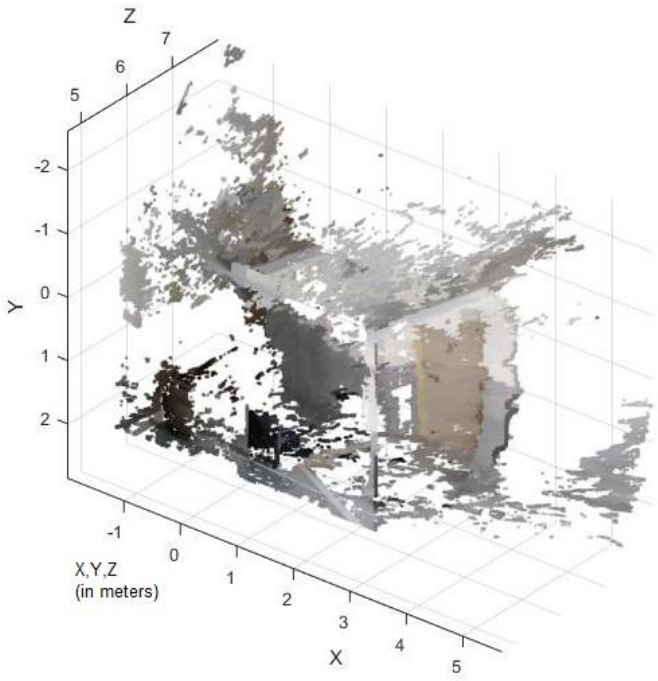


FIGURE 26: 3D SCENE RECONSTRUCTION (ISOMETRIC VIEW).

Here, actual depth of the objects has been recovered. It is able to reconstruct the depth of the objects kept in the range of 4.5-6 m. The actual size of the object with respect to the surrounding is recovered.

Objects	Actual distance (m)	Reconstructed distance (m)
Left Bag	4.94	4.719
Right Bag	5.52	4.719
Wall	7.2	6.35

TABLE 3: COMPARISION OF ACTUAL AND RECONSTRUCTED DISTANCES FOR SETUP 3A

Stereo images captured from mobile phone with baseline distance **7cm.** Here , the objects are kept nearby in the range of **3 to 4 m** perpendicular distance from the camera plane.

a. Left Image b. Right Image

FIGURE 27:STEREO IMAGES FROM MOBILE PHONE WITH BASELINE DISTANCE 7CM

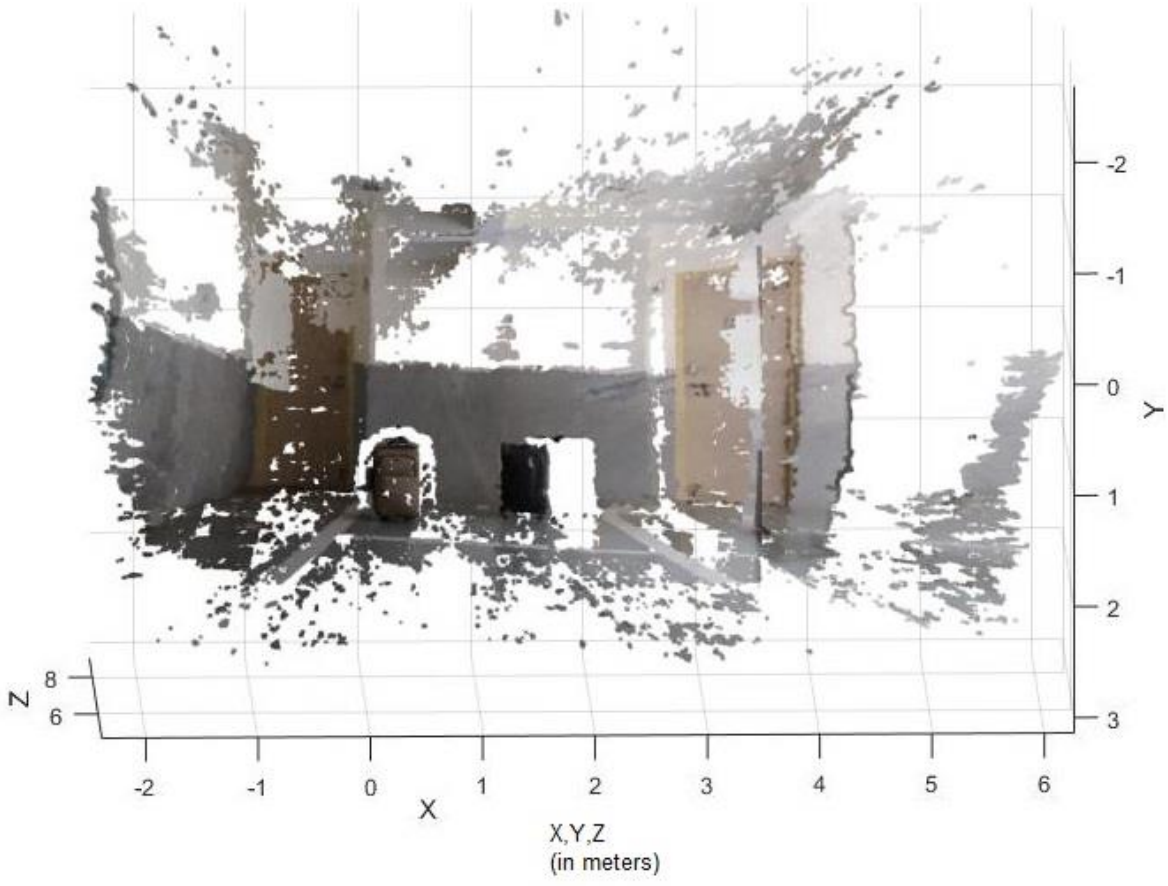


FIGURE 28: 3D SCENE RECONSTRUCTION (FRONT)

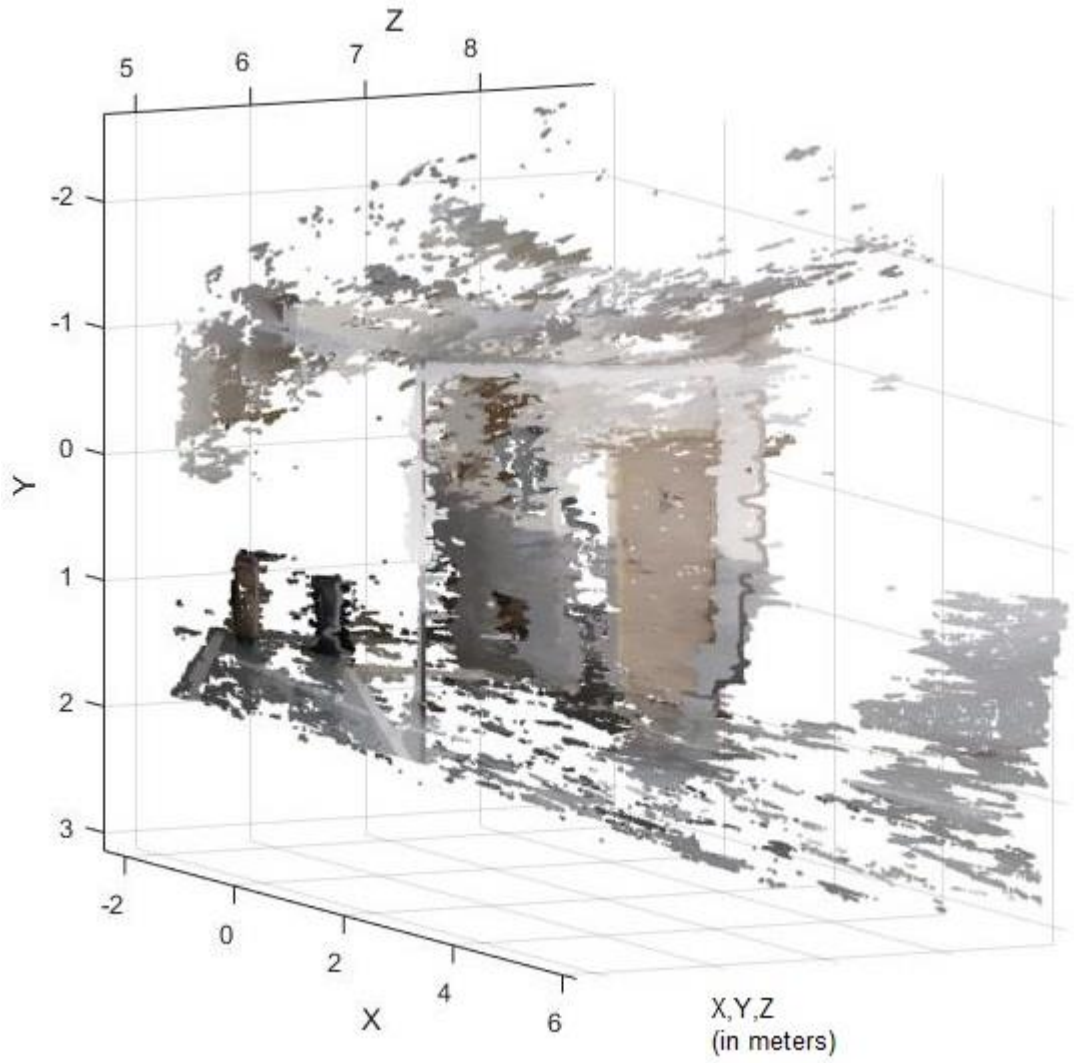


FIGURE 29: 3D SCENE RECONSTRUCTION (ISOMETRIC VIEW).

Here, actual depth of the objects has been recovered. It is able to reconstruct the depth of the objects kept in the range of **3 to 4 m**. The actual size of the object with respect to the surrounding is recovered.

Objects	Actual distance (m)	Reconstructed distance (m)
Left Bag	3.1	3.2
Right Bag	3.7	3.8
Wall	7.2	6.43

TABLE 4: COMPARISION OF ACTUAL AND RECONSTRUCTED DISTANCES FOR SETUP 3B

CHAPTER 6 INFERENCE

6.1 ACCURACY

From table it is observed that for a baseline 14cm

- Object at 3.1m mapped at 4.179m

  Accuracy = (1 - (1.079/3.1))*100 = 65.19%
- Object at 3.7m mapped at 4.179m

  Accuracy = (1 - (0.479/3.7))*100 = 87.05%
- Object at 4.94m mapped at 4.812m

  Accuracy = 97.4%
- Object at 5.52m mapped at 5.42m

  Accuracy = 98.18%

This shows that the reconstruction for a baseline 14cm started at 3.1 m and accurate reconstruction was done at 4.8m .After increasing the distance, again the reconstruction got distorted after 6.5m. Therefore, range of reconstruction for a baseline of 14cm is between 4.5 and 6.5 m.

6.2 EFFECT ON RECONSTRUCTED IMAGES BY VARYING THE BASELINE

It is observed that for 7cm baseline distance the range of accurate 3D reconstruction is 3 to 4.5m & for 14cm baseline distance, it is 4.5 to 6.5m..

This shows that as baseline is increased, objects at a greater distance get reconstructed appropriately.

But this is true only for a limited range of baseline. As baseline gets further large, stereo matching between images reduces and for smaller baselines the objects get distorted.

Experiments were carried out for a baselines between 4cm to 14cm.

6.3 WEBCAM VS MOBILE

From reconstructed scenes as shown in fig 16 and fig 22, it is observed that

- the scene reconstruction from webcam system was better than that from the mobile for the same baseline distance.

- The auto-focus of mobile camera results in different focus during the calibration process leading to slight errors in calculating the Camera Parameters.
- The focus of the webcam stays constant, one essential feature used in the calibration technique and hence better results.

6.4 DISTORTION AT SCENE BORDER

Since there is no stereo matching between the images at the borders ,it is not possible to reconstruct this part of the image.

7. FUTURE SCOPE

Calibration is done with the checker-board pattern throughout this project. This method wouldn't be possible in a dynamic environment. As mobile is a gadget which isn't used in a static environment, a different approach such as Self Calibration would make this idea practically implementable.

Another area of development would be with the auto focus option in the mobiles. Varying focus during calibration makes it difficult in estimating the Camera Parameters accurately leading to errors in the reconstruction process.