PRECISE RECONSTRUCTION OF INDOOR ENVIRONMENTS USING RGB-DEPTH SENSORS

WALID ABDALLAH ABOUMANDOUR DARWISH

PhD

The Hong Kong Polytechnic University

2018

The Hong Kong Polytechnic University

Department of Land Surveying and Geo-Informatics

Precise Reconstruction of Indoor Environments Using RGB-Depth Sensors

Walid Abdallah Aboumandour Darwish

A thesis

submitted in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

May 2018

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text

_____(Signature)

Walid Abdallah Aboumandour Darwish (Name of Student)

Dedication

Dedicated with love and gratitude to my parents and my small family Shereen,

Juwayriah, Mariah

Abstract

Commercial RGB-D cameras (e.g., Kinect) have been widely used in the gaming industry as non-touch remote controllers. RGB-D cameras are designed for maximum three-meter range applications where geometric fidelity is not of utmost importance. Recently, Structure Sensor was released in the commercial market as the first mobile RGB-D camera. As this promising camera has great potential to be used in indoor navigation and 3D modelling, precise calibration of their depth information, working range, and geometric sensor parameters should be thoroughly obtained.

In this study, we propose a novel calibration method for Structured Light (SL) RGB-D cameras. The calibration method uses a novel distortion model for the captured depth images. The depth distortion model consumes the distortion effects of both IR sensors. The method calibrates the geometric parameters of each RGB-D camera lens. Moreover, the method extends to modelling the systematic depth bias resulting from imaging conditions and IR sensors' baseline. The method can thoroughly calibrate the SL RGB-D cameras' full range independently of the IR sensors' baseline. The calibration procedure was normalized and designed to be automatic. The proposed calibration method can calibrate the full range of the sensor and achieve a relative error of 0.8%, while ordinary calibration methods can only calibrate up to 34% of the sensor's range and achieves a relative error of 4.0%.

Due to indoor scalability, many RGB-D frames were collected and registered together to form a complete colored 3D model. The Simultaneous Localization And Mapping (SLAM) technique is used to track the RGB-D camera. The scene structure, the depth range, and feature types are the dominant elements affecting registration accuracy and

thus SLAM performance. Those elements can easily force SLAM into a severe drift or terminate the tracking status (lost tracking). Current SLAM systems use visual matched point features to compute the camera pose; therefore, those systems suffer from lost tracking problems and inevitable drift.

To minimize the probability of lost tracking and drift, strong features (lines, planes) were added to the SLAM tracking core. In this context, a new procedure to detect, extract, describe, and match those 3D features was proposed. Line features were extracted using RGB and depth images while plane features were extracted using the depth image. The procedure uses a novel descriptor which adopted both visual and depth information to describe the 3D features for further matching. A new RGB-D SLAM system is proposed to utilize the valuable 3D matched features. The Fully Constrained RGB-D SLAM (FC RGB-D SLAM) system minimizes the combined geometric distance of 2D and 3D matched features to estimate the camera pose, then to enhance 3D model quality, the system applies a global refinement stage to refine the estimated camera poses based on indoor geometric constraints. Also, the system adopts the graph-based optimization technique to correct the closure error whenever a loop closure is detected. The results show that compared to visual RGB-D SLAM systems, FC RGB-D SLAM can achieve significant improvements in 3D model accuracy with and without loop closure constraints.

Publications arising from the thesis

During 3 years of work, several peer-reviewed and non-peer reviewed papers were published, the following is the publication list arising from this thesis.

1. W. Chen, **W. Darwish**, S. Tang, and W. Li, "Precise Calibration and Error Modeling For Indoor 3D Modelling Sensors," in *The 9th International Symposium on Mobile Mapping Technology MMT2015*, Sydney, Australia, 2015.
2. **W. Darwish**, W. Chen, S. Tang, and W. Li, "Full Parameter Calibration for Low Cost Depth Sensors," presented at the Melaha 2016 International Conference and Exhibition, Cairo, Egypt, 2016.
3. S. Tang, Q. Zhu, W. Chen, **W. Darwish**, B. Wu, H. Hu*, et al.*, "Enhanced RGB-D Mapping Method for Detailed 3D Indoor and Outdoor Modeling," *Sensors,* vol. 16, p. 1589, 2016.
4. S. Tang, Q. Zhu, W. Chen, **W. Darwish**, B. Wu, H. Hu*, et al.*, "ENHANCED RGB-D MAPPING METHOD FOR DETAILED 3D MODELING OF LARGE INDOOR ENVIRONMENTS," *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences,* vol. 3, 2016.
5. **W. Darwish**, W. Li, S. Tang, and W. Chen, "Coarse to Fine Global RGB-D Frames Registration For Precise Indoor 3D Model Reconstruction," in *Localization and GNSS (ICL-GNSS), 2017 International Conference on*, 2017, pp. 1-5.
6. **W. Darwish**, S. Tang, W. Li, and W. Chen, "A New Calibration Method for Commercial RGB-D Sensors," *Sensors,* vol. 17, p. 1204, 2017.

7. **W. Darwish**, W. Li, Y. Li, S. Tang, and W. Chen, "Constrained RGBD SLAM for Robust 3D Model Reconstruction of Indoor Environment," presented at the International Symposium on GNSS (IS-GNSS) 2017, Hong Kong, 2017.

8. *S. Tang, W. Chen, W. Wang, X. Li,* ***W. Darwish****, W. Li, et al., "Geometric Integration of Hybrid Correspondences for RGB-D Unidirectional Tracking," Sensors (Basel, Switzerland), vol. 18, 2018*

9. **W. Darwish**, W. Li, Y. Li, S. Tang, and W. Chen, "A Robust Calibration Method for Consumer Grade RGB-D Sensors for Precise Indoor Reconstruction" submitted to IEEE access (May 2018).

10. **W. Darwish**, W. Li, Y. Li, S. Tang, and W. Chen, "Fully Constrained RGBD SLAM for Precise 3D Model Reconstruction of Large Indoor Environment" in progress.

11. S. Tang, W. Li, **W. Darwish,** Y. Li, and W. Chen, "A Vertex-to-Edge Weighted Closed-Form Method for Achieving Drift-Free Dense RGB-D Indoor SLAM " submitted to neurocomputing.

Acknowledgments

I wish to thank all the person who supervised, guided, and help me during my PhD studies. First of all, I would like to appreciate the wise guidance and useful critiques of my chief supervisor Prof. Wu CHEN. My grateful thanks extended to my co-supervisor Dr. Bo WU for his useful help. I would like to thank Prof. Bruce King for his help in the field of Terrestrial Laser Scanner technology.

I would like to thank all my colleagues in the department of Land Surveying and Geo-Informatics (LSGI). I would like to express my special thanks to our 3D mapping group, Wenbin Li, Yaxin Li, Shengjun Tang, for their help and support.

Table of contents

List of figures

List of tables

List of Abbreviation

BIM	Building an information modeling
DLT	Direct Linear Transform
ICP	Iterative Closest Point
LSGI	Land Surveying and Geo-Informatics
PBA	Photogrammetric Bundle Adjustment
PCA	Principal Component Analysis
RANSAC	Random sample consensus
SC	Stereo Cameras
SFM	Structure from Motion
SIFT	Scale Invariant Feature Transform
SL	Structured Light
SLAM	Simultaneous Localization And Mapping
SSD	Sum of Squared Differences
SVD	Singular Value Decomposition
ToF	Time of flight
CCD	Charge Coupled Device
GPU	Graphics Processing Unit
IMU	Inertial Measurement Unit

IR	InfraRed
RGB	Red Blue Green (visible bands)
RGBD	Red Blue Green Depth
RMSE	Root Mean Square Error
SDK	Software Development Kit
AR	Augmented Reality
VR	Virtual Reality

List of Symbols

Z_i	The perpendicular distance between the IR sensor's baseline and feature point
f	The focal length of the IR camera
w	The baseline between IR camera and IR projector
Z_o	The depth of standard plane
d_i	The measured disparity
σ_z	The precision of depth
σ_d	The precision of disparity
s	The scale factor
x, y	The image point coordinates in pixels
X, Y, Z	The ground point coordinates
R	Rotation matrix
T	Translation vector
K	Camera intrinsic matrix
f_x	The camera focal lengths in x direction in pixels
f_y	The camera focal lengths in y direction in pixels
c_x, c_y	The coordinate of camera principal point in pixels
e	The skew between x and y direction
x_d, y_d	The Coordinates of distorted image point

k_1, k_2, k_3 The radial distortion parameters of camera lens

p_1, p_2 The tangential distortion parameters

H The Homography matrix

L_i The DLT coefficient

d_{ti} The true disparity

d_i The measured disparity

d_e The error resulting from IR camera and projectors lenses distortion

δ_{tang}^{c} The tangential distortion effect of IR camera

δ_{tang}^{p} The tangential distortion effect of IR projector

δ_{rad}^{c} The radial distortion effect of IR camera

δ_{rad}^{p} The radial distortion effect of IR projector

d_{sys} The systemic depth error remaining after applying the distortion

A, B, C, D The polynomial coefficients to be determined from calibration

d The undistorted depth.

D_{3d} The descriptor of each 3D feature

D_{2d} The descriptor of the 2D feature

F_j The 3D feature information

P_k The coordinate of projected matched SIFT point to 3D point cloud

PL_i Parameters define the ith plane

$PLc_{(i+1)}$ Point cloud defines (i+1)th plane

P Point cloud generated from a depth image

$thres_{pts}$ The distance threshold defining the outlier points

ϕ Function computes the orthogonal distance between points P and plane PL_i

Ω Function sorts the detected planes PL based on the number of inliers

I Indices of sorted planes based on the point inliers

$thres_{pls}$ The distance threshold needed to filter out the identical planes

ω The function returns the orthogonal distance between two planes

R_{plane} The percentage of point cloud of the (i+1)th plane lies inside the ith plane within $thres_{pls}$ distance

PL Cell array contains parameters defining planes

$thres_{nom}$ The threshold defining the overlap between two planes

PL_{nom} The nominated planes' parameters

D_i, D_k Descriptors of features i and k, respectively

σ_{di}, σ_{dk} Descriptors of standard deviation of features i and k, respectively

S_{ik} Matching score between descriptors i and k

cov Covariance between descriptors i and k

dx_i The image gradient of sub block (i) along x direction

dy_i The image gradient of sub block (i) along y direction

f_1, f_2 Point features existing in the first and second image, respectively

$SSD_{f_1 f_2}$ The sum of squared difference distances between point features' descriptors

D_{f_1} The descriptor of point features located on the first image

D_{f_2} The descriptor of point features located on the second image

C_{l1}, C_{l2} Matched line's center point of RGB-D frames 1 and 2, respectively.

C_{n1}, C_{n2} Matched plane's center point for RGB-D frames 1 and 2, respectively

D_1, D_2 Direction vectors of matched lines between RGB-D frames 1 and 2, respectively.

N_1, N_2 Normal vectors of matched planes between RGB-D frames 1 and 2, respectively

$\hat{R}$, $\hat{T}$ Estimated camera rotation and translation, respectively

E_p, E_l, and E_n Point, line, and plane features reprojection error, respectively

PE The function that detects the peaks in a time series

G Gaussian filter that functions to smooth the gradient of y axis rotation

θ_y Rotation angle around y axis

gM Global model to be smoothed

sM Sub models divided by turned frames

$rpose$ The refined camera poses after the refinement stage

rM The refined global model after the refinement stage

S Spatial constrained information of indoor environments

Con Constrained function reinforces the predefined spatial information S

Chapter 1: Introduction

1.1 Motivation

Mapping indoor environments is important for numerous applications related to the construction and video game industries. The indoor three-dimensional (3D) model is a basic component of building an information modeling (BIM) system used to simulate the indoor environment conditions, such as temperature and humidity (Pătrăucean et al., 2015; Tang et al., 2010). Also, this model is a critical component in building virtual reality and augmented reality (Litomisky, 2012) for both gaming and industrial applications. In addition, precise floor plans of indoor environments are essential input for indoor navigation systems (Darwish et al., 2017a; Lee et al., 2012; Wang et al., 2014a; Yamazoe et al., 2012). For precise applications, a 3D model can be used as an as-built drawing (known as AB BIM) and can be used to prepare the final drawings for maintenance and operation (Pătrăucean et al., 2015). During the construction process, a 3D model can be used to monitor the state of progress (Gupta, and Li, 2017; Kahn et al., 2013; Omar, and Nehdi, 2016).

Many techniques are used to obtain 3D models of indoor environments, and some of them were already used long ago. Back in the 1840s, Aimé Laussedat proposed a photogrammetry system which, after 22 years, was accepted by the science academy in Madrid in 1827 (Jiang et al., 2008) for surveying applications. Since that time, photogrammetry technologies have been well developed for many applications in surveying and engineering. Nowadays, closed-range photogrammetry with non-metric cameras has been widely used in computer vision and surveying applications (Fathi et al., 2015; Jia et al., 2012; Mallick et al., 2014). Stereo camera systems are used to

reconstruct the 3D model (Pillai et al., 2016). However, the optical image system is computer-intensive and highly dependent on the visual features of the scene. The system has been demonstrated in indoor environments and gives a promising results (Gupta, and Li, 2017). Instead of using stereo camera systems, 3D models can be reconstructed up to a scale factor from a single camera using structure from motion (SFM) concept (Wu, 2011, 2013).

Currently, terrestrial laser scanners are a commonly used for reconstructing 3D models for both indoor and outdoor environments (Lehtola et al., 2017). The working range and accuracy depend on the applications (Geosystems, 2016). While using the system indoors, the system faces problems of mobility and surveying cost. Compared to other mapping systems (e.g., cameras, RGB-D cameras), it costs around $HK 1-2M. Normally, the surveyors need around two square meters to smoothly function and operate the laser system, and this may not be possible in some indoor environments. Recently, laser and lidar systems have been developed to match indoor environment conditions. Size, cost, and mobility—all these factors have been considered for the laser systems development. The NavVis mobile mapping system (Navvis, 2018) is one such system; it combines three lidar sensor and six cameras, and it performs SLAM to build 3D models. The system cost is still high (i.e., the cost of one lidar sensor is around $HK 50K) compared with systems using RGB-D sensors (Matterport, 2018); moreover, the NavVis system cannot work in low texture environments (e.g., underground tunnels) due to SLAM failure.

Recently, newly developed RGB-D cameras have the potential to replace time consuming and expensive indoor mapping systems (Stachniss et al., 2017). Since 2010, when the first version of a Kinect sensor was released on the market as a remote controller for video games (Kinect), numerous research works have been trying to

adopt those cameras in precision surveying applications. 3D vision can be obtained through different principles, i.e. structured light (SL) (Khoshelham, and Elberink, 2012), time of flight (ToF) (Gokturk et al., 2004), stereo triangulation (Hirschmuller, 2005), and coded aperture (Martinello, and Favaro, 2011, 2012). Many sensors are already available on the commercial market. For example, Structure Sensor (Occipital, 2014) and ASUS Xtion Pro Live (Asus, 2017) sensors are based on the SL concept while Tango (google, 2016) and Kinect version two are based on ToF, and both are available.

Adopting RGB-D cameras in surveying applications can solve the cost and mobility problems of the current indoor mapping systems. For example, Structure sensor only costs around $HK3000. This cost includes the sensor cost, SLAM software, and SDK. It works with different operation platforms (e.g., window, IOS); therefore, it is highly compatible with different systems. The sensor is originally attached to iPad device with a several APPs for instant use. Arguably, structure sensor could be the cheapest, fully mobile, indoor mobile mapping system. However, despite these valuable advantages, the sensor suffers from serious problems which prevent it from being used commercially in surveying applications. Two major problems exist in such commercial cameras: accuracy deteriorates as depth increases—creating a limited depth operation range, and the SLAM method used to reconstruct 3D models (Ahmed et al., 2015; Bose, and Richards, 2016; Camplani et al., 2013; dos Santos et al., 2016; Dryanovski et al., 2013; Hu et al., 2012; Khoshelham et al., 2013; Newcombe et al., 2011; Whelan et al., 2013; Whelan et al., 2015). The sensor basically uses IR patterns to compute the depths of objects, and for its compatibility with gaming purposes and virtual reality applications the baseline between IR sensors is very short (i.e., 6.5cm). Due to the limited pattern of the IR projector and the short baseline between IR camera

and IR projector, the produced depth from such cameras does not exceed several meters (i.e., maximum nine meters) and the optimal working range is around quarter of the maximum detected depth (i.e., two meters) because of the depth resolution restriction (Basso et al., 2014; Chow, and Lichti, 2013; Darwish et al., 2016; Herrera et al., 2012; Lachat et al., 2015). The SLAM of such cameras is basically based on visual features detected from aligned RGB and depth images. In case of distant or few point features, the SLAM system can easily drift or lose tracking, which results in the SLAM failing to reconstruct 3D model from captured RGB-D frames. The reason for the SLAM failure is lack of nearby visual points.

In light of what we know about using RGB-D cameras as 3D mapping devices, we believe that with an appropriate precision calibration method and with an adequate SLAM algorithm, the RGB-D camera can effectively replace the expensive and time-consuming mobile 3D modelling technologies used to map indoor environments.

1.2 Thesis objectives

In order to convert low cost and flexible RGB-D cameras from gaming to construction applications, two main objectives are addressed in this thesis. The main objectives are divided into sub-objectives as follows:

- Thorough calibration of RGB-D cameras to improve measurement accuracy and working range.

To achieve this objective, we proposed a novel calibration method which deals with all the sensors implemented in RGB-D cameras and their geometric relations.

    - Relative calibration between RGB camera and IR camera to precisely co-register depth and color information.

- Recalibration of the manufacturer's constants to reveal the bias from IR camera and IR projector baseline.
- Calibration of IR projector lens distortion, then model its effect on depth measurement.
- Calibration of the depth measurement in accordance with the effects of rounding-off disparity, correlation algorithm, incident angle, and depth range.

- Reducing the lost tracking rate and improving tracking accuracy by developing a new RGB-D SLAM algorithm.

We enhanced the tracking ability of RGB-D SLAM by adding more features. Those features, like lines and planes, are automatically extracted and described in the RGB-D domain. The additional features can overcome the lost tracking problem and minimize the SLAM drift. After performing SLAM, indoor structural constraints were automatically extracted and applied into the global optimization stage for further enhancement of the 3D model. This objective is divided into these minor objectives:

- Extracting, descripting, and matching of 3D features of RGB-D frames such as lines and planes.
- Developing the tracking core of SLAM to contain point, line, and plane features.
- Automatically define the structural constraints of indoor environments and correctly implement those constraints in SLAM.

1.3 Thesis contributions and outlines

The contribution of this research can be divided into two parts. The first part is related to RGB-D camera calibration, while the second part is related to the SLAM system. Concerning the RGB-D calibration contributions, a novel distortion model for both the IR camera and IR projector was proposed. The distortion model is mathematically derived from the basic distortion characteristics of both IR sensors. Also, a new calibration method for depth measurements based on the structured light concept is proposed. The method thoroughly calibrates the depth by considering several factors such as IR projector distortion, disparity rounding-off, IR sensors baseline, and incident angle. Also, the method adopts a 3D checkerboard to estimate the relative baseline between RGB and IR cameras. The calibration method and algorithms have been implemented in MATLAB, thus, an automatic calibration toolbox is designed to handle any RGB-D camera based on SL concept regardless of IR baselines.

Regarding SLAM algorithms, a new method to extract and describe 3D features from RGB-D frame is proposed. The new method uses a novel description function for line and plane features based on both RGB and depth information, so the resulting description algorithm overcomes the problems of point cloud deteriorated quality. Features extraction, description, and matching all are implemented in MATLAB for later addition to the SLAM tracking core. The SLAM is enhanced based on two aspects. The first is the tracking algorithm. Besides 2D features, 3D features are added to estimate the camera's pose. The second is a SLAM refinement stage. A new method is proposed to extract the constraints of the indoor environment based on camera pose information. The proposed SLAM method applies the global constraints stage before possibly applying a loop closure correction. The SLAM method is fully implemented in MATLAB as a post processing SLAM for RGB-D data.

The thesis is organized as follows: **Chapter one** shows a brief introduction to the thesis and gives both motivations and objectives for this research. **Chapter two** illustrates the current developments in RGB-D cameras done to adopt those cameras to surveying applications. This chapter surveys the current calibration methods and the latest depth enhancement models. It also describes the current SLAM algorithms used to produce 3D models from RGB-D data. **Chapter three** shows in detail the current calibration models used to calibrate RGB-D cameras. The chapter includes our novel distortion model for IR sensors and systematic depth error model, as well as conventional DLT and Homography calibration methods. **Chapter four** presents our novel method for calibrating RGB-D cameras. This chapter also states the calibration results of two different RGB-D cameras. Quantitative and qualitative assessment results of our calibration method based on the proposed algorithms are also provided. **Chapter five** surveys the existing features in RGB-D frames. The features are divided into two categories: 2D and 3D. This chapter also introduces a new method with a novel description function to extract and describe in order to match 3D features from RGB-D frames. This chapter also shows the impact of using both 2D and 3D features on the RGB-D frame registration results. **Chapter Six** introduces a new RGB-D SLAM method for precisely reconstructing indoor 3D models. This chapter also provides detailed descriptions of each step of the proposed SLAM, and it presents some of the key results that were used to assess the method. **Chapter seven** summarizes the conclusions and offers future recommendations.

Chapter 2: Recent development on 3D modeling using RGB-D sensors

2.1 Introduction

In this chapter, the basic components of SL RGB-D cameras and their working principles are introduced first. The calibrated parameters of SL RGB-D cameras will be clearly defined. And the effects of calibration parameters on both depth accuracy and point cloud quality will be discussed. Current calibration techniques for SL RGB-D cameras and their applications will be reviewed. Some limitations of the existing methods for RGB-D camera calibration and 3D mapping applications will be also highlighted in this chapter.

2.2 Principal of RGB-D sensors

RGB-D depth measurement based on SL concept consists of two basic IR sensors: an IR camera and an IR projector. These two IR sensors are responsible for the depth computation, and unlike the ToF, the baseline between IR camera and IR projector restricts both minimum and maximum working range of RGB-D sensors based on SL concepts. Figure 2.1 shows the basic components of an RGB-D sensor based on SL concepts.

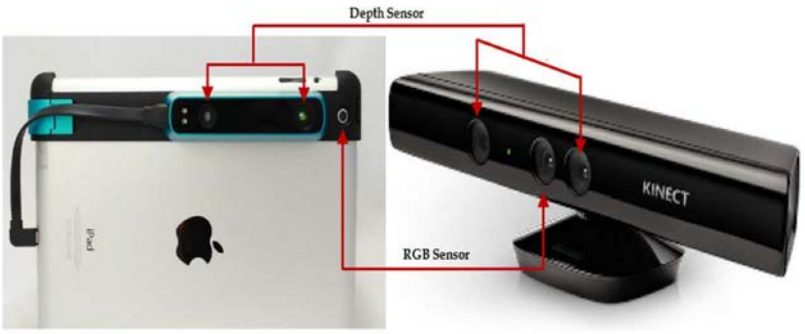


Figure 2.1: The elements of RGB-D sensors based on SL concepts: Left is S.S. and Right is Kinect (Darwish et al., 2017c)

RGB-D cameras use their IR sensors to produce the depth information of each pixel in the observed scene and they use a normal RGB camera for coloring the depth information (i.e., when reconstructing 3D models using Kinect Fusion (Newcombe et al., 2011)) or for face detection and image processing applications for gaming purposes (i.e., for use as an advanced remote controller in video games). As the depth information does not depend on the RGB camera, the RGB-D sensor can be treated as a low-cost depth sensor without RGB information. This is clearly illustrated in the latest released senor (i.e., Structure Sensor (Occipital, 2014)). Structure Sensor combines only IR camera and IR projector which can be used alone without an RGB sensor. The manufacturer has built the sensor with features enabling it to be attached to any mobile device (e.g., iPad), where it can use any existing RGB sensor in a mobile device to color the point cloud and with features that prepare it for advanced implementation in RGB-D SLAM systems (Tang et al., 2016).

Two different stages are followed to compute the real pixel depth from RGB-D sensors. The first stage is the in-factory preparation stage (Zhang, and Zhang, 2014). At this stage, the sensor is placed parallel to a planar surface and positioned at a predefined distance (normally one meter). This distance is known as the distance of reference plane (Z_0). Next, the IR camera and projector were switched on and the

sensor started capturing the image of the reference planar surface. The output of this stage is the reference pattern used to produce the disparity which can be converted to the depth for each pixel (Shpunt et al., 2010).

The second stage is real-time producing depth information. At this stage, the RGB-D camera simulates the in-factory stage, both IR sensors are switched on and the IR camera captures the projected pattern from the IR projector through the reflection of an object. By comparing the reflected captured pattern and the corresponding reference pattern stored in the sensor firmware, the sensor calculates the current depths of observed pixels. The concept of depth perception using an RGB-D sensor is illustrated in Figure 2.2. Assuming that the sensor has a factory-installed calibration process (Zhang, and Zhang, 2014), the manufacturer's parameters are represented as follows: 1) Reference plane depth (Z_0), 2) Predefined standard pattern ($x_{i,0}^c$) of the existing feature point (Q_i), 3) Focal length of IR sensors (f), and 4) Baseline between IR camera and IR projector (w) are known and stored in the sensor firmware (Khoshelham, 2011; Khoshelham, and Elberink, 2012). In addition to the RGB-D measurements (i.e., the captured IR pixel location (x_i^c) of the feature point (Q_i) projected by IR projector), the depth of the imaged feature point (Q_i) can be computed using the triangulation geometry from Figure 2.2.

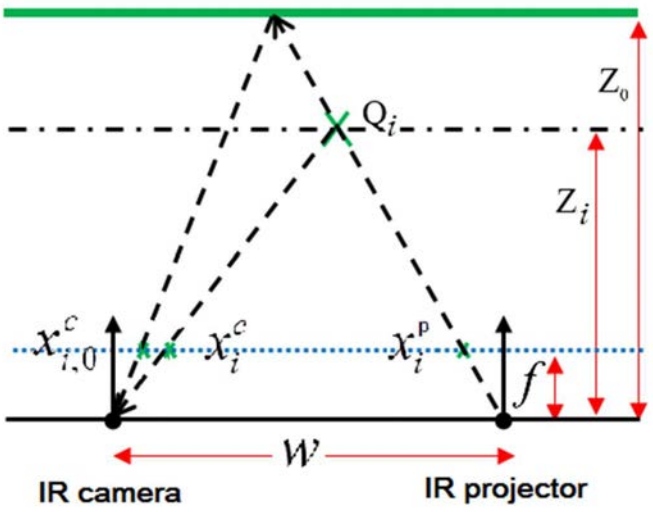


Figure 2.2: RGB-D sensor depth perception concept (Darwish et al., 2017c)

Based on the factory-installed parameters and the measured location of the feature point (Z_i), the depth of the feature (Z_i) can be determined by triangulation (Khoshelham, 2011; Yamazoe et al., 2012). Equations (2.1) and (2.2) illustrate the relationship between the depth of the feature point and the measured and in-factory parameters.

$$x_{i,0}^c = x_i^p + fw/Z_0 \tag{2.1}$$

$$x_i^c = x_i^p + fw/Z_i \tag{2.2}$$

The difference between the IR standard pattern location ($x_{i,0}^c$) and the measured projected pattern of the IR projector is called the disparity (d_i), where ($d_i = x_i^c - x_{i,0}^c$). In applying this concept to (2.1) and (2.2), the relationship between the disparity (d_i) and the unknown depth can be written as:

$$Z_i = \frac{fw}{\frac{fw}{Z_0} + d_i} \tag{2.3}$$

where

Z_i is the perpendicular distance between the IR sensor's baseline and feature point

f, which is the focal length of the IR camera or IR projector

w the baseline between IR camera and IR projector

Z_o the depth of standard plane

d_i the measured disparity

Instead of measured disparity, the RGB-D sensor provides a normal disparity (d_i^n). The normal disparity is a normalized value ranging between 0 and 2047. Two linear factors (α and β) are used to convert the measured disparity into normal disparity. The relationship between normal disparity and measured disparity can be written as $d_i^n = \frac{1}{\alpha}(d_i - \beta)$, while replacing the disparity in (2.3) by the normal disparity, equation (2.3) can be rewritten as follows:

$$Z_i = \frac{1}{\left(\frac{1}{Z_0} + \frac{\beta}{fw}\right) + \left(\frac{\alpha}{fw}\right) d_i^n} \tag{2.4}$$

By combining all the constants in (2.4) and assigning them to two factors a and b, equation (2.4) can be rewritten as follows:

$$Z_i = \frac{1}{(a) + (b) d_i^n} \tag{2.5}$$

where a and b are constants related to the in-factory calibration process and the linear factors which convert measured disparity to normal disparity. The a and b factors can be represented as follows:

$$a = \left(\frac{1}{Z_0} + \frac{\beta}{fw}\right) \tag{2.6}$$

$$b = \left(\frac{\alpha}{fw}\right) \tag{2.7}$$

For every pixel in a SL RGB-D sensor, the sensor measured the actual disparity then used the predefined in-factory parameters to compute the depth of each pixel. The previous formulas, provided in (2.1) to (2.7), showed that all the effects of in-factory parameters can be reduced to dominant factors a and b. This is very important when the user re-calibrates the in-factory parameters of RGB-D sensors. The point cloud ($X\ Y\ Z$) of the observed scene can be computed from the depth image by using the geometric parameters (focal length (f), and the principal point (c_x^c, c_y^c) of the IR camera as follows:

$$\begin{aligned} Z_i &= \frac{1}{(a) + (b)d_i^n} \\ X_i &= \frac{(x_i^c - c_x^c)Z_i}{f} \\ Y_i &= \frac{(y_i^c - c_y^c)Z_i}{f} \end{aligned} \tag{2.8}$$

For precise applications (mm-level precision for near depth or cm-level precision for far depth), in-factory parameters as well as the relative baseline between IR and RGB cameras must be recalibrated precisely. Beyond the calibration of the geometric parameters of the sensors involved in the RGB-D cameras, a rigorous calibration of the depth information must be able to handle distortions from both IR camera and IR projector.

2.3 Influence of IR sensors baseline on depth precision

The disparity concept ($d_i = x_i^c - x_{i,0}^c$) is the principal of computing the depth in such sensors. The definition of disparity is the difference between the predefined pattern and the reflected pattern from the objects, stored in RGB-D firmware and captured by the IR camera, respectively. Therefore, the computed disparity is influenced by the distortion of both IR camera (receiver) and IR projector (emitter). This means that the depth distortion results from both IR camera and IR projector. In addition to the IR patterns, the manufacturer's geometric constants (i.e., baseline between IR camera and projector) have a great effect on the sensor's depth precision.

The 6DoF geometric distance between IR camera and IR projector has a great influence in depth precision of RGB-D sensors. To investigate the effect of the baseline on depth precision, the covariance error propagation concept was adopted to depth observation equation (2.3). The relationship between disparity and depth variances can be stated as follows:

$$\sigma_z = \frac{Z^2}{fw}\sigma_d \tag{2.9}$$

where

σ_z and σ_d are the precision of depth and disparity, respectively

Z is the calculated depth

f is IR sensor focal length

w is the baseline between IR camera and IR projector

We assume that we have two SL RGB-D sensors and that they have different physical properties, especially different IR sensors' baselines. Figure 2.3 shows two different

RGB-D sensors capturing the same scene from the same distance. The figure illustrates the difference in depth precision against the IR sensors' baselines.

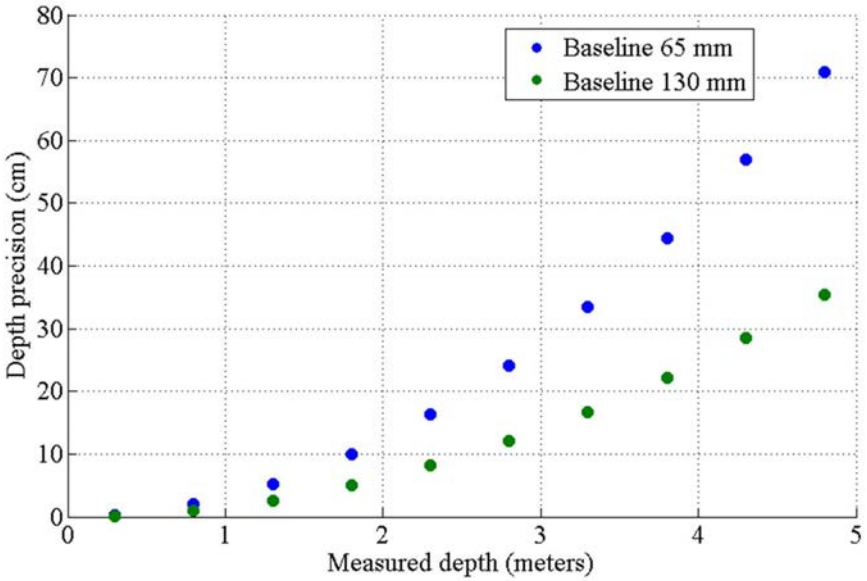


Figure 2.3: Depth uncertainty versus IR sensors baselines

2.4 Review on current RGB-D calibration methods

RGB-D sensors usually have visual and depth information for each pixel in the captured images. Thanks to this valuable information, numerous applications for RGB-D sensors have been proposed since 2010. In order to enhance both the visual and depth information of RGB-D cameras, numerous RGB-D calibration methods have been investigated for their potential to thoroughly calibrate both RGB and depth information of such sensors (Chow, and Lichti, 2013; Haggag et al., 2013; Herrera et al., 2012; Shibo, and Qing, 2012; Wang et al., 2014a; Wang et al., 2014b). The calibration procedure of RGB-D sensors depends on the application accuracy requirements. For a specific application, a different level of data quality is required. For example, 3D computer vision applications are mostly concerned with 3D shape consistency; Virtual and Augmented Reality (VR & AR) applications are highly

focused on tracking and coloring information; and robot guidance applications give more attention to depth quality, especially near range depths, to create the Six Degrees of Freedom (6DoF) navigation data for the robot.

Due to the nature of indoor environments (i.e., the systemic painting color and less distinctive 3D features), some indoor environments can be treated as 2.5D (Turner et al., 2015; Turner, 2015) instead of rich 3D (e.g., for the human body and objects). For 3D reconstruction of indoor environments, usually the calibration procedure deals with the full range calibration as well as the distortion effect from the three sensors involved in the RGB-D cameras. Arguably, the RGB-D calibration for 3D indoor reconstruction can be assumed to integrate all calibration procedures because of the required high accuracy (cm-level precision) of 3D applications. The depth accuracy must be fully calibrated and corrected up to cm-level accuracy. While adopting RGB-D sensors to reconstruct the 3D models of indoor environments, usually the RGB-D SLAM algorithm is utilized to build 3D model. RGB-D SLAM uses the corresponding depth of detected RGB features to perform tracking; therefore, the relative calibration between IR camera and RGB camera must achieve pixel-level accuracy.

The calibration process of RGB-D cameras is divided into two major parts; however, both parts can be carried out in one pipeline (Chow, and Lichti, 2013; Herrera et al., 2012). The first part is related to the external baseline calibration between RGB and IR cameras; thus, this part is entirely a stereo calibration problem between RGB and IR cameras (Zhang, and Zhang, 2014). The second part is related to the depth calibration, which is mainly concentrated on the modeling of the depth distortion and systemic error (Darwish et al., 2016). The following sections will discuss the existing procedure to conduct both parts.

2.4.1 RGB-IR cameras baseline calibration

The baseline calibration between RGB and IR cameras can be considered a stereo calibration problem. The main purpose of this calibration is to find the relationship between RGB and IR camera working reference frames. As the stereo calibration problem has been solved for closed range photogrammetry camera calibration using the Photogrammetric Bundle Adjustment (PBA) concept (Abdel-Aziz, and Karara, 1971; McGlone, 1989) with precise compensation distortion models (Fryer, 1989; Fryer, and Brown, 1986)), the algorithm is well developed and can be applied to any stereo camera, even to small baseline camera pairs. When using commercial non-topographic cameras (Harley, 1967), instead of PBA, the Direct Linear Transform (DLT) method (Abdel-Aziz, and Karara, 1971) is adopted into a computer vision community (known as Homography calibration), which can be simplified to use a 2D checkerboard for calibration (Bouguet, 2000; Heikkilä, and Silvén, 1997; Morvan, 2009; Zhang, 2000). Lack of information regarding CCD size in commercial RGB-D camera lenses and the acceptable accuracy of DLT calibration results are two reasons given for the superior ability of DLT to calibrate RGB-D cameras, especially in computer vision and robotics applications.

Despite the pinhole camera model solution differences (i.e., solving linear (Abdel-Aziz, and Karara, 1971) or nonlinear equations systems (Zhang, 2000)), the implemented observations in the calibration procedure strongly impacts the calibration result (Darwish et al., 2017c). Depth image or IR image along with the RGB image are used to calibrate the relative baseline between IR camera and RGB camera. The physical meaning of the distortion parameters as well the focal length must be clearly stated, as the output calibration parameters from the relative calibration step will be used in the depth calibration step. Three ways can be followed to calibrate RGB-D

cameras. **Firstly,** an IR camera image along with an RGB image can be used to precisely calibrate the baseline between RGB and IR cameras; however, the distortion of the IR projector is neglected in this case, and depth distortion cannot be thoroughly calibrated. **Secondly,** depth image produced from IR camera and IR projector can be used to thoroughly calibrate the RGB-D camera's lenses; however, the distortion of the depth image must be adjusted to include the effects of both the IR projector and the IR camera. **Finally,** the RGB-D camera can be calibrated using RGB camera, IR camera, and IR projector data. Due to the unavailability of IR projector data, the calibration method should optimize IR projector data during the calibration process. Based on these three ways, the stereo calibration process of an RGB-D camera can be divided into three main categories.

The first category uses the images of the RGB and IR cameras captured for a checkerboard and solves the camera pinhole model by using the image and ground points coordinates (Bouguet, 2000). In this method, the IR projector is switched off (Tang et al., 2016) or covered by tape (Macknojia et al., 2013) in the early version of the Kinect v1 sensor. The output of this method is a complete set of each camera's geometric parameters: focal lengths, principal point, distortion parameters, and the relative 6DoF baseline between the RGB and IR cameras. Normally, a 2D checkerboard is used for the calibration, and Zhang's method (Zhang, and Zhang, 2011; Zhang, 2000) is utilized with the pinhole camera model as the core mathematical model and with a maximum likelihood estimation method to estimate geometrical parameters of both RGB and IR cameras. The output calibration results from this category can only apply to the RGB camera to project the color information to the depth information (i.e., point cloud coloring and co-registration between RGB and depth images). As the depth image results from both the IR camera and the IR

projector, the estimated distortion parameters of the IR camera cannot fully consume the depth distortion.

The second category is designed to overcome the problem of IR projector distortion by using an empirical depth distortion model based on depth measurements. It adopts a depth image instead of an IR image (by keeping IR projector on) and an RGB image. This method uses different checkerboards to simultaneously work for both color and depth images (Gui et al., 2014; Herrera et al., 2012; Raposo et al., 2013). It applies the pinhole camera model to both RGB and depth images; however, the depth image is created from both IR sensors. Herrera C et al. (2011) uses an ordinary checkerboard attached to a planar surface, then the Homography method was adopted to the extracted point features from depth and color images. Only four points extracted from the depth images are used to estimate the internal parameters of the IR camera. In this method, the depth distortion of the depth image is not considered. Subsequently, the method was modified with an empirical distortion model for depth sensors (Herrera et al., 2012; Raposo et al., 2013). The major problem of this method is that the estimated depth geometric parameters are estimated from only four points. To overcome those issues, different approaches were proposed for using checkerboards that produce more accurate points. For example, Herrera et al. (2012) added a high-resolution camera which was synchronized with the Kinect sensors. The high-resolution camera is responsible for providing the true depth measurement needed to calibrate the depth distortion and depth bias. Jung et al. (2015) adopted a 2.5D checkerboard to enrich the points in the depth images and extract them automatically. A 3D calibration environment with manually selected points in the depth images were adopted by Gui et al. (2014) and Khoshelham et al. (2013). Lastly, Liu et al. (2012) adopted a 3D line based calibration. In this category, the distortion model depends on the depth error

behavior; thus the error performance of the RGB-D sensor must be investigated in advance before carrying out the calibration procedure.

The third category applies the pinhole camera concept separately to the IR camera and projector as well as to the RGB camera (Chow, and Lichti, 2013); consequently, the calibration tries to overcome the distortion problem of IR projectors. As it is difficult to obtain an IR projector's raw data, the author estimated such data from the disparities and approximated a baseline between the IR camera and projector. The two main limitations of this method are the data dependency of IR projectors and the unreliability of estimated IR distortion parameters. Other research deals with the implementation of a new mathematical distortion model for combined IR camera and projector (Yamazoe et al., 2012). The model only compensates the radial distortion effect of both IR sensors. It concentrates on depth distortion calibration using the fitted plane as a reference depth.

Calibration of the relative baseline between RGB camera, IR camera, and the internal distortion of both RGB camera and IR sensors is crucial as the following step of the calibration procedure of depth is highly dependent on the estimated parameters of the RGB and IR cameras baseline calibration step. For the second category, the distortion parameters resulting from this calibration step can be applied to the depth image. However, in the first and third categories, the estimated distortion cannot be directly applied to the depth distortion as they are estimated based only on the IR camera or on both the IR camera and the IR projector, for first and third categories respectively.

2.4.2 Depth calibration

To fully calibrate RGB-D cameras, after calibrating the relative baseline between the RGB and IR cameras, any systematic depth errors and depth distortion must be

investigated and modeled. This section will highlight the traditional methods for calibrating the depth of RGB-D camera.

The main characteristic of RGB-D cameras is their ability to simultaneously deliver the registered depth of an RGB image pixel (if the relative calibration was obtained). In fact, this beneficial characteristic is the major reason preventing RGB-D cameras from being applied in precise surveying applications. To overcome the lack of precision in depth information, systematic depth error and depth distortion must be robustly calibrated. Enormous research work has been done. The common methods use the concept of images (depth and color) rather than sensors (RGB, IR projector, IR camera) to investigate the distortion and systematic depth error of RGB-D sensors. These methods are largely applied to applications involving pattern recognition and single frame data interpretations (Han et al., 2013). When adopting RGB-D cameras in precise applications, both far and near depth errors should be investigated and modeled.

After reviewing and summarizing the research work dealing with depth calibration methods, these methods can be divided into three major groups (Darwish et al., 2017c). The first group deals with each sensor involved in RGB-D camera (RGB camera, IR camera, IR projector) separately. This method assumed that the pinhole camera model is valid for three different sensors. This method calibrates the relative baseline between the IR and RGB cameras and calibrates the depth distortion in one mathematical model known as PBA (Chow, and Lichti, 2013). This method can individually model the IR camera and IR projector distortions; nevertheless, the distortion models lack rigorous IR projector distortion parameters. It modeled the systematic depth error as a function of radial distortion parameters. It can delete the artifacts of depth and enhance depth precision. The method was only applied to calibrate the normal working range of the

RGB-D cameras; it does not investigate the calibration of both near and far ranges of the RGB-D cameras.

The second group calibrates the distortion and systemic error of depth by adopting an empirical distortion model for depth sensors (combining both IR camera and IR projector). This method begins with the disparity image provided by IR sensors and calibrates the manufacturer's parameters (Herrera C et al., 2011). This method added a high-resolution camera for calibration with the system. Based on the pinhole camera model, the method calibrates the external RGB camera and RGB sensor of the RGB-D camera and calibrates the distortion model for the depth sensor (Herrera et al., 2012; Raposo et al., 2013). The distortion model not only compensates for the distortion of the IR camera and IR projector, but it also models systematic depth error. When the empirical distortion model was applied to the Kinect v1 sensor it produced a significant improvement in near range depth precision. As the distortion model is empirical, it is highly dependent on the depth error behavior of the sensor, thus the baseline between the IR camera and IR projector. The main limitation of the distortion model is its incompatibility with RGB-D cameras that have different baselines between IR sensors.

The third group is completely concentrated on minimizing depth error by combining the distortion of IR sensors with the systematic depth error (Haggag et al., 2013), or by separating distortion from systematic error (Basso et al., 2014). Several empirical models based on disparity or depth information have been discussed by Mallick et al. (2014). The model can work well with each unique sensor; however, the repeatability and the durability of the error model must be checked especially with different work environments (e.g., indoor or outdoor mapping) (Andújar et al., 2017). Recently, a new error model based on the covariance propagation concept for measurement models of RGB-D cameras was introduced in order to improve the error model (Lachat et al.,

2015; Pagliari, and Pinto, 2015). Since the depth error in outer margins of the depth image cannot be fully modelled, the error model is used to reconstruct a 3D model for small objects. Table 2.1shows the comparison between the existing calibration methods used to calibrate both baseline between RGB camera and IR camera and depth sensor.

Table 2.1: Calibration methods requirements and algorithms.

Category of calibration method	RGB camera-IR camera	RGB camera-Depth sensor	RGB camera-IR projector-IR camera	Ours
Captured data	Stereo pairs of IR image (IR projector-off) and RGB image	Stereo pairs of Depth image (IR projector-on) and RGB image	Stereo pairs of IR image (IR projector-off) and RGB image	Stereo pairs of IR and RGB images + Stereo pairs of RGB and Depth images
Algorithm	Homography	Homography	Bundle adjustments	Direct Linear Transform
Depth distortion models	Not applied	Empirical (Depend on the sensor behavior)	Browns models for Both IR camera and IR projector	Combined IR camera and IR projector Browns distortions effects
Depth systematic model	Not applied	Not applied	Modeled as function of radial distortion parameters	Estimated as 3rd order polynomial model
Depth manufacturer constants	Not applied	Applied as two factors (a, b)	Applied separately (f, w_d, w_r, Z_0)	Applied as two factors (a, b)
Limitations	Has limitations on baseline accuracy between RGB and IR cameras	Has limitations on baseline accuracy between RGB and IR cameras	Has a limitation on the reliability of IR projector distortion parameters	Automation process for full range application
Applied for full depth range	Not applied	Not applied	Not applied	Applied
Examples	Herrera C et al. (2011), Kim et al. (2015), and Zhang and Zhang (2011)	Herrera et al. (2012) and Raposo et al. (2013)	Chow and Lichti (2013)	Darwish et al. (2017)

To thoroughly calibrate RGB-D cameras' depth information, depth calibration methods should follow the following pipeline. **Firstly**, the manufacturer's parameters involved in depth computation concept must be calibrated to ensure that the systematic error resulting from an inaccurate baseline between IR camera and projector is recognized and corrected. **Secondly**, the remaining error related to the distortion effect of the IR camera and projector must be handled and mathematically modeled independent of the baseline between the IR camera and projector. **Finally**, the

undistorted depth error remaining after correcting for distortion and systematic error (due to calibration) should be modeled for further enhancement of depth precision for precise (cm-level of accuracy) applications.

2.5 RGB-D sensor applications in surveying and mapping

Surveying engineers and researchers have tried to switch RGB-D cameras from gaming purposes to surveying and mapping applications (Bell, and Gausebeck, 2014; Bell et al., 2016). Surveying applications, such as 3D modeling of existing structures, indoor base maps, and BIM models, can be done using RGB-D cameras even in static or kinematic mode (Lehtola et al., 2017). However, the sensors can produce better models in the post-processing mode (Halber, and Funkhouser, 2017). Major research has concentrated on adopting those sensors in kinematic mode. This research is done because of the potential **a)** to decrease surveying time **b)** to produce real or near real time 3D models of indoor environments (Dai et al., 2017), and **c)** to use those cameras to replace IMU in indoor navigation (Chow et al., 2014).

The Kinect Fusion system (Newcombe et al., 2011) was the first trial to reconstruct the 3D model from the RGB-D data. This opened up research into using these low-cost sensors to produce 3D rich models to fit surveying applications. The system depends on the depth information with ICP algorithm to register successive RGB-D frames ignoring the visual information of RGB images. The system uses only the RGB data for coloring the final 3D model; it does not apply the loop closure concept (e.g., g2o (Kümmerle et al., 2011) or bundle adjustment (Dai et al., 2017)). Two major disadvantages plagued the Kinect fusion system: computational cost and drift error. A lot of research works have been done to develop the existing system. Whelan et al. (2015) used volumetric fusion and implemented GPU to achieve real time SLAM

performance. This method uses both photometric and geometric information to estimate the camera pose in addition to successively searching for loop closure to enhance the 3D reconstructed model (Whelan et al., 2013; Whelan et al., 2015). Zeng et al. (2012) used GPU implementation with an octree based structure for voxel to reduce memory consumption and increase mapped volumeZeng et al. (2012). This system can map an area eight times larger than the Kinect fusion system can map. Chen et al. (2013) proposed a sparse data structure to extend the mapping volume as well enrich the fine details. Unfortunately, all the above methods still suffer from the inevitable drift of camera pose (Mur-Artal, and Tardos, 2017).

For visual features in RGB-D frames, another SLAM system is proposed for reconstructing a 3D model from RGB-D data. Henry et al. (2010) defined the basic visual RGB-D SLAM system by considering that it uses both ICP and visual features combined with loop closure correction. Several research works considered adding all other possible features to enhance RGB-D SLAM performance. They started by adding simple matched visual features (e.g., SIFT (Cornelis, and Van Gool, 2008), SURF (Bay et al., 2008)) and by applying the concept of SFM (Koenderink, and Van Doorn, 1991) to recover the relative transformation between each successive RGB-D frame. dos Santos et al. (2016) used a disparity-based model with maximum stable color region to estimate the relative movement between two successive RGB-D frames.

Regarding to visual RGB-D SLAM concept, many research efforts have explored the possibility of integrating different algorithms and different methodologies to enhance RGB-D SLAM performance for specific applications (Stachniss et al., 2017). To apply a visual-based RGB-D SLAM with continuous searching for loop closure optimization between each key frame, investigating the possibility of closing current frame to be close to the previous frames. This procedure was introduced as DVO SLAM (Kerl et

al., 2013). Recently, Dai et al. (2017) designed a system which can handle on-the-fly RGB-D SLAM system by implementing sparse points and dense model optimization.

A plethora of SLAM algorithms have been published (Stachniss et al., 2017) with specific performance reports based on available observations, details regarding the optimization technique used, and the identifications of applications (e.g., surveying, robotics and navigation, indoor and outdoor navigation, looped environments). Endres et al. (2012) proposed a system that implements the visual matched features with local loop closure using g2o (Kümmerle et al., 2011). The system was tested on a room environment with many distinctive visual near-depth features. The system gave precise results in both modeling and navigation. Fioraio and Konolige (2011) used bundle adjustment with ICP (Besl, and McKay, 1992b) and used graph optimization for final pose optimization. The common limitation of these methods is that the operation distance should be less than three meters and the number of matching features should be more than five to reliably recover pose information. When the matched features are farther from the camera (i.e., depth is greater than three meters) and the scene lacks distinguishing visual details, the visual SLAM system can easily fail.

Instead of depending on visual features to compute the relative camera pose, the Edge RGB-D SLAM (Bose, and Richards, 2016) introduces edge detection based on depth images with the ICP algorithm to evaluate the relative pose. This method can work well in 3D spaces with 3D lines and edges even if they have less visual features. The main constraints of this method are the mapping speed and local minima problem of ICP, as the edges were extracted from depth images without further matching.

As the point clouds produced from depth images are noisy, especially if the points used have a depth more than three meters, the classical way to detect and extract and match 3D features (Díez et al., 2015) has not yet been implemented because the descriptor of

the 3D features is based on the surrounding structure of the 3D features and its greatly affected by the depth noise. Hsiao et al. (2017) used the planar constraint to decrease the drift problem of the visual SLAM system for a 30 Hz frame rate. To overcome the RGB-D depth precision problem of remotely matched visual points, integration of the SFM technique and the RGB-D SLAM system was carried out by many research works (Concha, and Civera, 2017; Dai et al., 2017; Kerl et al., 2013; Melbouci et al., 2015; Stückler, and Behnke, 2012). For example Kerl et al. (2013) minimized both geometric (depth) and photometric (color) distances with the g2o algorithm as a global optimization container, the system achieved a 3cm error average compared with a 4cm error average for the MRSMap system (Stückler, and Behnke, 2012).

Other research works have concentrated on the offline enhancement of RGB-D SLAM performance. Those studies have mainly focused on surveying applications and 3D model reconstruction. Halber and Funkhouser (2017) introduce an off-line planar constraint method to refine the reconstructed 3D model by adopting the predefined geometry of the model (e.g., orthogonality, parallelism). The system can work well in reconstructing large indoor spaces; however, the user must provide the system the with pre-known structural data for the environment to be surveyed. The system mainly used the planar constraints to iteratively refine the global 3D model. Tang et al. (2016) presented a method for integrating the concept of SFM with the visual RGB-D SLAM system to produce consistent and complete 3D models of close and far range surveying. To some extent, this method can be applied in outdoor environments. Darwish et al. (2017b) introduced a method for computing the camera's pose from 3D line features. It can be the first attempt to match the depth features (lines and edges) extracted from RGB-D frames before deploying them in the ICP algorithm, which can effectively solve the local minima problem of ICP.

Arguably, on existing disadvantage of RGB-D SLAM algorithms is the applying of the manually extracted conditions from the mapped indoor environments (e.g., a wall is always perpendicular, a ceiling is perpendicular to a wall, a floor is parallel to a ceiling), and this is a time-consuming step, particularly in large indoor environments. The main disadvantage is that the extraction stage demands extensive interaction from the users.

2.6 Summary

Recently, many research efforts have been done to adopt RGB-D cameras to surveying applications. For reconstructing 3D models from captured RGB-D frames, the procedure is divided into three stages. The first stage is the data collection, the second stage is the tracking algorithms which include both feature tracking and tracking optimization models. Thirdly, the refinement stage is applied on post-processing.

In the first step, the data collected from RGB-D cameras always suffers from depth distortion, lens distortion, systematic depth error, and the limited working range of those cameras (usually three meters). A rigorous calibration must be carefully applied to RGB-D cameras to produce the best quality of both RGB and depth images. The depth range of the RGB-D cameras is also the major problem of existing calibration procedure, as the conventional calibration methods can calibrate the depth up to three meters only; however, RGB-D cameras can produce depth information up to nine meters. This is possible for two reasons: they ignored the calibration of the baseline between the IR camera and projector and calibrated depth by only considering the depth distortion, ignoring the rounding-off and correlation algorithm uncertainty of disparity, which together are major causes of depth far range distortion. The calibration parameters of RGB-D cameras are **1)** the relative baseline between the RGB and IR

cameras, **2)** the geometric parameters of the RGB camera and its distortion coefficient, **3)** the systematic depth error, **4)** the distortion and geometric parameters of both the IR camera and projector.

In the second and third steps, the features implemented in the camera pose computation have a great impact of the camera pose quality. 3D line and plane features exiting in both RGB and depth images must be detected, extracted, described, then matched to solve the local minima problem of ICP algorithm. To simplify the implementation of constrained conditions, those conditions must be automatically extracted from the observation data. The RGB-D SLAM process and its tracking core strongly impact the reconstructed 3D models and their accuracy.

Chapter 3: RGB-D Cameras Calibration Models

3.1 Introduction

The main problems affecting RGB-D calibration are lens distortion and the depth error modelling of RGB-D sensors. This chapter will survey the current calibration models used to simultaneously calibrate the distortion effect of both the IR camera and projector along with the distortion effect of the RGB camera. The chapter will also introduce the basics about the Direct Linear Transform (DLT) and Homomorphy methods used to solve the calibration problem of non-metric camera based on pinhole camera model. Potential combination between stereo cameras calibration and depth distortion calibration models will also be discussed. Finally, the chapter will address a newly proposed distortion model which compensates for both IR camera and IR projector distortion effects.

3.2 RGB and IR stereo cameras calibration

The RGB-D camera combines data from three sensors: the RGB and IR camera as well as the IR projector. The depth is reported according to the IR camera reference frame while the RGB image is captured according to the RGB camera reference frame. To accurately align the depth pixel to the corresponding RGB pixel, a precise calibration of the baseline between the IR and RGB cameras must be obtained. Moreover, the distortion of both lenses should be corrected. The stereo calibration method is a commonly used method for obtaining both geometric camera parameters and an external baseline between two-fixed camera systems. Due to the unavailability of the initial parameters of the camera lens (e.g., CCD size), Homography and DLT methods are the widely used methods to calibrate commercial cameras. The following sections

will illustrate the concepts of pinhole camera model in addition to Homography and DLT calibration methods.

3.2.1 Pinhole camera model

The pinhole camera model is a widely applicable model used in close-range photogrammetry (Raposo et al., 2013). Figure 3.1 shows the relationship between image coordinate system and ground coordinate system.

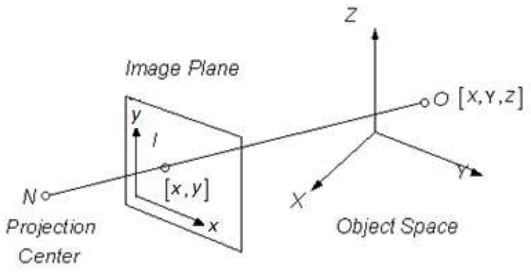


Figure 3.1: Camera coordinate system versus object coordinate system definitions.

The pinhole camera model expresses the relationship between the image point coordinates (x, y) and the corresponding ground point coordinates (X, Y, Z).

$$s[x \quad y \quad 1] = [X \quad Y \quad Z \quad 1]\begin{bmatrix} R \\ T \end{bmatrix}[K] \tag{3.1}$$

where,

s the scale factor

x, y the image point coordinates in pixels

X, Y, Z the ground point coordinates

R 3x3 rotation matrix

T 3x1 translation vector; where $T = [dx \quad dy \quad dz]$

K 3x3 intrinsic matrix $K = \begin{bmatrix} f_x & e & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$

f_x and f_y the focal length in pixels

c_x and c_y the coordinate of the principal point in pixels

e the skew between x and y direction

To simplify, treat the image point coordinates as $p = [x \quad y \quad 1]$ and the corresponding ground point coordinates as $P = [X \quad Y \quad Z \quad 1]$. Thus, equation (3.1) can be simplified as,

$$sp = P \begin{bmatrix} R \\ T \end{bmatrix} [K] \tag{3.2}$$

Equation (3.2) is the fundamental mathematical model used to calibrate both internal and external parameters for a specific camera. If the ground control points and corresponding image point coordinates are known, all geometric parameters related to camera lens can be estimated. As camera lens suffer from distortion effects, radial and tangential distortion models (Fryer, 1989; Fryer, and Brown, 1986) are proposed to compensate for the effect of both manufacturing imprecision (tangential distortion (3.3)) and poor quality lens material (radial distortion (3.4)). Usually, Brown's model is used to describe both distortion patterns as follows:

$$x_d = x(1 + k_1 r^2 + +k_2 r^4 + k_3 r^6)$$
$$y_d = y(1 + k_1 r^2 + +k_2 r^4 + k_3 r^6) \quad (3.3)$$

$$x_d = x + \left(2yp_1 + p_2(r^2 + 2x^2)\right)$$
$$y_d = y + \left(2xp_2 + p_1(r^2 + 2y^2)\right) \quad (3.4)$$

where

x_d and y_d are the distorted image point coordinates

x and y are the coordinates of the free distortion points

k_1, k_2, and k_3 are the radial distortion parameters

$p_1$and p_2 are the tangential distortion parameters

Equations (3.3) and (3.4) represent the radial and the tangential distortion effects, respectively. The full vector of distortion models is $[k_1\ k_2\ k_3\ p_1\ p_2]$, which are treated as the internal parameters for the camera. The Homography and DLT methods are two different techniques that can solve (3.2) and obtain both the internal parameters of camera and the external parameters of each captured image.

3.2.2 Homography calibration method

The Homography method was proposed by Zhang (2000) and implemented in MATLAB (Bouguet, 2000). The procedure is discussed in detail by Morvan (2009). The procedure estimates the Homography matrix and then divides it into external and internal parameters for each camera based on the pre-known properties of rotation matrix. The pinhole camera model is used as a cost function. It is minimized using least square method during the calibration procedure. The method can be divided into three steps. The first step is the geometric camera calibration stage in which focal

length, principal point, and distortion parameters are estimated. The second step is the estimation of the external baseline between two stereo cameras. The third stage is the global refinement stage.

3.2.2.1 Estimating the internal parameters

Combining both internal and external parameters in one matrix which represents the relationship between image frame and the corresponding ground coordinate system, an equation (3.2) can be rewritten as follow:

$$sp = PH \tag{3.5}$$

where H is the Homography matrix: $H = \begin{bmatrix} R \\ T \end{bmatrix} [K]$

The Homography matrix is a 4x4 matrix when the calibration adopts 3D ground control points, but it can be reduced to a 3x3 matrix ($H = [h_1; h_2; h_3]$) for a 2D checkerboard as the Z coordinate is zero. We can eliminate the scale factor by just applying the cross product of both sides of (3.5) by vector p, The results are shown in (3.6), where i refers to the point number.

$$p_i \times (sp_i) = p_i \times (P_i H) = 0_3 \tag{3.6}$$

By applying the cross product and knowing the linear combination of (3.6) as, $h_1 P_i = h_1^t P_i^t$ we can rewrite equation (3.6) as:

$$p_i \times (P_i H) = p_i \times \begin{bmatrix} h_1 P_i \\ h_2 P_i \\ h_3 P_i \end{bmatrix} = \begin{bmatrix} 0 & -1 & y_i \\ 1 & 0 & -x_i \\ -y_i & x_i & 0 \end{bmatrix} \begin{bmatrix} h_1 P_i \\ h_2 P_i \\ h_3 P_i \end{bmatrix} = 0_3 \quad (3.7)$$

$$p_i \times (P_i H) = p_i \times \begin{bmatrix} h_1 P_i \\ h_2 P_i \\ h_3 P_i \end{bmatrix} = \begin{bmatrix} 0 & -P_i^t & y_i P_i^t \\ P_i^t & 0 & -x_i P_i^t \\ -y_i P_i^t & x_i P_i^t & 0 \end{bmatrix} \begin{bmatrix} h_1^t \\ h_2^t \\ h_3^t \end{bmatrix} = 0_3 \quad (3.8)$$

Equation (3.8) is the fundamental formulae for estimating the Homography matrix (H). The system can be represented as $AH = 0$ where $A = [0 \quad -P_i^t \quad y_i P_i^t; \quad P_i^t \quad 0 \quad -x_i P_i^t; \quad -y_i P_i^t \quad x_i P_i^t \quad 0]$ thus, Singular Value Decomposition (SVD) (Pilu, 1997) or Principal Component Analysis (PCA) (Corke, 2011) can be adopted to solve (3.8).

The Homography matrix (H), according to its definition, already contains the camera's internal parameters (K), the rotation matrix $R = [r_1 \quad r_2 \quad r_3]$, and the translation vector $T = [dx \quad dy \quad dz]$. T and R are combined and they are referred to as the camera pose. For 2D checkerboards, the constraint of $Z = 0$ is applied; thus, the Homography matrix becomes $H = [K][r_1 \quad r_2 \quad T^t]$. From (3.2), we can realize that $H = [K][r_1 \quad r_2 \quad T^t] = s[h_1^t \quad h_2^t \quad h_3^t]$, and the relationship between (h_1^t and h_2^t) and (r_1 and r_2) can be written as follows:

$$K[r_1 \quad r_2] = s[h_1^t \quad h_2^t] \quad (3.9)$$

$$s^t[r_1 \quad r_2] = K^{-1}[h_1^t \quad h_2^t] \quad (3.10)$$

Based on the prior knowledge of rotation matrix properties, r_1 and r_2 are orthonormal and perpendicular to each other and they share an equal norm (Zhang, 2000). This can be used as a constrained condition illustrated in (3.11).

$$r_1^t . r_2 = 0$$
$$r_1^t . r_1 = r_2^t . r_2 \tag{3.11}$$

$$r_1 \cong K^{-1} h_1^t$$
$$r_2 \cong K^{-1} h_2^t \tag{3.12}$$

$$h_1 K^{-t} K^{-1} h_2^t = 0$$
$$h_1 K^{-t} K^{-1} h_1^t = h_2 K^{-t} K^{-1} h_2^t \tag{3.13}$$

As h_1 and h_2 are estimated beforehand, the camera matrix (K) can be recovered using (3.13), and by letting $B = K^{-t} K^{-1}$. This leads to:

$$h_1 B h_2^t = 0$$
$$h_1 B h_1^t = h_2 B h_2^t \tag{3.14}$$

$$B = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix} = \begin{bmatrix} \frac{1}{f_x^2} & -\frac{e}{f_x^2 f_y} & \frac{ec_y - f_y c_x}{e^2 f_y} \\ -\frac{e}{f_x^2 f_y} & \frac{e^2 + f_x^2}{f_x^2 f_y^2} & \frac{-c_y(e^2 + f_x^2) + f_y e c_x}{f_x^2 f_y^2} \\ \frac{ec_y - f_y c_x}{e^2 f_y} & \frac{-c_y(e^2 + f_x^2) + f_y e c_x}{f_x^2 f_y^2} & \frac{c_x^2(e^2 + f_x^2) - 2e f_y c_y c_x + c_y^2 f_x^2 + f_x^2 f_y^2}{f_x^2 f_y^2} \end{bmatrix} \tag{3.15}$$

Considering the unrepeated elements of the B matrix, B can be reduced to a b vector as follows:

$$b = [B_{11} \quad B_{12} \quad B_{22} \quad B_{13} \quad B_{23} \quad B_{33}]^t \tag{3.16}$$

Vector b can be applied in (3.14); thus, (3.14) can be converted into $Vb = 0$, where $V_{ij} = [h_{1i}h_{1j} \quad h_{1i}h_{2j} + h_{2i}h_{1j} \quad h_{2i}h_{2j} \quad h_{3i}h_{1j} + h_{1i}h_{3j} \quad h_{3j}h_{2j} + h_{2i}h_{3j} \quad h_{3i}h_{3j}]^t$, and where i and j are the column index for H matrix:

$$\begin{bmatrix} V_{12}^{t} \\ (V_{11} - V_{22})^{t} \end{bmatrix} [b] = 0 \tag{3.17}$$

Equation (3.17) can be solved by the SVD concept to create a simplified vector b. Then, the internal parameter of the camera are as follows:

$$\begin{aligned} c_y &= \frac{B_{11}B_{13} - B_{11}B_{23}}{B_{11}B_{22} - B_{12}^2} \\ s &= B_{33} - \frac{B_{12}^2 + c_y(B_{12}B_{13} - B_{11}B_{23})}{B_{11}} \\ f_x &= \sqrt{\frac{s}{B_{11}}} \\ f_y &= \sqrt{\frac{sB_{11}}{B_{11}B_{22} - B_{12}^2}} \\ e &= -\frac{B_{12}f_x^2 f_y}{s} \\ c_x &= \frac{ec_y}{f_y} - \frac{B_{13}f_x^2}{s} \end{aligned} \tag{3.18}$$

3.2.2.2 Estimating camera external parameters

To estimate the pose information during the calibration process, the relationship between the Homography matrix columns and rotation matrix columns (3.19) can be used, as the camera matrix a K is already determined. For a complete rotation matrix, the third column is reassembled as $r_3 = r_1 \times r_2$, since the rotation matrix is orthogonal.

$$\begin{aligned} r_1 &= s_1 K^{-1} h_1 \\ r_2 &= s_2 K^{-1} h_2 \\ T &= s_3 K^{-1} h_3 \end{aligned} \tag{3.19}$$

In (3.19), the scaling factors are s_1, s_2 and s_3 are equal to s, but due to the inaccuracy involved in estimating the corresponding points from an image, differences might arise

(Morvan, 2009). But, this can be compensated for by applying three different scale factors as follows:

$$s_1 = \frac{1}{\|K^{-1}h_1\|}$$
$$s_2 = \frac{1}{\|K^{-1}h_2\|} \quad (3.20)$$
$$s_3 = \frac{s_1 + s_2}{2}$$

3.2.2.3 Nonlinear refinement using pinhole camera model

After estimating both internal and external parameters for the IR and RGB cameras using a 2D checkerboard, a general cost function was applied to globally refine the calibration results. The output of this step includes the refined internal and external parameters in addition to the external baseline between both cameras. The cost function of stereo calibration is presented as follows:

$$min \sum_{n}^{N} \sum_{m}^{M} \left(\left\| p_{mn} - \left(P_{mn} \begin{bmatrix} R_n \\ T_n \end{bmatrix} K \right) \right\|_{color}^{2} + \left\| p_{mn} - \left(P_{mn} \begin{bmatrix} R_n \\ T_n \end{bmatrix} K \right) \right\|_{IR}^{2} \right) \quad (3.21)$$

where

N is the total number of images

M is the total number of points

The cost function (3.21) minimizes the pinhole camera model for the RGB and IR cameras and produces a full set of calibration parameters. The algorithms were already implemented in MATLAB (Bouguet, 2000) with the cost function.

3.2.3 Direct Linear Transform calibration method

The Direct Linear Transform (DLT) method was first proposed by (Abdel-Aziz, and Karara, 1971) with basic camera geometric parameters, focal length, principal point as well as external parameters. Then, the DLT method was modified by (Fryer, 1989) to take into account the full distortion parameters of camera lens. Due to the unknown manufacturer's properties of commercial cameras (e.g., CCD size) and the unavailability of initial parameters needed to initiate the PBA solution, the DLT method has become the most extensively used method for calibrating commercial cameras. Moreover, the DLT solution has acceptable accuracy compared to the PBA solution when working in close range applications (McGlone, 1989).

The DLT method directly solves the camera pinhole model (3.1) using the collinearity concept. According to Figure 3.1, the ray connected the focal point (N) and the image point (I) in the camera coordinate system corresponds to the ray connected to focal point (N) and the corresponding ground point (O) of image point (I). Assuming that **1)** the camera coordinate system is centered at the principal point (c_x, c_x) and the x-axis and y-axis are perpendicular to each other in the image plane and the z axis is perpendicular to image plane, **2)** the ground coordinate system can be assumed to be centered on an arbitrary point (X_G, Y_G, Z_G). Therefore, the focal point and image point coordinates will be (c_x, c_x, f) and (x, y, 0), respectively, where f is the focal length. According to the camera coordinate system, the corresponding coordinates of focal and image points will be (X_N, Y_N, Z_N) and (X, Y, Z), respectively. The relationship between both vectors can be expressed as

$$\overrightarrow{NI}_{image} = sT\overrightarrow{NO}_{ground} \tag{3.22}$$

where

$\overrightarrow{NI}_{image}$ the vector connecting the focal point to image point in image frame

$\overrightarrow{NO}_{ground}$ the vector connecting the focal point and corresponding ground point in ground frame

s scale factor

T the transformation between image coordinate frame and ground coordinate frame

Using the known coordinates of points N, I, and O, (3.22) can be expanded and formulated as:

$$\begin{bmatrix} x - c_x \\ y - c_y \\ 0 - f \end{bmatrix} = s \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix} \begin{bmatrix} X - X_N \\ Y - Y_N \\ Z - Z_N \end{bmatrix} \tag{3.23}$$

To expressing the matrix equation as a linear equation, (3.23) can be expressed as

$$\begin{aligned} x - c_x &= s\big(m_{11}(X - X_N) + m_{12}(Y - Y_N) + m_{13}(Z - Z_N)\big) \\ y - c_y &= s\big(m_{21}(X - X_N) + m_{22}(Y - Y_N) + m_{23}(Z - Z_N)\big) \\ -f &= s\big(m_{31}(X - X_N) + m_{32}(Y - Y_N) + m_{33}(Z - Z_N)\big) \end{aligned} \tag{3.24}$$

From (3.24), we can replace s as a function of the internal and external parameters of the camera as follows:

$$s = \frac{-f}{\big(m_{31}(X - X_N) + m_{32}(Y - Y_N) + m_{33}(Z - Z_N)\big)} \tag{3.25}$$

Substituting (3.25) into (3.24) leads to the following coplanarity equation:

$$x - c_x = -f\frac{\left(m_{11}(X - X_N) + m_{12}(Y - Y_N) + m_{13}(Z - Z_N)\right)}{\left(m_{31}(X - X_N) + m_{32}(Y - Y_N) + m_{33}(Z - Z_N)\right)}$$
$$y - c_y = -f\frac{\left(m_{21}(X - X_N) + m_{22}(Y - Y_N) + m_{23}(Z - Z_N)\right)}{\left(m_{31}(X - X_N) + m_{32}(Y - Y_N) + m_{33}(Z - Z_N)\right)} \quad (3.26)$$

Equation (3.26) suffers from unit inconsistency as the image units are in pixels and the ground units are in metric (e.g., mm). $S_x$and S_y are two conversion factors for x and y axis, respectively. The conversion factors were added to solve the unit homogeneity problem.

$$x - c_x = \frac{-f}{S_x}\frac{\left(m_{11}(X - X_N) + m_{12}(Y - Y_N) + m_{13}(Z - Z_N)\right)}{\left(m_{31}(X - X_N) + m_{32}(Y - Y_N) + m_{33}(Z - Z_N)\right)}$$
$$y - c_y = \frac{-f}{S_y}\frac{\left(m_{21}(X - X_N) + m_{22}(Y - Y_N) + m_{23}(Z - Z_N)\right)}{\left(m_{31}(X - X_N) + m_{32}(Y - Y_N) + m_{33}(Z - Z_N)\right)} \quad (3.27)$$

Equation (3.27) can be expressed as a function of unknowns as follows:

$$x = \frac{XL_1 + YL_2 + ZL_3 + L_4}{XL_9 + YL_{10} + ZL_{11} + 1}$$
$$y = \frac{XL_5 + YL_6 + ZL_7 + L_8}{XL_9 + YL_{10} + ZL_{11} + 1} \quad (3.28)$$

where L_i is the DLT coefficients containing both internal and external camera parameters.

Equation (3.28) can be used for camera calibration if, and only if, both ground coordinates (X, Y, Z) and corresponding image coordinates (x, y) are known in advance. Thus L_i (where i=1, 2, ….11) will represent the unknowns and they can be computed as follows:

$$L_1 = -\frac{c_x m_{31} - \frac{f}{S_x}m_{11}}{X_N m_{31} + Y_N m_{23} + Z_N m_{33}}$$

$$L_2 = -\frac{c_x m_{32} - \frac{f}{S_x} m_{12}}{X_N m_{31} + Y_N m_{23} + Z_N m_{33}}$$

$$L_3 = -\frac{c_x m_{33} - \frac{f}{S_x} m_{13}}{X_N m_{31} + Y_N m_{23} + Z_N m_{33}}$$

$$L_4 = -\frac{\left(\frac{f}{S_x} m_{11} - c_x m_{31}\right) X_N + \left(\frac{f}{S_x} m_{12} - c_x m_{32}\right) Y_N + \left(\frac{f}{S_x} m_{13} - c_x m_{33}\right) Z_N}{X_N m_{31} + Y_N m_{23} + Z_N m_{33}}$$

$$L_5 = -\frac{c_y m_{31} - \frac{f}{S_y} m_{21}}{X_N m_{31} + Y_N m_{23} + Z_N m_{33}}$$

$$L_6 = -\frac{c_y m_{32} - \frac{f}{S_y} m_{22}}{X_N m_{31} + Y_N m_{23} + Z_N m_{33}}$$

$$L_7 = -\frac{c_y m_{33} - \frac{f}{S_y} m_{23}}{X_N m_{31} + Y_N m_{23} + Z_N m_{33}}$$

$$L_8 = -\frac{\left(\frac{f}{S_y} m_{21} - c_y m_{31}\right) X_N + \left(\frac{f}{S_y} m_{22} - c_y m_{32}\right) Y_N + \left(\frac{f}{S_y} m_{23} - c_y m_{33}\right) Z_N}{X_N m_{31} + Y_N m_{23} + Z_N m_{33}}$$

$$L_9 = -\frac{m_{31}}{X_N m_{31} + Y_N m_{23} + Z_N m_{33}}$$

$$L_{10} = -\frac{m_{32}}{X_N m_{31} + Y_N m_{23} + Z_N m_{33}}$$

$$L_{11} = -\frac{m_{33}}{X_N m_{31} + Y_N m_{23} + Z_N m_{33}}$$

As commercial cameras suffer from a severe tangential and radial distortion effects, a distortion vector of five parameters represented by L_j (where j=12, 13…. 16), can be added to the DLT method to help model both tangential and radial distortion. The solution of (3.28) can be carried out using the least square algorithm. A MATLAB

code was designed to automatically calibrate stereo cameras based using the DLT method with a 3D checkerboard. After estimating both internal and external parameters, the general cost function illustrated in (3.21) is adopted to globally refine the calibration parameters.

Joint calibration methods, which can be adopted to calibrate short baseline stereo cameras, provide geometric camera parameters as well as the distortion models' coefficients. Moreover, the baseline between both cameras is computed from the recovered pose of each camera. In the following section, a new distortion model of depth sensor is presented based on the critical review of the distortion dilemma of the depth sensors showed in 2.4.2. Therefore, the RGB-D camera can now be geometrically fully calibrated.

3.3 Depth sensor distortion calibration

Disparity, as defined in 2.2, is the raw measurements the RGB-D sensor used to measure depth. The disparity measurement is originally computed from three manufacturer's constants (i.e., IR camera focal length (f), baseline between IR camera and projector (w), and the standard projector depth (Z_0)), and two measurements (projected IR pattern from IR projector (x_i^p) and reflected IR pattern received by IR camera (x_i^c)). Both measured values are affected by both the IR camera and projector distortion. Consider both measured disparity and true disparity, the relationship between true and measured disparity can be expressed as follows:

$$d_{ti} = d_i - d_e \tag{3.29}$$

where

d_{ti} is the true disparity

d_i is the measured disparity

d_e represents the error resulting from IR camera and projectors lenses distortion

As each IR sensor suffers from tangential and radial distortion, the combined effect of such distortion, which is the disparity error, can be expressed as follows:

$$d_e = \left(\delta^c_{tang} + \delta^c_{rad}\right) - \left(\delta^p_{tang} + \delta^p_{rad}\right) \tag{3.30}$$

where

δ^c_{tang} and δ^p_{tang} are the tangential distortion effect for the IR camera and projector, respectively

δ^c_{rad} and δ^p_{rad} are the radial distortion effect for the IR camera and projector, respectively.

Factors p_1, and p_2 are used to compensate tangential lens distortion based on Brown's model (Fryer, and Brown, 1986; Heikkilä, and Silvén, 1997), and other two factors k_1, and k_2 are adopted to model the radial distortion error based on extended Brown's model (Fryer, 1989). Equations (3.31) and (3.32) consider the tangential and radial distortion effects implemented in disparity computation, respectively. From the definition, the disparity is the difference between x-location of the IR camera and the IR projector image point, thus, the distortion of y-direction is not included when calculating disparity error.

$$\delta_{tang} = p_1\left((x_t^2 + y_t^2) + 2x_t\right) + p_2 x_t y_t \tag{3.31}$$

$$\delta_{rad} = x_t(k_1(x_t^2 + y_t^2) + k_2(x_t^2 + y_t^2)^2) \tag{3.32}$$

where

p_1, and $p_2$are the factors representing the tangential distortion model

k_1, and $k_2$are the factors representing the radial distortion model

x_t and y_t are the undistorted coordinates of the image point

Considering both tangential and radial distortion effects on pixel location, the relationship between true and measured pixel location can be expressed as:

$$x_m = x_t + \delta_{tang} + \delta_{rad} \tag{3.33}$$

where

x_m the measured x-coordinate

x_t the true x-coordinate

δ_{tang} tangential distortion effect along the x-axis

δ_{rad} radial distortion effect along the x-axis

After inserting (3.31) and (3.32) into (3.30), the disparity error model of RGB-D cameras can be written as follows:

$$\begin{aligned} d_e = {} & \left(p_1\left((x_t^2 + y_t^2) + 2x_t\right) + p_2 x_t y_t\right)_c - \left(p_1\left((x_t^2 + y_t^2) + 2x_t\right) + p_2 x_t y_t\right)_p \\ & + \left(x_t(k_1(x_t^2 + y_t^2) + k_2(x_t^2 + y_t^2)^2)\right)_c \\ & - \left(x_t(k_1(x_t^2 + y_t^2) + k_2(x_t^2 + y_t^2)^2)\right)_p \end{aligned} \tag{3.34}$$

where

p refers to the IR projector

c refers to the IR camera

Equation (3.34) is a complete disparity distortion model for SL RGB-D cameras. The distortion model combines a total of eight factors, four of them eliminate the radial

distortion of both the IR camera and projector, and others compensate for the tangential distortion of the IR camera and projector. To further simplify the distortion model (3.34), four factors (w_1, w_2, w_3, and w_4) are introduced to represent the full disparity distortion. w_1 and w_2 present the combined tangential distortion of IR camera and IR projector, while w_3 and w_4 present the combined radial distortion of IR camera and IR projector.

Two major assumptions are applied for further simplifying (3.34). They are as follows:

1- As the relative orientation between IR camera and IR projector is well fixed and pre-calibrated using the manufacturer's parameters ((2.6) and (2.7)), therefore, the y-axes for the IR camera and projector can be assumed to be identical.

2- Due to the absence of IR projector data, a combined radial distortion effect known as Seidal aberrations (Fryer, 1989) and the IR camera's pixel location are adopted to assign the x-distortion effect.

The above two assumptions give the constraints illustrated in (3.35).

$$\begin{aligned} y_t^c &\cong y_t^p \\ \delta_{rad}^{sensor} &= x_t^c F(x_t^c, y_t^c, w_3, w_4) \end{aligned} \tag{3.35}$$

Inserting (3.35) into (3.34), the global distortion model can be stated as follows:

$$\begin{aligned} d_e = 3w_1\left(x_c^2 - x_p^2\right) + y_c w_2\left(x_c - x_p\right) + x_c w_3\left(x_c^2 - x_p^2\right) \\ + x_c w_4\left(x_c^4 - x_p^4 + 2y_c^2\left(x_c^2 - x_p^2\right)\right) \end{aligned} \tag{3.36}$$

Equation (3.36) illustrates the depth sensor distortion as a function of new distortion parameters w_1, w_2, w_3, and w_4. Based on the concept of disparity, the relationship between image coordinates and disparity can be stated as $x_c - x_p = d_t$ and $x_c + x_p =$

$2\,x_c - d_t$. Therefore, the squared terms in (3.36) can be replaced by the following formulae:

$$\begin{aligned} x_c^2 - x_p^2 &= d_t(2\,x_c - d_t) \\ x_c^2 + x_p^2 &= d_t(2\,x_c - d_t) + 2\,(x_c - d_t)^2 \end{aligned} \tag{3.37}$$

Finally, the full distortion model for the SL RGB-D cameras can be stated as follows:

$$d_{ei} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}^t \begin{bmatrix} 3d_{ti}(2x_i - d_{ti}) \\ 2y_i d_{ti} \\ x_i\big(d_{ti}(2x_i - d_{ti})\big) \\ x_i\left(d_{ti}(2x_i - d_{ti}) + 2(x_i - d_{ti})^2\big(d_{ti}(2x_i - d_{ti})\big) + 2y_i^2 d_{ti}(2x_i - d_{ti})\right) \end{bmatrix} \tag{3.38}$$

With four parameters w_1, w_2, w_3, and w_4, the effect of the combined radial and tangential distortion can be model for the SL RGB-D cameras independently of the baseline between the IR camera and projector using (3.38).

3.4 Systemic depth error calibration

After calibrating the baseline between RGB and IR cameras, revealing the distortion models of depth sensor, and calibrating the manufacturer's constants (*a*, and *b*), the RGB and IR sensors can be geometrically calibrated. However, many other factors affect depth precision such as depth uncertainty due to a short baseline between IR sensors, correlation algorithms and rounding-off disparity, incident angle, and object distance (Khoshelham, 2011; Park et al., 2012). The baseline between IR camera and projector is a few centimeters, which is very short; thus, the depth precision is dramatically decreased in far range. Based on the covariance disparity error propagation concept, a depth error model which compensates for the remaining depth error resulting from imaging conditions and disparity related errors is proposed.

The geometric sensor calibration does not deal with the imaging conditions and properties of the imaged scene. Thus, an extended calibration depth model is proposed

to handle the non-geometric effect of RGB-D cameras. we adopted a polynomial function to calibrate the remaining depth error after correcting the bias from the IR sensors' baseline and the IR camera and projector distortion effects. The depth error model was proposed as follows:

$$d_{sys} = Ad^3 + Bd^2 + Cd + D \tag{3.39}$$

where

d_{sys} the systemic depth error remaining after applying the distortion

$A,\ B,\ C,\ D$ the polynomial coefficients to be determined from calibration

d the undistorted depth.

Equation (3.39) is proposed to calibrate the undistorted depth of the SL RGB-D camera. The resulting coefficients *A*, *B*, *C*, and *D* are basically per-pixel numbers which can form four calibration images. SL RGB-D cameras can be fully calibrated sensors when the calibration of the manufacturer's constant, depth distortion and systematic depth error all have been handled.

3.5 Summary

In this chapter, the pinhole camera model is introduced as a calibration mathematical model for RGB-D camera lenses, and, the stereo calibration concept for the baselines of separate lenses are introduced. Both Homography and DLT methods are discussed in detail as methods for optimizing the computing of internal and external camera parameters. Also, in this chapter, a novel distortion model for depth sensors is introduced. The depth distortion model considers the distortion effect of both IR camera and IR projector lenses. Moreover, a new depth error model is proposed to

compensate for errors which are irrelevant to geometric camera calibration (e.g., rounding-off disparity, disparity correlation algorithm, incident angle, and depth range).

Chapter 4: Calibration of commercial SL RGB-D

4.1 Introduction

In this chapter, the procedure used to robustly calibrate SL RGB-D cameras is proposed. The procedure is divided into two major threads. The first thread considers the calibration of the external baseline between RGB and IR cameras and the geometric parameters of both RGB and IR cameras. The geometric camera parameters include focal length, principal point, and lens distortion parameters. The second thread calibrates the depth measurement parameters, including manufacturer constants, depth distortion, and systemic depth error. The procedure is fully implemented in MATLAB, the code can calibrate any type of SL RGB-D.

4.2 RGB and IR cameras baseline calibration procedure

The baseline between RGB and IR cameras can be solved using either the DLT method or the Homography method. The difference between the two methods is automation. The Homography method is a fully automated method implemented in MATLAB (Bouguet, 2000); however, it has limited orientation parameter accuracy; this is due to the short baseline between RGB and IR cameras relative to the checkerboard captured distances (Khoshelham et al., 2013). The baseline and object distance strongly affect depth accuracy and therefore both internal and external camera parameters (Gallup et al., 2008; Kytö et al., 2011). In order to automate the DLT method, a new function is designated to extract a predefined 3D checkerboard. The 3D checkerboard combines two 2D checkerboards which are placed perpendicular to each other. First, the function detects one of two 2D checkerboards, then extracts the point features. The function replaces the detected 2D checkerboard with a blank area; therefore, the captured image

is processed again in order to detect a new checkerboard. If a new checkerboard exists, the algorithm detects the checkerboard and replaces with a blank area. The process ends when no checkerboard is found. The final step combines the extracted 2D checkerboards is applied to form the 3D checkerboard's point features. The baseline calibration methodology is presented in Figure 4.1.

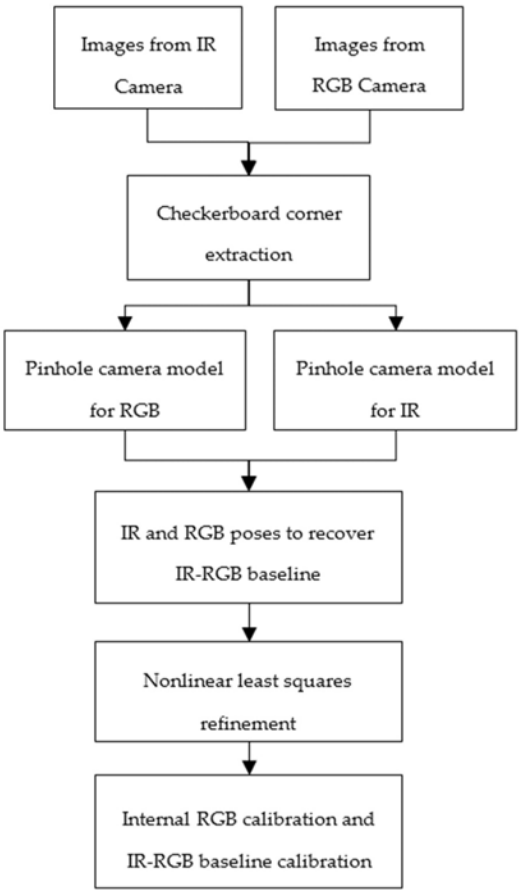


Figure 4.1: RGB-IR cameras baseline calibration methodology.

Figure 4.1 shows the general methodology for baseline calibration between stereo cameras. At this stage, the IR projector is covered or turned off, thus instead of a depth image an IR image is captured by the RGB-D camera. As indicated in Figure 4.2, a 3D checkerboard is designed to calibrate RGB-D cameras. After capturing the pair of images, the proposed function to extract the corner points is applied. By adopting the

equation series indicated in 3.2.3, a sole camera calibration is applied based on the DLT method. The method is applied separately to each camera image, which means that every camera has its own internal and external calibration parameters. Using this information, we extracted common images to be used to stereo calibrate RGB and IR cameras based on the residual image. Using the initial parameters obtained from a single camera calibration and a mean value for the external baseline between RGB and IR cameras, a nonlinear least squared method optimizes the pinhole camera model (3.2), illustrated in 3.2.1, including the camera distortion parameters. The optimization model (3.21) is illustrated in 3.2.2.3.

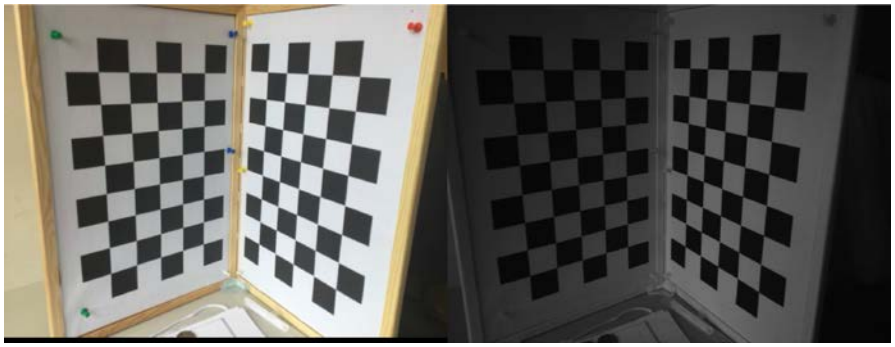


Figure 4.2: 3D designed checkerboard proposed to calibrate RGB-D cameras.

4.3 RGB-D depth calibration methodology

The depth calibration procedure handles three major components: **1)** manufacturer's constants, **2)** depth distortion, **3)** systematic depth error. Figure 4.3 shows the depth calibration methodology for RGB-D cameras. Three major parts are indicated in Figure 4.3; they are highlighted by dotted-red lines. The middle part deals with the calibration of manufacturer's constants, the lower part estimates the distortion model parameters, the left part models the remaining systematic depth error. The calibration of manufacturer's constants is proposed to compensate for the depth bias resulting

from the inaccuracy of the IR camera and projector baseline. This part of the calibration adopts the disparity and true ground depth to calibrate the manufacturer's constants (a and b), as indicated in (2.5) and described in detail in (2.6) and (2.7), respectively. Calibrating those constants is supposed to eliminate the effects of the IR camera and projector baseline, IR sensor focal length, and the mapping factors (α and β) which normalize the disparity. After calibrating the manufacturer's constants, the true depth image is converted back to the disparity domain; therefore, the ground truth disparity is generated. The second part of the calibration is calibrating the distortion of IR camera and projector lenses. This part adopts the proposed distortion model (3.38) which deals with RGB-D camera disparities. The least square method is used to obtain the distortion parameters (w_1, w_2, w_3, and w_4) of the depth sensor. Then the captured disparity is corrected regarding the IR sensors' lenses, therefore the resulting disparity from IR sensors baseline and IR lenses distortion is corrected. Finally, the depth error model is proposed based on (3.39). The ground truth depth image is used with the corrected disparity to compute the residual depth error. The polynomial model is adopted to fit the depth residual, then the model parameters (A, B, C, and D) are stored as correction images.

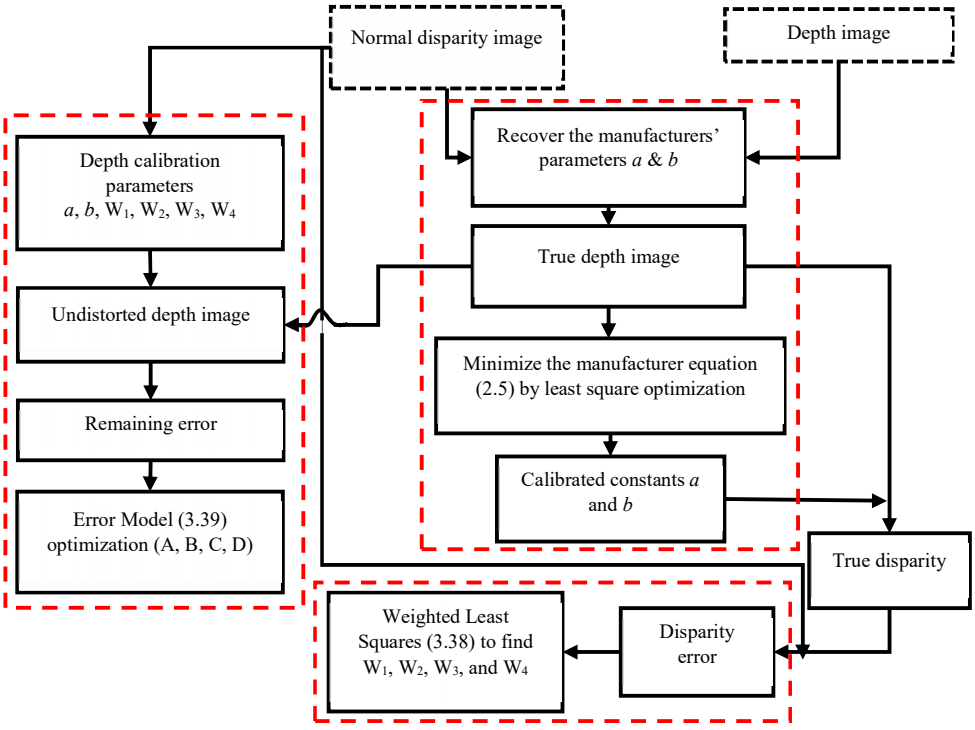


Figure 4.3: Depth calibration methodology divided into three threads

Based on the methodologies indicated in Figure 4.1 and Figure 4.3, a mobile APP is developed to extract the raw images (i.e., IR and RGB images, depth and disparity images). These images are saved to help thoroughly calibrate the SL RGB-D cameras. In the following section, the calibration results for different SL RGB-D cameras are illustrated to examine the performance of the calibration methodologies.

4.4 Experimental Design and Data Collection

4.4.1 RGB-D cameras calibration results

The required data to be captured by the sensor and the calibration results are listed in Table 4.1.

Table 4.1: Calibration data and calibration results for the SL RGB-D cameras.

Calibration thread	Required data	Output calibrated parameters
External baseline calibration	Pair of RGB and IR images for 2D or 3D checkerboard (IR projector switched off)	For both RGB and IR cameras: 1. Geometric camera parameters 2. Distortion camera parameters 3. RGB-IR baseline parameters
Depth calibration	RGB, depth, and disparity images for known surface (IR projector switched on)	1. Recalibrated manufacturer (*a*, *b*) parameters 2. Depth distortion parameters 3. Per-pixel systemic error model parameters

In this study, a two-step IOS APP is developed to aid the structure sensor in capturing the data required for thoroughly calibrating the sensor. In the first step, the APP switches the IR projector to off mode and starts capturing both RGB and IR camera images at the same time for RGB-IR camera baseline calibration. In the second step, the IR projector is switched on and a pair of RGB depth images are captured. For example, two structure sensors are calibrated using our method. The two structure sensors are attached to two different iPads to obtain RGB images. The first sensor (S.N. 26779) is attached to an iPad Air, while the second sensor (S.N. 27414) is attached to an iPad Air 2. Table 4.2 shows the data captured by both sensors.

Table 4.2: Data captured by SL RGB-D sensors

Sensor	Step 1 (RGB and IR images)	Step 2 (RGB and Depth images)
Sensor 1 (IPad Air)	53	90
Sensor 2 (IPad Air2)	59	44

For geometric camera and RGB-IR baseline calibration, the DLT method is used to solve the pinhole camera model. The pinhole model is refined using tangential and radial distortion models. The geometric camera parameters for both RGB and IR cameras as well for the RGB-IR baseline are given in Table 4.3. Camera focal lengths

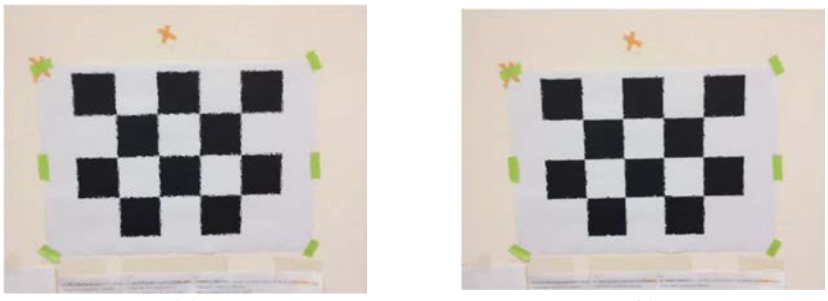
(Fx, Fy) and principal points (Cx, Cy) in pixels, three radial distortion parameters (K_1, K_2, K_3), and two tangential distortion parameters (P_1, P_2) are the calibrated geometric parameters for both IR and RGB cameras' lenses. The RGB-IR baseline is expressed as a 6DoF vector which includes translation components (dx, dy, dz) and rotational components (Rx, Ry, Rz). Translation is expressed in mm while rotation is quantified as Euler angles in rad.

Table 4.3: RGB-IR baseline calibration results (Sensor 1, and Sensor 2)

Sensor		Sensor 2 (iPad Air2)		Sensor 1 (iPad Air)	
		Value	**STD**	**Value**	**STD**
RGB camera internal parameters	**Fx**	552.690	0.490	550.060	0.620
	Fy	551.160	0.460	549.040	0.600
	Cx	315.850	0.620	321.380	0.640
	Cy	241.640	0.660	234.890	0.870
	K_1	0.160	0.007	0.149	0.010
	K_2	-0.338	0.051	-0.302	0.083
	P_1	0.002	0.001	-0.004	0.001
	P_2	0.003	0.001	0.009	0.001
	K_3	-0.005	0.110	-0.286	0.201
IR camera internal parameters	**Fx**	552.440	0.450	552.150	0.590
	Fy	550.850	0.440	550.470	0.570
	Cx	316.080	0.550	314.440	0.610
	Cy	238.930	0.690	233.640	0.850
	K_1	0.058	0.008	0.110	0.013
	K_2	-0.577	0.067	-1.090	0.120
	P_1	-0.001	0.000	-0.004	0.001
	P_2	0.004	0.000	0.005	0.000
	K_3	1.065	0.165	2.456	0.347
RGB-IR baseline	**dx**	37.502	0.121	36.399	0.180
	dy	2.877	0.113	3.357	0.192
	dz	21.539	0.464	18.812	0.770
	Rx	-0.010	0.001	0.005	0.002
	Ry	0.013	0.001	0.001	0.002
	Rz	-0.006	0.000	-0.008	0.000

The focal lengths, principal point, and distortion parameters vary significantly among both examined sensors. The manufacturer provided an SDK to adopt those cameras to

produce color point clouds from captured RGB-D frames, then the manufacturer assigns a value of 566.6 pixels for both Fx and Fy and uses the image center as the principal point while ignoring the radial and tangential distortions of cameras lenses. Comparing the calibrated results with the manufacturer's parameters, there are significant discrepancies between the calibrated parameter of both sensors and the manufacturer's default parameters. This leads to the conclusion that the sensors must be calibrated before adopting them for extremely precise applications. According to (2.8), the resulting point cloud is completely affected by the focal length and principal point. To evaluate the effects of focal lengths and principal point on the point cloud, one calibrated sensor (Sensor 1), is used to capture a scene with known control points for further quantitative assessment. The captured RGB-D images were converted to a color point cloud using the calibrated data shown in Table 4.1, and the original manufacturer's parameters. Figure 4.4 shows the point clouds of the observed scene for a square checkerboard with a width and length of 63.5mm. Sixty distances were measured from five captured frames. The frames were captured from a distance under one meter in order to overcome the effect of RGB image resolution on the accuracy of extracted checkerboard corners.



(a) (b)

Figure 4.4: Reconstructed point cloud (a) using the calibrated parameters of our method, (b) using the default parameters.

Sixty control distances were measured from the point cloud and compared to the ground truth distances. Figure 4.5 shows the error performance of both calibrated and default parameters.

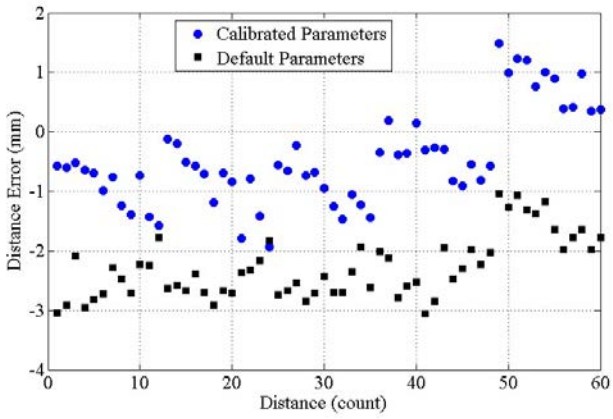


Figure 4.5: Distance residuals using both calibrated and default parameters.

Figure 4.5 clearly shows that the error bias is reduced, for when using the calibrated parameters, the distance bias is around -0.5mm; on the other hand, while using the default parameters, the bias is around -2.5mm. Between calibrated focal length and default focal length, a 16 pixel difference exists, which leads to a bias of 2mm. The effect of this bias will be severe in far range applications due to other factors effecting depth precision (e.g., depth resolution).

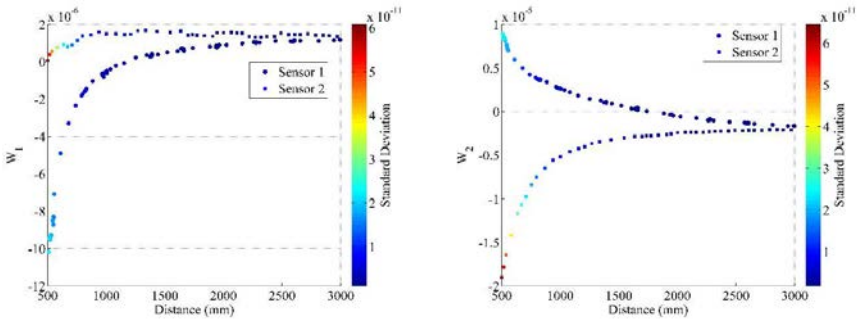
Regarding the depth calibration procedure illustrated in Figure 4.3, the true depth image used in the calibration process was produced as follows: The sensor was attached to an iPad device, four control points were placed on the iPad screen and another thirty points were placed on a designed wall. Then a total station was used to calculate the true depth image. For a range of nine meters, a structure sensor was placed at a designated station around 0.50 meters from other stations (12 stations are involved). The total station was placed on a remote station 15 meters from the wall and perpendicular to the iPad and the wall. For each station, the four control points

posted on the iPad were captured, then the true depth image was recovered using the information of the control points on the wall and the control points on the iPad. Using the methodology indicated in Figure 4.3, the manufacturer's parameters, distortion depth model, and the systematic error model have been recovered. The following table shows the manufacturer's and calibrated constants *a*, and *b*. These parameters represent the effect of inaccuracy of the baseline between the IR camera and projector.

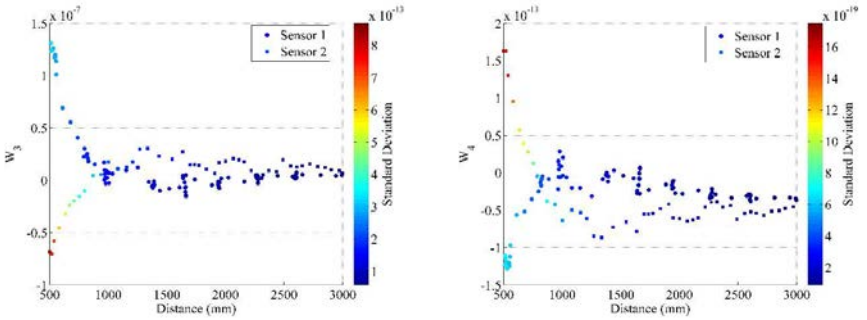
Table 4.4: Manufacturer's parameters *a*, and *b* for both sensors before and after the calibration process

Sensor	In-Factory Calibrated Value		Calibrated Value	
	a	b	a	b
1	−3.38807× 10−6	3.82665× 10−3	−3.42936 × 10−6	3.86688 × 10−3
2	−3.38649× 10−6	3.82538× 10−3	−3.34912 × 10−6	3.78253 × 10−3

After adopting the calibrated *a*, and *b* parameters, the depth distortion parameters were recovered. Figure 4.6 shows the depth distortion parameters of both sensors. From Figure 4.6, it can be clearly seen that the four depth distortion parameters tend to be the same value beyond a depth range of 2.50 m. This means a depth distortion model can be approximated for observation data of depths ranging to up to 2.5 meters. One reason for this is that beyond the 2.50-meter depth range, the dominant error affecting the depth value is not distortion but relative biases resulting from the depth uncertainty, rounded off disparity and disparity correlation (Khoshelham, 2011; Park et al., 2012).



(a) (b)

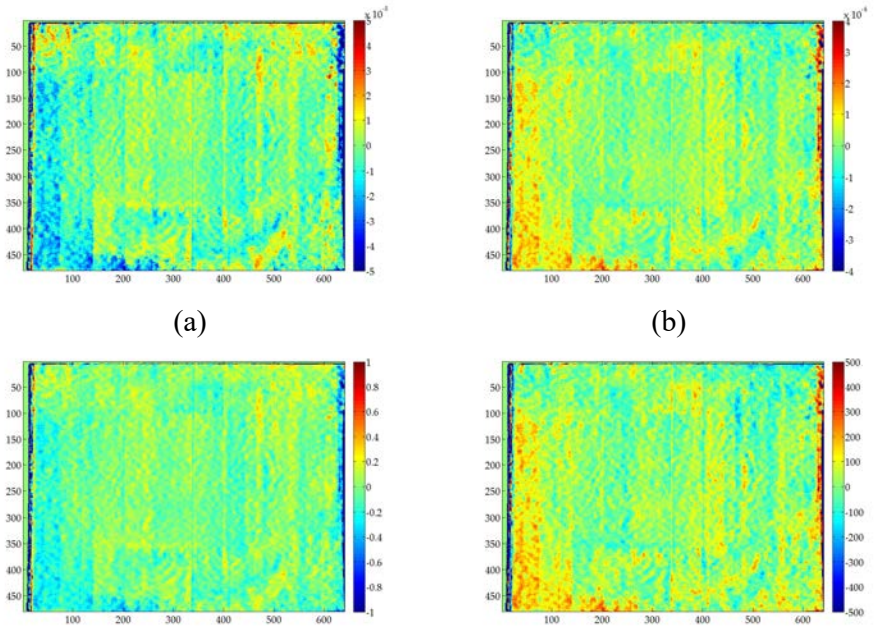


(c) (d)

Figure 4.6: Distortion parameters for both sensors. (a) is W_1; (b) is W_2; (c) is W_3; (d) is W_4.

Figure 4.6 shows that both sensors have an inverse performance regarding distortion parameters; this is mainly due to the initial bias of depth measurements. Apparently, sensor 1 has a negative bias in its measurement (see Figure 4.5); therefore, the positive value of the distortion parameters overcome this bias; in contrast, sensor 2's distortion parameters have a negative value to compensate for its positive bias.

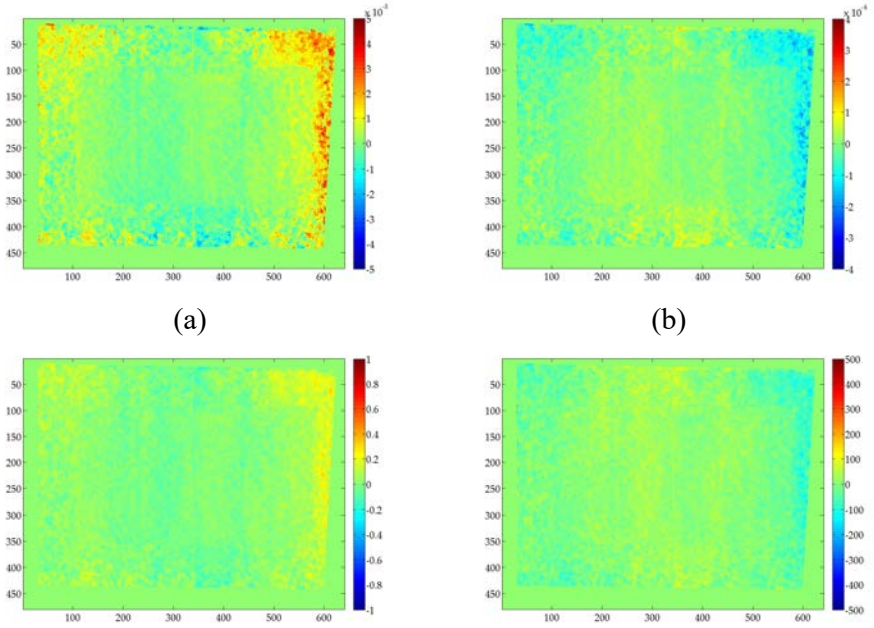
After computing the distortion parameters of the depth sensor, the systemic depth error remaining after the manufacturer's calibration and the distortion modeling are calibrated using the proposed polynomial model shown in (3.39). Figure 4.7 and Figure 4.8 show the depth error model coefficients for each pixel for sensor 1 and sensor 2, respectively.



(a) (b)

(c) (d)

Figure 4.7: The systemic depth error model coefficient for sensor 1; (a) represents A coefficient; (b) represents B coefficient; (c) represents C coefficient; (d) represents D coefficient



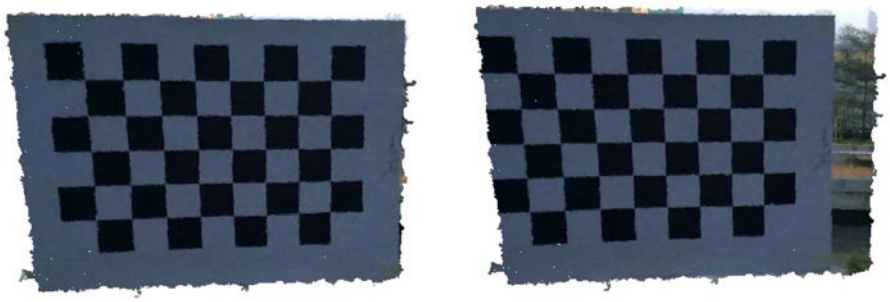
(a) (b)

(c) (d)

Figure 4.8: The systemic depth error model coefficient for sensor 2; (a) represents A coefficient; (b) represents B coefficient; (c) represents C coefficient; (d) represents D coefficient in equation

4.4.2 Calibration procedure validation

To examine the performance of RGB-IR baseline calibration, the depth and color images were collected using a structure sensor, then the calibrated parameters of the RGB-IR baseline were applied. Figure 4.9 shows the aligned point cloud before and after calibration.



(a) (b)

Figure 4.9: Calibration of the IR-RGB camera baseline effect; (a) after applying baseline calibration; (b) before applying baseline calibration.

For depth calibration verification, three different experiments were carried out to ensure the effectiveness of the calibration models. Three experiments were designated to examine depth precision, depth distortion, and RGB-IR camera baseline calibration accuracy. In the first experiment, a sensor was used to capture a distant plane from different distances. The RGB-D camera was used to capture the plane from distances ranging from 0.50 to nearly five meters. Then, for each step (0.50 meters), the calibrated and uncalibrated depth images were compared with the true depth images. Then the RMSE of the fitted plane was used to evaluate the depth performance. Figure 4.10 shows the depth performance of one of the examined RGB-D cameras.

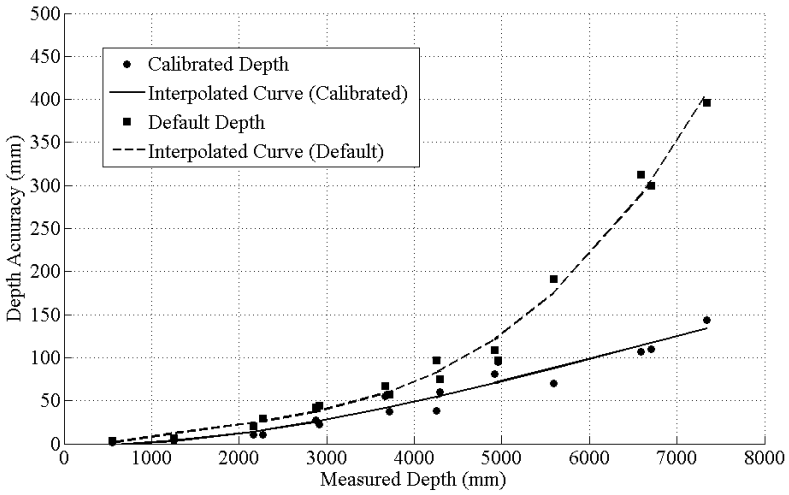


Figure 4.10: The default and calibrated depth precision performance of the examined RGBD sensors

Figure 4.10 shows that after applying our proposed calibration method, the depth precision did not exceed 0.4% of the measured depth, while for the default calibrated depth, the depth precision dramatically decreased, and exceeded 1%.

The second experiment examined a part of a room using only one RGB-D frame to compute the angle between ceiling and wall. The data was captured using one of the calibrated sensors (Sensor 1). The sensor was placed an average distance of three meters from the walls. The minimum and maximum depths were two and five meters, respectively. The planes of both wall and ceiling were extracted and the angle between them was computed. Figure 4.11 shows the difference between calibrated and uncalibrated depth. It reveals that the deformations in the wall and ceiling were corrected using calibrated data, and significant improvements in depth distortion—especially in corners—was noted. Table 4.5 shows the recovered angle using both default depth and calibrated depth. Using random sample consensus (RANSAC) with different thresholds to compute the angle between ceiling and wall, the calibrated depth can measure the angle as 89.897 ± 0.37, while it was 90.812 ± 7.17 for the default depth data.

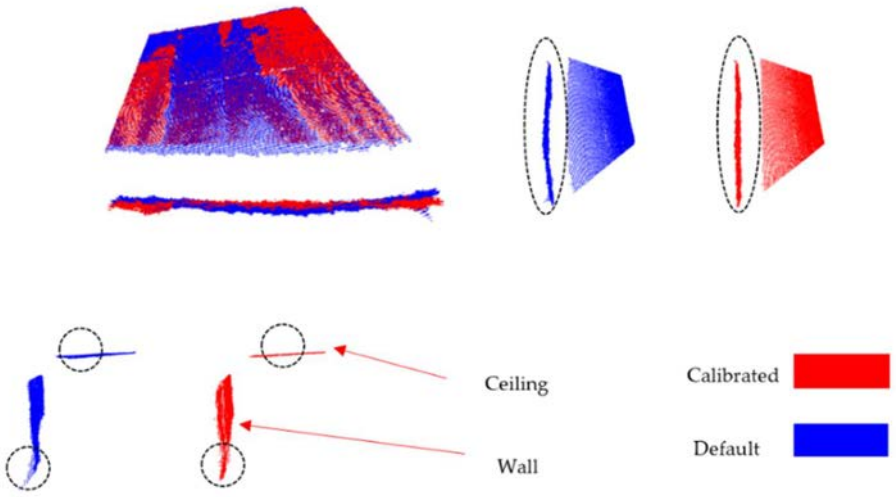


Figure 4.11: Ceiling and wall point cloud for both calibrated (red) and uncalibrated (blue) depth; the highlighted black-doted circles show the calibration impact on the point cloud.

Table 4.5: Recovered angle between two perpendicular planes using the calibrated and uncalibrated depth images.

RANSAC Threshold (m)	Recovered Angle (Degrees)	
	Default Depth	Calibrated Depth
0.001	79.8288	89.8004
0.002	99.8740	89.3294
0.005	91.5966	89.9098
0.010	92.2871	90.2850
0.020	90.4728	90.1596

The third experiment showed the full impact of the calibration procedure on 3D model quality. The thorough calibration process should produce high quality depth information as well as an accurate information between depth and RGB images. These two aims help visual RGB-D SLAM (Dryanovski et al., 2013; Hu et al., 2012; Whelan et al., 2013; Whelan et al., 2015) produce reliable cm-precision level 3D models for indoor environments (e.g., offices, corridors, rooms). Using a calibrated RGB-D camera to collect several RGB-D frames to survey an office measuring 4.5x3.5 meters. Figure 4.12 and Figure 4.13 show the reconstructed 3D model for the surveyed office with and without calibration.

Figure 4.13 (calibrated depth result), clearly shows that the edges of both chair and door as well as bookshelves all are perfectly straight lines and the projected office's wall to the floor plane is more accurate than the model produced from uncalibrated images. On the opposite side, Figure 4.12 represents the reconstructed 3D model using the uncalibrated depth. It reveals that the borders of the objects (e.g., chairs, bookshelves, and door) have a lot of noise and they cannot be easily recognized.

Figure 4.12: 3D model reconstruction of an office using the uncalibrated data of sensor 1

Figure 4.13: 3D model reconstruction of an office using the calibrated data of sensor 1

For quantitative analysis purposes, eleven distances were selected and marked. They were used to conduct the accuracy evaluation of the reconstructed models. The total station was used to measure distances for further quantifying the model's accuracy. Table 4.6 presents the comparison between the distance errors of the default and calibrated models. The results show a significant improvement in model accuracy. The overall accuracy of the 3D model reconstructed using the calibrated data is 0.8% compared to 4.0% using the uncalibrated data.

Table 4.6: Comparison between calibrated and default data for room reconstruction (meters)

Check distances	Distances measured by a total station	Distance from calibrated data	Error	Relative E (%)	Distance from default data	Error	Relative E (%)
d1	0.800	0.807	0.007	0.867	0.767	-0.033	-4.329
d2	0.802	0.792	-0.011	-1.346	0.835	0.033	3.968
d3	0.947	0.942	-0.005	-0.520	0.841	-0.106	-12.56
d4	1.110	1.113	0.003	0.298	1.076	-0.034	-3.145
d5	2.337	2.299	-0.039	-1.675	2.173	-0.165	-7.571
d6	2.560	2.564	0.004	0.159	2.441	-0.119	-4.888
d7	3.067	3.071	0.004	0.131	2.909	-0.158	-5.434
d8	3.360	3.356	-0.004	-0.131	3.291	-0.069	-2.102
d9	3.402	3.423	0.02	0.597	3.262	-0.141	-4.317
d10	3.855	3.858	0.003	0.079	3.720	-0.135	-3.638
d11	4.670	4.672	0.002	0.043	4.478	-0.192	-4.292
Mean	--	--	-0.001	-0.136	--	-0.102	-4.392
RMSE	--	--	0.015	0.771	--	0.068	3.942

The average error between the true and measured distances using calibrated and uncalibrated models were 1.5cm and 6.8cm, respectively. This improvement comes from two main reasons, the first reason is the depth precision enhancement. The second reason is the accurate registration between depth and RGB images which served to precisely assign the matched feature points to their corresponding depth. These two reasons highly affected the visual RGB-D SLAM results as assigned wrong depth to

matched features or having a depth bias may cause failure, resulting in lost tracking or severe drift of the whole reconstructed 3D model.

The experiments have shown the significance of RGB-D camera calibration effects on reconstructing the 3D model of indoor environments. The calibration procedure mainly addresses two problems: the alignment between RGB image and depth image, and the depth accuracy of the matched feature points. Miss-alignment between matched features or/and huge errors in depth information can be lead to SLAM failure or severe drift in the reconstructed 3D models. The experiments show that with calibrated data the accuracy of the resulting model is less than 1% relative error within 2cm absolute error for an indoor area (about 14 square meters). The resulting model can be efficiently used in surveying applications requiring cm-level precision 3D models.

4.5 Summary

In this chapter, the calibration procedure for RGB-D cameras is developed and implemented. The method calibrates the geometric parameters of RGB camera, IR camera, and IR projector lenses. The method also calibrates the external baseline between RGB and IR cameras in addition to the baseline between IR sensors. The method achieves a relative accuracy of 1% error for regular indoor spaces with 80% error improvement. The calibration method can enhance the RGB-D cameras' measurements' precision. RGB-D measurements obtained from calibrated cameras can be adopted in 3D reconstructions with a range of nine meters instead of a mere two meters for uncalibrated cameras.

Chapter 5: Line and plane features of RGB-D frame

5.1 Introduction

The registration between successive RGB-D frames is a critical issue for SLAM, which requires building the whole environment from successive RGB-D frames. Due to the limited field of view and working range of RGB-D cameras, applying such cameras in relatively wide spaces is a challenging task. To overcome the limited depth range, the registration of successive RGB-D frames requires the matching of their common features. There are three types of features in RGB-D frames: point, line, and plane features. Visual point features are commonly used to compute the relative transformation between two successive RGB-D frames; thus, the corresponding spatial coordinates of those feature have a severe effect on registration accuracy (Tang et al., 2016). There are many reliable and precise methods and algorithms available for extracting and describing visual point features on RGB space. In this chapter, we will concentrate on the line and plane features exiting in RGB-D frames which have not been fully investigated before. A novel method to extract, describe, and match the line and plane features in RGB-D frames is proposed. Examples are given to demonstrate that the new method can significantly improve RGB-D frame registration.

5.2 Features in RGB-D frames

Registration between two successive RGB-D frames can be handled using the color information integrated with the depth information which is well co-registered after an adequate calibration. The registration concept is extended from the Structure From Motion (SFM) (Koenderink, and Van Doorn, 1991) concept and applied to RGB-D data. The Scale Invariant Feature Transform (SIFT) (Lowe, 2004) detection algorithm

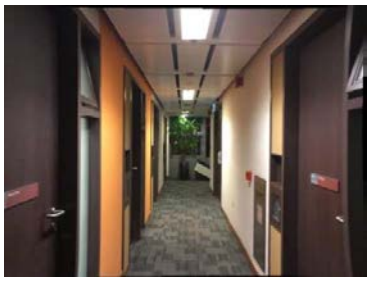
has been widely used for color images and the corresponding depths are used to recover the scale in registration (Darwish et al., 2017b). After estimating the coarse registration between two successive RGB-D frames, the Iterative Closest Point (ICP) (Besl, and McKay, 1992b) method can be applied to refine the transformation. This registration method has been widely adopted in many RGB-D SLAM systems (Henry et al., 2010; Kerl et al., 2013; Whelan et al., 2013; Whelan et al., 2015).

The main problems of RGB-D frame registration are the convergence and local minima problems of ICP; and the corresponding depth value of SIFT points. As the sensor is moving fast or dealing with processing key frames, the overlap between two successive RGB-D frames is weak, and consequently the matched SIFT points decrease. If enough matched points are exist to process the relative transformation (more than 5 points) the corresponding depths will significantly affect the accuracy of the registration.

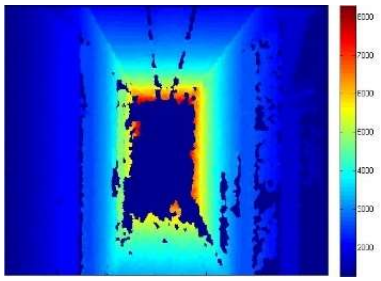
Many research efforts have been done to overcome these problems. Most of them deal with working factors such as moving the camera slowly when scanning or putting some targets on the scenes to enrich SIFT features. Other adopted line and edge to be integrated with SIFT features (Bose, and Richards, 2016) or virtual points based on planar surfaces (Ahmed et al., 2015), or using the disparity instead of depth to improve the registration state (dos Santos et al., 2016). The RGB-D frames contain several features besides SIFT features which can be implemented in the registration process. However, those features must be extracted and described for further matching before adopting them in the registration process.

RGB-D frame combines two different types of data. The first type is visual information of the scene stored in three-dimensional array representing the color intensities in three visual bands which are Red, Green, and Blue. The second type is the distance of each

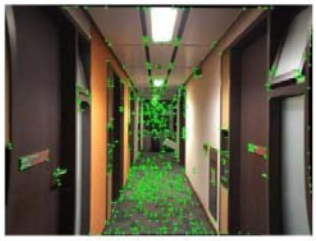
pixel from the focal point of the IR camera. Considering the geometric calibration of RGB-D cameras, each RGB-D frame can be converted to a 3D colored point cloud. Figure 5.1 shows the possible features existing in RGB-D frames.



(a) RGB image

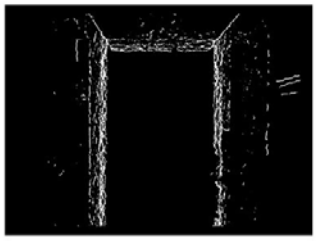


(b) Depth image

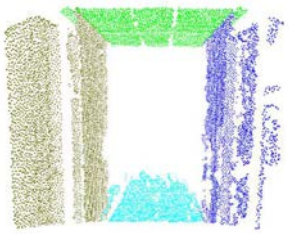


(c) SIFT features

(d) Line features from RGB



(e) Line features from depth



(f) Plane features from Depth

Figure 5.1: RGB-D frame data and features

Figure 5.1 demonstrates that all possible data can be found in one RGB-D frame. It is divided into three rows. The first row presents RGB and depth, the original data captured by RGB-D cameras. For further processing, the second row presents SIFT features' points and line features extracted from an RGB image using canny method. The third row shows the line and plane features extracted from a depth image using surface normal and RANSAC, respectively.

Depending only on visual features extracted from RGB-D may lead to unreliable registration between successive RGB-D frames. Considering the output colored point cloud instead of images, there are many additional features that can be extracted such as lines and planes. Due to the deteriorated quality of the point cloud, traditional methods to define the 3D features cannot be applied (Darwish et al., 2017a; Díez et al., 2015). In this chapter, we introduce a novel method to detect, extract, and describe 3D features like lines and planes existing in RGB-D frames. We divided the features into two main categories: the first category has 2D features in which the features—mainly visual point features—are extracted and described based on RGB images only. This kind of feature is intensively investigated before in many research starts, and the widely used methods to extract and describe these features are SIFT (Lowe, 2004), SURF (Cornelis, and Van Gool, 2008), and ORB (Rublee et al., 2011). The second category, which is our main concern, has 3D features in which the features are extracted and described based on both visual and spatial data. The following sections will discuss in detail our proposed methods to obtain line and plane features.

5.2.1 Linear features

To obtain fully functioning linear features from RGB-D frames, the detected feature should be well-defined using proper parameters, filtered using a suitable nomination criterion, then each nominated feature must be labeled using a distinctive descriptor. This procedure can be expressed in three steps which are detection, nomination, and description. Once fully described linear features are extracted, matching successive RGB-D frames is possible. Figure 5.2 shows the methodology used to detect, extract, and describe the line features.

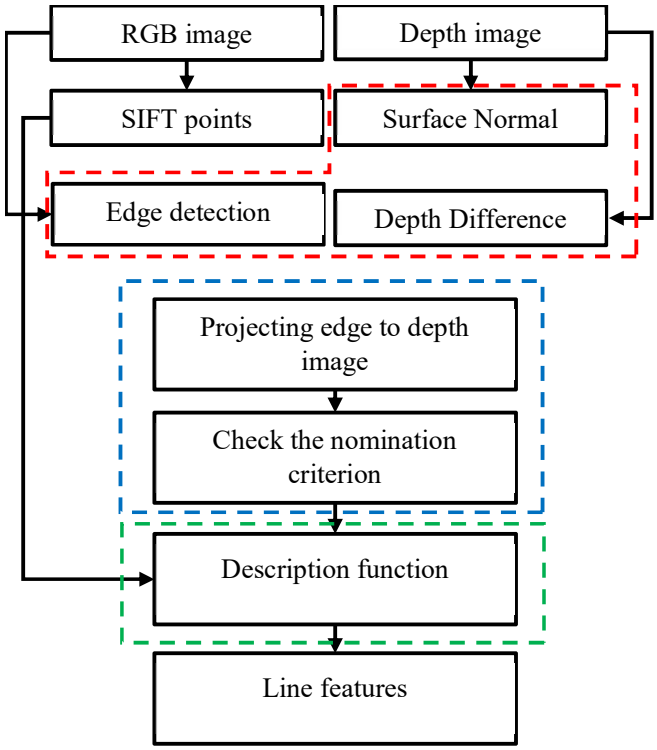


Figure 5.2: Line feature determination methodology: red-dotted line indicates detection stage, blue-dotted line indicates nomination stage, green-dotted line indicates description stage.

The detection stage is presented by a red-dotted line, in which line's properties are based on three different sources: RGB image, surface normal image, and depth image. The blue-dotted line describes the nomination stage in which all the detected line features were defined using two points. Then, based on the feature length, the detected features are filtered out. The green-dotted line presented the description function which is based on the matched visual SIFT points.

Detection step:

Line features can be extracted using RGB or depth image or both. In this method, we adopted both RGB and depth images to extract all possible line features. Firstly, RGB images were processed using the canny edge detection method to extract possible lines from the visual characteristics of the scene (Canny, 1987). The detected edges were

projected to the depth image to obtain 3D line information. on the other side, considering the depth image, the line existed in depth images generated from two cases. Firstly, when a significant change of the surface normal (e.g., line between wall and ceiling) occurred; secondly, when a noticeable depth difference is detected even within a normal surface (e.g., T.V screen border on the wall). Utilizing those two cases based on surface normal and depth difference images, a new set of line features were extracted. Those linear features can be very useful in registering RGB-D frames; however, they cannot be directly applied without matching. They may have similarity problems and some of them are just noise and too short to be considered a feature. So, the nomination stage must be carried out based on a certain criterion.

Nomination step:

The nomination stage is responsible for filtering out the weak line features. Basically, we use the length to consider the strong line features. Line features exceeding a certain length (e.g., 50cm) will be nominated as a feature.

Description step:

Due to depth noise, the description stage is crucial for the line features extracted from RGB-D frames. Many conventional ways have been developed to describe features in 3D space, especially points (Díez et al., 2015). These methods use the properties of surrounding structures to describe features. PCA is one of these methods (Ilin, and Raiko, 2010; Pacella, and Colosimo, 2013). The traditional method cannot be applied to describe the line features extracted from RGB-D frames, because of the deteriorated spatial quality of the point cloud (Darwish et al., 2017b). Instead of depending on spatial information only, we proposed a new description function combining both

visual features and line features (5.1). The description vector is based on the Euclidian distances between the matched SIFT points and the nominated line features.

$$D_{3d} = \prod_{k\in m}^{j\in n} \left\| F_j - P_k \right\| \tag{5.1}$$

where

D_{3d} the descriptor of each 3D feature, presented as line either extracted from an RGB image and projected back to the point cloud or directly extracted from a depth image based on normal

F_j the 3D feature information line uses two points

P_k the coordinate of projected matched SIFT point to 3D point cloud

m and n are the total number of matched SIFT points and extracted 3D features, respectively

5.2.2 Planar features

In addition to line features, other planar features can be extracted and described as well. In contrast to line feature extraction, RGB information cannot be simply used to detect the planar object. However, the revolution of deep learning can be helpful in the long run. In this way, we only used the depth image to detect planar surfaces, then we defined a nomination criterion to overcome the similarity problem between features. We adopted the same function used to describe the line features in the description stage.

<u>Detection step:</u>

We combined the depth image and RANSAC method to extract the possible planes in the observed scene (Fischler, and Bolles, 1981). Equation (5.2) indicates the output of the RANSAC method. Using a certain threshold to fit the planar objects, the detected planes were sorted based on the number of points.

$$I = \Omega\left(\sum_{i=1}^{m} \phi(PL_i, P) \leq thres_{pts}\right) \tag{5.2}$$

where

PL_i parameters define the i[th] plane

P point cloud generated from a depth image

$thres_{pts}$ the distance threshold defining the point outliers

ϕ function computes the orthogonal distance between points P and plane PL_i

Ω function sorts the detected planes PL based on the number of inliers

I indices of sorted planes based on the point inliers

m the total number of detected planes

Nomination step:

To keep only distinguishing planar features, we used two concepts to define the nomination criterion. The first concept is the number of points belonging to plane features and the second concept is the distance from the nearest plane. We defined the relationship between each detected plane using the Euclidean distance between every pair of points. The percentage of similarity between each plane can be carried out using (5.3). The final nominated plane features based on number of points and surrounding distance can be obtained using (5.4).

$$R_{plane} = \frac{1}{PLc_{I(i+1)}} \sum_{i=1}^{m-1} \omega\left(PL_{I(i)}, PLc_{I(i+1)}\right) \geq thres_{pls} \quad (5.3)$$

where

$PL_{(i)}$ parameters define the i[th] plane

$PLc_{(i+1)}$ point cloud defines (i+1)[th] plane

$thres_{pls}$ the distance threshold needed to filter out the identical planes

ω the function returns the orthogonal distance between two planes

m the total number of detected planes

I indices of sorted planes based on the point inliers

R_{plane} the percentage of point cloud of the (i+1)[th] plane lies inside the i[th] plane within $thres_{pls}$ distance

$$PL_{nom} = PL\left(1, R_{plane} \geq thres_{nom}\right) \quad (5.4)$$

where

R_{plane} the percentage of point cloud of the (i+1)[th] plane lies inside the i[th] plane

PL cell array contains parameters defining all detected planes

$thres_{nom}$ the threshold defining the overlap between two planes

PL_{nom} the nominated planes' parameters

Description stage:

The nominated plane features are defined by three points which can be used to reconstruct the plane information. In order to describe the plane features, (5.1) is adopted to find description vector. F_j is replaced by 3D information about plane features.

5.2.3 Feature matching

After detection, extraction, and description of plane and line features, the matching step is applied to remove the outliers. The description vector of each nominated feature is based on the Euclidean distances between the nominated feature and the position of matched SIFT points (Darwish et al., 2017b). The descriptor length depends on how many matched SIFT points exist between two successive RGB-D frames. In practice, the minimum number of matched SIFT points to construct a distinctive descriptor is ten points (Darwish et al., 2017b). As the matching between the nominated features is based on their descriptors, the features are matched based on the normalized Pearson's cross correlation concept (5.5).

$$S_{ik} = \frac{cov(D_i, D_k)}{\sigma_{di}\sigma_{dk}} \tag{5.5}$$

where

S_{ik} matching score between descriptors i and k

cov covariance between descriptors i and k

D_i and D_k descriptors of features i and k, respectively

σ_{di} and σ_{dk} descriptors of standard deviation of features i and k, respectively

Assuming 20 features were extracted from first image and 30 features from the second image, a new matrix called matching matrix is constructed to carry out the matching process. The matching matrix consists of 20 rows and 30 columns with each element presented as a value of S_{ik} between two corresponding features, then the searching among the rows to select the maximum S_{ik} value corresponding to the matched feature presented as a column index.

5.3 Effect of feature types on RGB-D frames registration

To evaluate the effect of the proposed extraction and matching methods on the RGB-D frame registration accuracy, some examples of RGB-D images are collected in different indoor spaces. Those frames are registered by using both visual 2D features as a conventional way to combine them, and by using the proposed method wherein line and plane features are added to conduct the registration between RGB-D frames. Three different data sets are collected for indoor environments and each set is combined from ten RGB-D frames. The following figures show the difference between 2D visual and 3D feature registration methods.

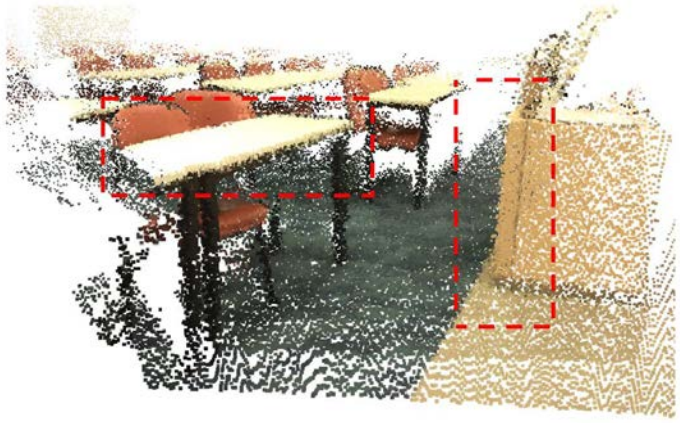


Figure 5.3: Classroom model reconstructed using the point, line, and plane features of the proposed registration method.

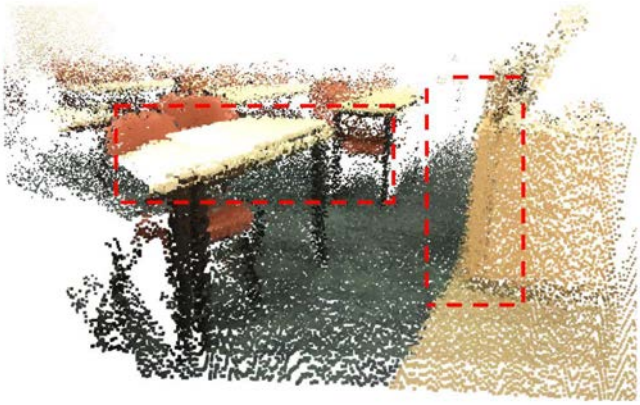


Figure 5.4: Classroom model reconstructed by the 2D visual registration method.

Figure 5.3 and Figure 5.4 show a reconstructed point cloud from ten RGB-D frames captured for a class room. The scanned environment has a lot of distinctive visual 2D features which can produce a relatively good 3D model; however, the resulting model has a mismatching problem due to the accuracy of the corresponding depth and the 3D geometry of the 2D visual features: neither is rigorous enough to produce reliable registration between RGB-D frames. It can be clearly seen that the line and plane features method is helping to reconstruct a precise model for studying desk and lectern edges (highlighted in red-dotted line).

Figure 5.5: Reconstructed model of part of large space (lift area) using point, line, and plane features registration method, (a) and (b) are different views.

(a) (b)

Figure 5.6: Reconstructed model of part of large space (lift area) using the visual 2D features registration method. (a) and (b) are different views.

Figure 5.5 and Figure 5.6 represent an example of a corridor area. The examined part of the corridor has an orange glass wall and few paintings on the opposite wall; consequently, this is considered a hard environment for RGB-D camera to produce an accurate 3D model. The difficulty of this environment stems from the scene structure as it combines mainly planar surfaces and its captured features averaged a distance of around five meters. According to our calibration method, stated in 4.4.2, the depth precision is around three centimeters. The orange wall is a glass wall and the poster on the right wall is also made of glass; since the glass scatters the IR patterns, the detected depth becomes unreliable. As we can see from the figures, the reconstructed point cloud from 2D visual points has a lot of problems such as lack of alignment between captured frames. The problem is highlighted by red-dotted line. On contrast to the visual 2D registration method, adopting the point, line, and plane features registration method can achieve a precise model for these hard environments.

Figure 5.7 and Figure 5.8 show the tested data for part of a corridor. The typical corridor has only a few distant 2D distinctive features on the ceiling and floor. Thus, using the 2D visual features registration method may lead to a severe bias. This can be

clearly seen from the presented results. The red-dotted line indicates the wall and door edge. The model reconstructed using the point, line, and plane features registration method is more precise than the one reconstructed based on the 2D visual features registration method.

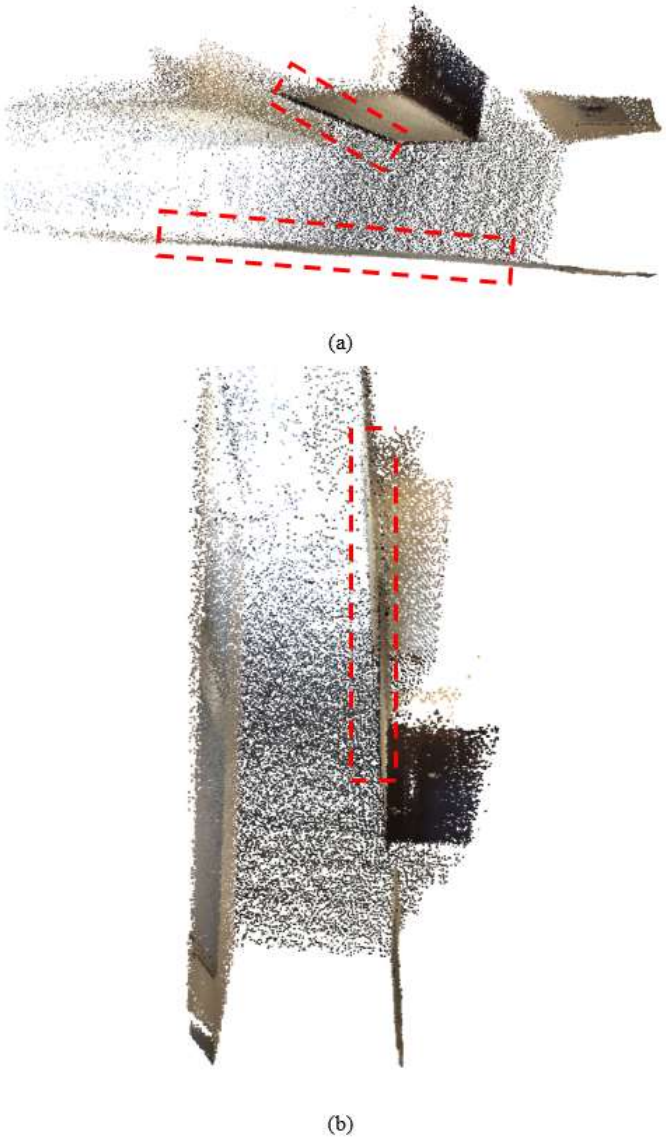


Figure 5.7: Reconstructed corridor using the point, line, and plane features registration method, (a) and (b) are different views.

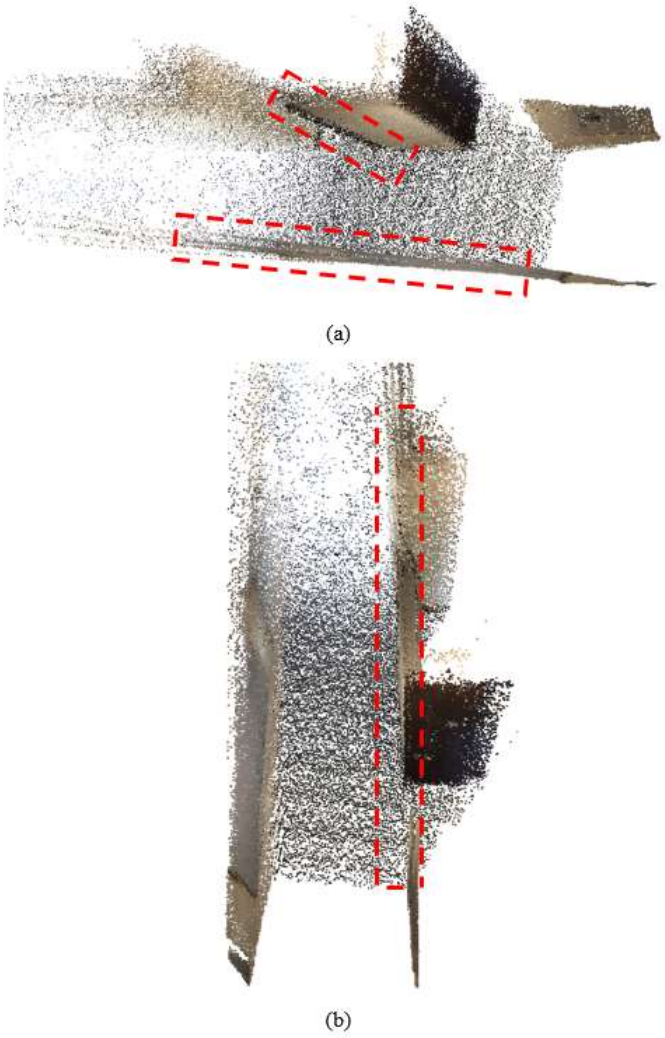


Figure 5.8: Reconstructed corridor using the 2D visual features registration method. (a) and (b) are different views.

5.4 Summary

Feature extraction and description are crucial steps for handling RGB-D frame registration problems. In contrast to the visual 2D features existing in RGB images, 3D features, presented as planes and lines, existing in RGB-D frames have not been thoroughly investigated. This study proposes a new method for extracting and describing 3D features including lines and planes. The extraction method uses both

depth and RGB images to firstly detect all 3D features exiting in RGB and depth images, then, secondly, to apply a certain criterion to the detect features to filter out the weak features. The description method uses a novel description function which combines both RGB and depth information to distinctively describe the nominated feature. The tested data demonstrated the advantage of adding 3D features to register RGB-D frames. With respect to the tested classroom, wide indoor space and corridor, the performance of the new registration method always proved better than the conventional 2D visual features registration method.

Chapter 6: Indoor reconstruction using RGB-D cameras

6.1 Introduction

The reconstruction of indoor environments using RGB-D cameras requires registering the collected RGB-D frames. The registration between successive RGB-D frames can be carried out by the RGB-D SLAM system. RGB-D SLAM is one kind of simultaneous localization and mapping algorithm which mainly deals with the data produced from RGB-D cameras. The earliest method used to reconstruct the 3D model of indoor spaces from RGB-D frames is the Kinect Fusion system (Newcombe et al., 2011), which uses the depth images of RGB-D cameras and adopts the Iterative Closet Point (ICP) algorithm (Besl, and McKay, 1992a; Rusinkiewicz, and Levoy, 2001) to register two successive RGB-D frames. Also, the system uses RGB images to color the final 3D reconstructed model. Recently, the fusion of both depth and RGB image information is being integrated with existing RGB-D SLAM algorithms.

The RGB-D SLAM function depends on different applications. The first type of application involves enabling robots to avoid obstacles in indoor environments. For this type of application, SLAM must be achieved in real time with accurate camera pose estimates (Endres et al., 2014; Huang et al., 2017). The second type of application involves reconstructing the 3D models of certain indoor environments. This application mainly relates to surveying application with high precision while post-the processing option is acceptable (dos Santos et al., 2016; Tang et al., 2016; Tsai et al., 2015).

The general framework of current RGB-D SLAM systems can be divided into three major threads. Firstly, the system adopts a proper calibration procedure to eliminate

the systemic visual and depth error from the captured RGB and depth images respectively. Secondly, RGB-D frames are registered using visual point features extracted from RGB images. Extracting matched visual point features with their depth information from two successive RGB-D frames converts the tracking problem into a rigid transformation problem (Bay et al., 2008; Cornelis, and Van Gool, 2008; Lowe, 2004). Thirdly, the global optimization step is introduced to mitigate the loop closure error. The loop closure can be detected based on either visual or geometric features (Zhang et al., 2015). These three threads form the framework of current RGB-D SLAM systems.

As the current RGB-D SLAM system uses only point features to compute the camera pose, so the reliability and accuracy of the camera pose is highly affected by depth error and the geometric distribution of the point features. In case of brittle RGB-D frames or distant feature points, the RGB-D SLAM system can easily drift or lose its location (lost tracking). Lost tracking and drift can be overcome or minimized while using more and stronger features exiting in RGB-D frames. As highlighted in the previous chapter, many features can be extracted from RGB-D frames such as lines and planes. The new system overcomes the problems of current SLAM systems, drift and lost tracking, by utilizing both 2D and 3D features to register successive RGB-D frames. In case of less textured RGB-D frames, the SLAM can continuously keep tracking using the 3D features; thus, lost tracking is overcome. Instead of 2D features, both 2D and 3D features are used to compute the camera pose and the drift problem is accordingly minimized (Darwish et al., 2017a). For further refinement of the 3D reconstructed model, the proposed SLAM system automatically extracts the environment's structural constraints for further applying them in the global refinement stage. Then, in case of a loop closure constraint, the system applies the loop closure

correction based on a graph-based optimization (Kümmerle et al., 2011). The full description of the system is indicated in the following section.

6.2 Constrained RGB-D SLAM

In this study, we propose a Fully Constrained (FC) RGB-D SLAM system which considers all possible exiting features in both RGB and depth images to precisely reconstruct a 3D model of an indoor environment. Figure 6.1 shows the major functions of the FC RGB-D SLAM system; these functions have five major threads. The first thread is calibration, which is mandatory for improving depth precision and for eliminating lens distortion. The disparity-based calibration model (Darwish et al., 2017c) is adopted to compensate for the systematic depth error and lens distortion. This method is described in 3.4 and 3.3, respectively. The second thread mainly focuses on extracting features from both 2D space (RGB) and 3D space (point cloud). The third thread deals with the description of the nominated features for further matching. Before the global optimization thread (i.e., in case of loop closure correction), the fourth thread is added to deal with the tracking algorithm which keeps the RGB-D camera in the same frame-work. The following subsections describe in detail the aforesaid threads. The major contributions of this chapter consist of outlining both a strategy of applying both 3D and 2D features in the tracking core and in describing the global constraints optimization stage of the SLAM system.

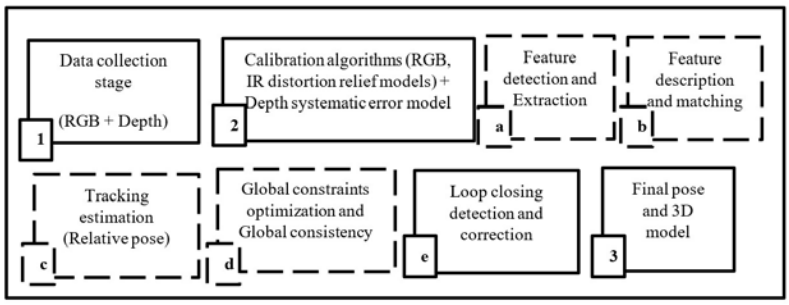


Figure 6.1: FC RGBD SLAM method threads.

In Figure 6.1, the proposed RGB-D SLAM is divided into several blocks. In the first block (1), we develop a mobile APP to capture both processed and raw data for further offline processing. In the second block (2), we adopt the method mentioned in 4.2, and 4.3 to calibrate both RGB and depth images. In the third and fourth blocks (a, b), 2D and 3D features are extracted and described then matched based on the procedure indicated in 5.2. In the fifth block (c), the general cost function which minimizes the geometric distances between matched features is applied to estimate the relative camera transformation. In the sixth block (d), the global consistency of perpendicular and parallel planes is generated and optimized. The seventh block (e) presents loop closure detection and correction. We adopt the method presented in (Kümmerle et al., 2011). The final block (3) presents the final camera pose and 3D model outputs.

6.2.1 Feature detection and extraction

FC RGB-D SLAM depends on extracted features from both RGB and depth images. The extracted features are divided into two categories. The first category is presented as 2D features which can be obtained from a SIFT algorithm (Cornelis, and Van Gool, 2008). The second category is 3D features and contains lines, edges and planes extracted from both RGB and depth images. The method described in 5.2 is used to extract 3D features existing in RGB-D frames. The color gradient based on the hessian

matrix (Bay, 2006) is also applied to detect 2D features (Bay et al., 2008; Cornelis, and Van Gool, 2008).

6.2.2 Feature description and matching

After detecting and extracting the strongest 2D and 3D features, the describing process for each feature is applied for further matching as well as for removing outliers. For 2D features, a SIFT descriptor based on the color gradient is used with a correction of scale and orientation of each nominated feature point (i.e., the SIFT descriptor length is 64 bit). The description vector of each selected 3D feature is based on the Euclidean distances between the 3D feature and the position of matched SIFT points (Darwish et al., 2017b). Equation (6.1) shows the descriptors of 2D features. As the matching between the nominated features is based on the descriptors, 2D feature matching is based on the Sum of Squared Differences (SSD) between each pair of descriptors, and the best matching is based on the minimal of SSD (Cornelis, and Van Gool, 2008). Equation (6.2) presents the matching concept between 2D features. 3D features are matched based on the normalized Pearson's cross correlation concept.

$$D_{2d} = \coprod_{block=1}^{block=16} \left(\sum_{i=1}^{i=4} dx_i , \sum_{i=1}^{i=4} |dx_i| , \sum_{i=1}^{i=4} dy_i , \sum_{i=1}^{i=4} |dy_i| \right) \tag{6.1}$$

where

D_{2d} the descriptor of the 2D feature image point

dx_i the image gradient of sub block (i) along x direction

dy_i the image gradient of sub block (i) along y direction

Normally, the SIFT descriptor uses a 4x4 pixel sub block size with a global block of 4x4 of the sub blocks. This means that the descriptor has a vector length of 64.

$$SSD_{f_1 f_2} = \sum_{i=1}^{i=64} \left(D_{f_1}(i) - D_{f_1}(i) \right) \tag{6.2}$$

where

f_1 and f_2 point features existing in the first and second image respectively

$SSD_{f_1 f_2}$ the sum of squared difference distances between point features

D_{f_1} the descriptor of point features located on the first image

D_{f_2} the descriptor of point features located on the second image

6.2.3 Tracking core

Computing the relative movement between two captured RGB-D frames is crucial step for continuous tracking of RGB-D cameras. The visual RGB-D SLAM system minimizes the geometric distance of corresponding SIFT matched points between RGB-D frames to compute the camera pose (Tang et al., 2016). The proposed FC RGB-D SLAM system uses all the geometric information to compute the relative pose between RGB-D frames. Three types of information are extracted from two successive RGB-D frames including point features from RGB images and matched 3D points from the point cloud, line features extracted from both RGB and depth information, and planes extracted from the point cloud.

Annotations: For each extracted feature type, we adopted a different representation. Thus, different relationships are introduced to compute the relative movement between two RGB-D frames depending on the feature types. This assumes that each frame has

(m) matched point features and (n) matched line features and (q) matched plane features. For 3D point features with correspondences between two successive frames, we present P_1 and P_2 as two matrices with the dimension of mx3, and each matrix contains points information, and each row has $[X_i\ Y_i\ Z_i]$. For the extracted lines, we present L_1 and L_2 as two matrices with the dimension of nx6, and each matrix contains the line information as $[XC_{li}\ YC_{li}\ ZC_{li}\ XD_{li}\ YD_{li}\ ZD_{li}]$ where the first three elements refer to the coordinates of the center point of line (we choose the nearest matched SIFT point to the line and compute its projected coordinate to the line), and the next three elements refer to the direction vector of the extracted line. For the matched plane features, we present PL_1 and PL_2 as two matrices with the dimension of qx6, each row represents the plane information as $[XC_{ni}\ YC_{ni}\ ZC_{ni}\ NX_{ni}\ NY_{ni}\ NZ_{ni}]$ where the first three elements refer to the center point of the plane, we used the same concept of line to detect such point, the next three elements refer to the normal vector of the plane feature.

R and T are the rotation and translation of the rigid relative transformation between two RGB-D frames, respectively. The Möller and Hughes (1999) method is adopted to compute the relative rotation between two corresponding vectors. Three geometric quantities should be minimized during the pose estimation process: first, E_p is the back-projection error of the matched 3D point features between the RGB-D frames (6.3); secondly, E_l is the residuals vector between matched line features (6.4); and finally, E_n is the vector of residuals between matched plane features (6.5).

$$E_p = \sum_{i=1}^{i=m} \|RP_1 + T - P_2\|^2 \tag{6.3}$$

where

m total number of matched point features

P_1 and P_2 mx3 matrix contain coordinates of all matched points for RGB-D frame 1 and 2, respectively.

$$E_l = \sum_{j=1}^{j=n} \left\| \left(Rf(D_1) - f(D_2) \right) + (RC_{l1} + T - C_{l2}) \right\|^2 \quad (6.4)$$

where

n total number of matched line features

f function converts the direction vector to a normal vector

C_{l1} and C_{l2} matched line's center point of RGB-D frames 1 and 2, respectively.

D_1 and D_2 direction vectors of matched lines between RGB-D frames 1 and 2, respectively.

$$E_n = \sum_{k=1}^{k=q} \| (RN_1 - N_2) + (RC_{n1} + T - C_{n2}) \|^2 \quad (6.5)$$

Where

q total number of matched plane features

C_{n1} and C_{n2} matched plane's center point coordinates for RGB-D frames 1 and 2, respectively

N_1 and N_2 normal vectors of matched planes between RGB-D frames 1 and 2, respectively

The global motion estimation of the RGB-D camera can be represented as in (6.6), as from the geometry principals, line and plane features introduce more constraints on

rotation rather than the translation opposite to the point features. Here, we apply a weighting parameter α, which gives more weight to lines and planes than to points. The general formula can be written as

$$\{\hat{R}, \hat{T}\} = arg\ min\left(E_p + \alpha(E_l + E_n)\right) \tag{6.6}$$

where

$\hat{R}$ and $\hat{T}$ estimated camera rotation and translation, respectively

E_p, E_l, and E_n point, line, and plane features reprojection error, respectively.

α weighting factor

First, the system initializes the pose between two RGB-D frames by adopting the concept of visual RGB-D SLAM. The point features' correspondences are used to calculate the relative pose, then SLAM used this information to match all extracted planes and lines extracted from both depth and RGB images. The final pose is calculated through (6.6). After optimizing the pose information between successive RGB-D frames, the global constraints stage is introduced for further smoothing the reconstructed 3D models.

6.2.4 Global constraint

The global constraint stage is the final refinement process before the loop closure concept can be applied. In this stage, the spatial relations (i.e., perpendiculars, parallels) are basically generated from the camera pose to roughly determine the 3D shape constraints. Thus, the global model (gM) is divided into separate sub models (sM) for enhancing the accuracy of alignments. The global constraints stage is based on the planar objects between the successive sub models. The proposed method is

based on the camera pose information. Figure 6.2 shows the coordinate system of the structure sensor camera.

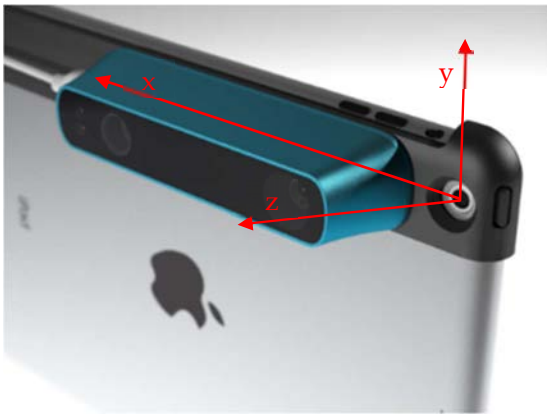


Figure 6.2: Structure sensor coordinate system.

Regarding the sensor coordinate system, the rotation around y direction (θ_y) indicates planar movement constraints in a 2D floor plan. In an indoor environment, the dominant axis which controls the scanning process is the y axis. Equation (6.7) shows the formula used to detect the turned frames from the y-axis rotation. The turned frames were detected because they are the most likely to contain constrained structures. Based on these turned frames, the global model is divided to several sub models using (6.8) for further refinements.

$$N = \left[1, PE\left(G\left(\frac{\partial \theta_y}{\partial y}\right)\right), end\right] \tag{6.7}$$

where

N IDs of turned RGB-D frames

PE the function that detects the peaks in a time series

G Gaussian filter that functions to smooth the gradient of y axis rotation

θ_y rotation angle around y axis

$$sM = \coprod_{i=1}^{length(N)-1} gM\big(N(i):N(i+1)\big) \tag{6.8}$$

where

N IDs of turned RGB-D frames

gM global model to be smoothed

sM sub models divided by turned frames' indices N

Once all sub models are constructed, the global refinement stage is carried using the pre-known spatial relations between each sub model. The spatial relations are stored in (S), which contains the turn angles around the three axes. S is reconstructed using the planar relation between two successive sub models, i.e. perpendiculars, parallels, and artificially defined angles (e.g., $\pi/4$, $\pi/2$, $3\pi/4$). The global refinement stage deals with forcing back these artificial angles. Equation (6.9) presents the formula used in the global refinement stage.

$$\{rpose, rM\} = Con(sM, S) \tag{6.9}$$

where

$rpose$ the refined camera poses after the refinement stage

rM the refined global model after the refinement stage

S spatial constrained information

Con constrained function reinforces the predefined spatial information S

After estimating and refining the relative pose between successive RGB-D frames, the detection and correction of loop closure, if any, is performed.

6.2.5 Loop closure

Loop closure is a basic concept for correcting a closed mapped space. The common method is a graph optimization technique based on nonlinear least square optimization (Kümmerle et al., 2011). This method constructs a graph problem based on nodes and edges. Each node represents the pose information of each RGB-D frame, while the edges represent the 6DoF relative baseline between two successive RGB-D frames. The method is adopted to most existing RGB-D SLAM algorithms, and produces stable results. Thus, we also adopt this approach in our proposed FC RGB-D SLAM.

6.3 Three-dimensional model reconstruction

Two experiments are conducted to evaluate the FC RGB-D SLAM performance. The first experiment consists of scanning an open (no loop closure) corridor. The second experiment consists of surveying a closed (loop closure existed) corridor. Additional experiments for constructing small and big rooms are illustrated. In the experiments, the performance of FC RGB-D SLAM is evaluated by comparing its results to those of other existing RGB-D SLAM systems (i.e., visual RGB-D SLAM, SensorFusion).

6.3.1 Scanning of an open environment

In this experiment, we use the proposed SLAM to precisely reconstruct 3D models of indoor corridors which has long length with less distinctive 2D and 3D features. The data are captured by structure sensor. The sensor is attached to an iPad Air 2 to capture and process the data. The Structure sensor has its own processing framework (SDK)

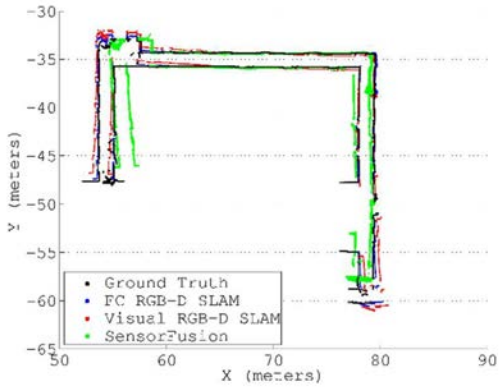
for processing the captured depth images, color images, and the IMU data from the iPad to produce the 3D model of the captured environment (we note this SDK as SensorFusion system). Then the data are processed by both the proposed FC RGB-D SLAM system and by visual RGB-D SLAM for post-processing. The data for a narrow corridor measuring 58m in length, 2.5m in height and 1.5m in width are captured. Two kinds of scanning methods associated with iPad position are used: vertical scanning and horizontal scanning methods (see Figure 6.3). A laser scanner was used to capture the ground truth of the corridor for further quantitative evaluation. Four different experiments are conducted for the examined corridor.

(a) (b)

Figure 6.3: Scanning methods for structure sensor, vertical (a) and horizontal (b)

(a) (b)

(c) (d)



(e)

Figure 6.4: The scanned corridor for the vertical scanning method, (a) model from SensorFusion; (b) model from FC RGB-D SLAM; (c) model from visual RGB-D SLAM; (d) laser scanner model (ground truth); (e) projected wall to the ground of four models (black is ground truth, blue is FC RGB-D SLAM, red is Visual RGB-D SLAM, and green is SensorFusion)

The accuracy assessment is based on model quality. Each model is compared to the ground truth; thus, the error of each point is quantified and a summary histogram of each model is used to check for model accuracy. The error of each mapped point is quantified based on the Euclidean distance difference between the mapped point and the corresponding ground truth point obtained from the laser scanner. Figure 6.5 shows the quantitative results among the three different RGB-D SLAM systems. It can be clearly seen that the model of FC RGB-D SLAM can enhance overall model accuracy and its alignment. The model is divided into patches A, B, and C. These three patched are completely perpendicular. Corner angles are perfect right angles in the reconstructing model using FC RGB-D SLAM; however, they are not right angles in either visual RGB-D SLAM or SensorFusion methods. Although visual SLAM can

achieve better accuracy compared to SensorFusion system, part A and C show severe drift in both SensorFusion and visual RGB-D SLAM results, while the drift is significantly reduced by FC RGB-D SLAM. The largest error existed in the corner between A and B because at that spot the corridor has a glass window. Thus, the depth data produced by the RGB-D camera has a lot of noise. Furthermore, the error existing at the end of part C is due to the structure of the E area (highlighted by black-dot line). This area lacks 2D and 3D feature, as it is an open space.

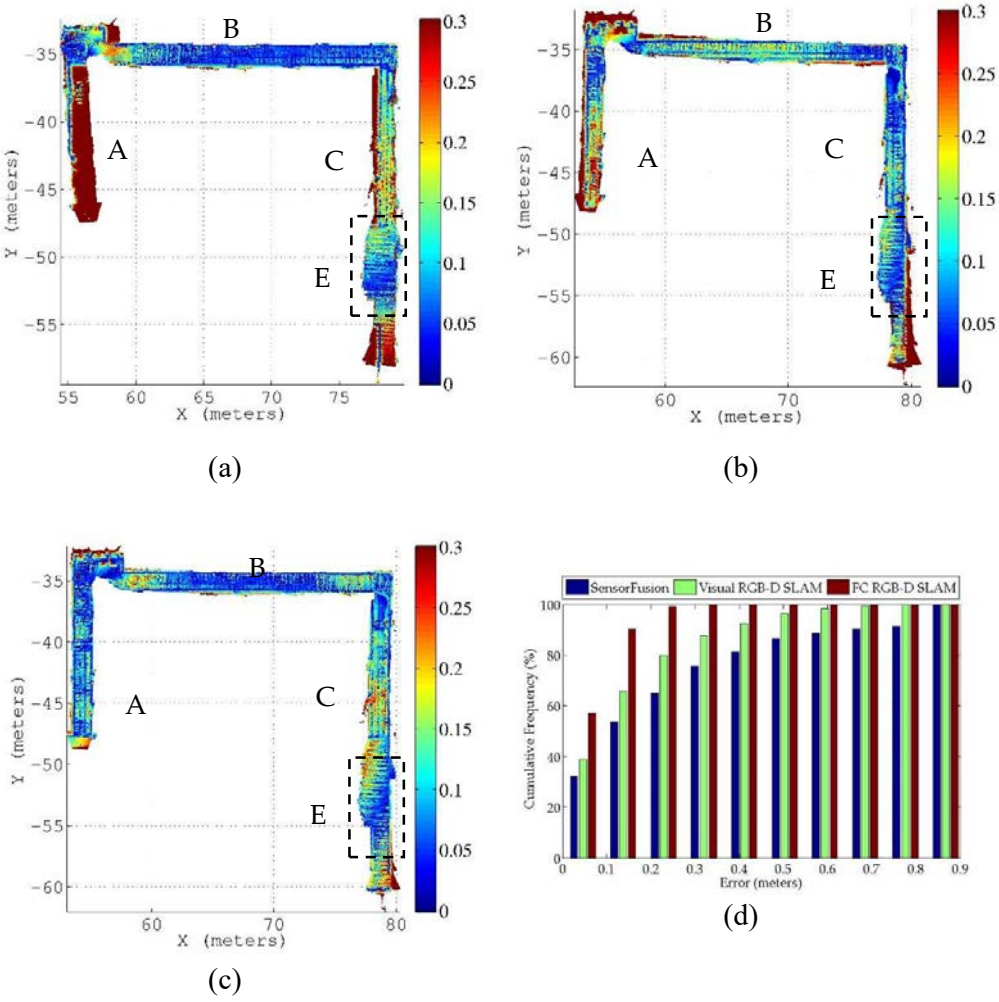


Figure 6.5: The spatial error distributions for corridor scanned in vertical scanning mode, (a) model from SesnorFusion; (b) model from Visual RGB-D SLAM; (c) model from FC RGB-D SLAM; (d) error histogram of the three different systems.

To validate the FC RGB-D SLAM method, three more data sets are captured using the same sensor with different data collection procedures: one set used a horizontal position, another set used a horizontal iPad position with a low frame rate (5 fps), and the third set used a horizontal iPad position facing the ground (uncaptured ceiling). To summarize the results of the four experiments, the average cumulative error histogram is presented in Figure 6.6. It can be clearly seen that the error rate for 95% of points does not exceeded 0.20m for the FC RGB-D SLAM, which is better than the 1.00m and 1.20m error rates for the visual RGB-D SLAM and SensorFusion methods, respectively.

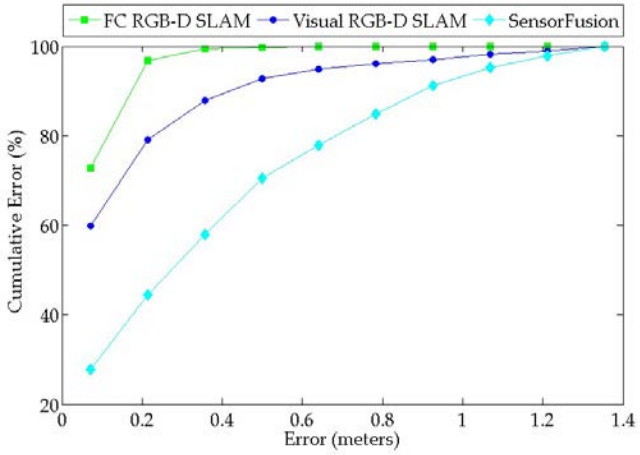


Figure 6.6: The average cumulative error histogram for all captured experiments.

6.3.2 Scanning of a closed environment

The purpose of this experiment is to evaluate the FC RGB-D SLAM performance in closed spaces, which means the camera will be forced to revisit the first captured scene. A corridor measuring 80m in length, 2.5m in height and 1.5m in width is captured and processed using SensorFusion, visual RGB-D SLAM, and FC RGB-D SLAM. For accuracy evaluation, the mapped corridor ground truth (used as reference) is captured using laser scanner. Figure 6.7 shows the comparison between the results of the three

systems. In Figure 6.7, it can be clearly seen that the performances of FC RGB-D SLAM and visual RGB-D SLAM are quite similar. This is because the D area (highlighted by a black dotted line in Figure 6.7) of the corridor has a glass wall; thus, depth from structure sensors are almost missing, and only the ceiling and floor of the corridor can be mapped. Furthermore, the loop closure correction, which optimizes the global camera's locations based on the closure error, is applied to both cases.

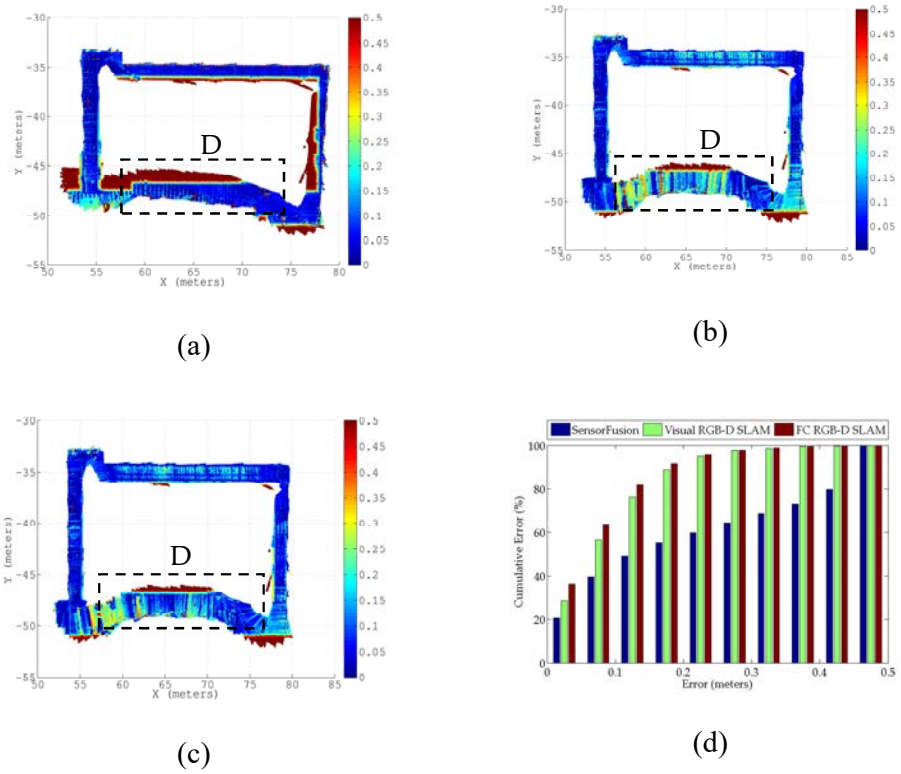


Figure 6.7: The spatial error distributions for a scanned corridor using the vertical scanning mode, (a) model from SesnorFusion; (b) model from visual RGB-D SLAM; (c) model from FC RGB-D SLAM; (d) error histogram of the three different systems.

Other experiments involving closed space with different environmental conditions are tested for a printing room measuring 3.5mx2.5m and a classroom measuring 11.4mx6.7m. The FC RGB-D SLAM system's results are compared to the results of visual RGB-D SLAM. Qualitative and quantitative assessments of the resulting

models are investigated using ground truth measurements (Laser scanner) for the examined environments.

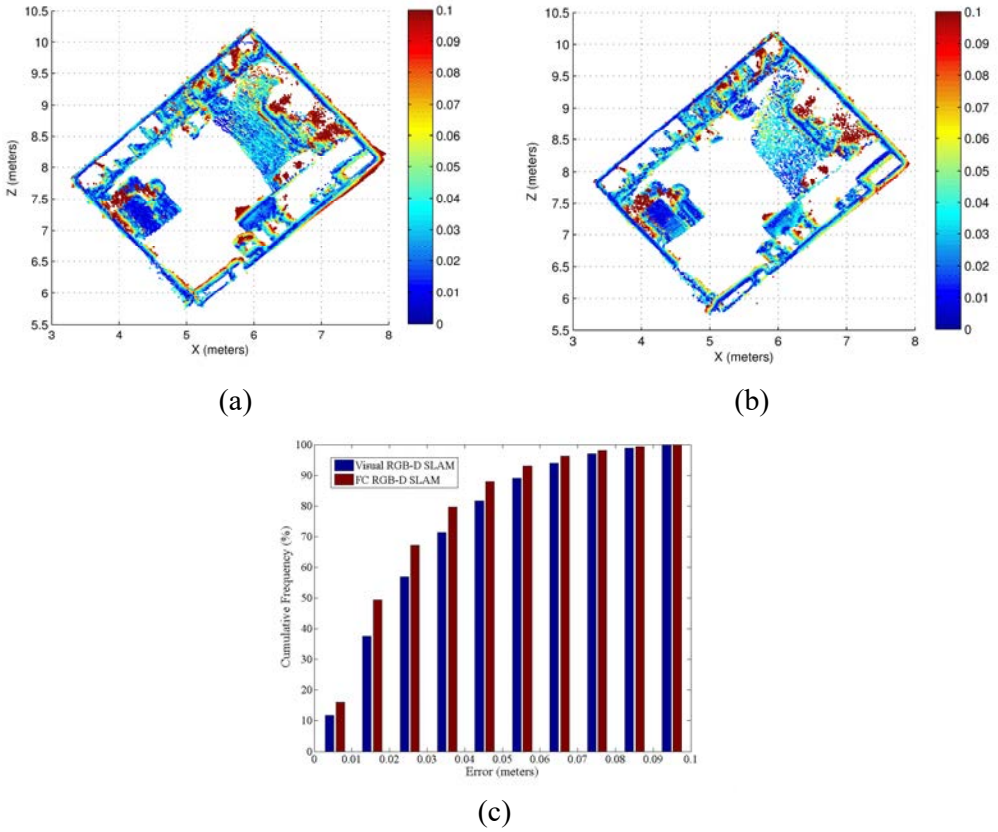


Figure 6.8: Error of printing room reconstructed model; (a) using visual RGB-D SLAM; (b) using proposed FC RGB-D SLAM; (c) the error histogram of both methods.

Figure 6.8 shows the point cloud errors of the printer room projected in the horizontal plan from the visual RGB-D SLAM method and the FC RGB-D SLAM method, and compares them to the results from the terrestrial Laser Scanner. By comparing Figure 6.8 (a) with Figure 6.8 (b), it can be clearly seen that large errors (brown colors) are significantly reduced at the corners of the room and around objects in the room. This is because in those places there are many line and plane features and FC RGB-D SLAM can utilize those features to improve 3D model accuracy. Figure 6.8 (c) compares the point cloud error distributions with the visual RGB-D SLAM and the FC

RGB-D SLAM methods. It is clearly demonstrated that with the FC RGB-D SLAM method the 3D model is more accurate than with the visual RGB-D SLAM method.

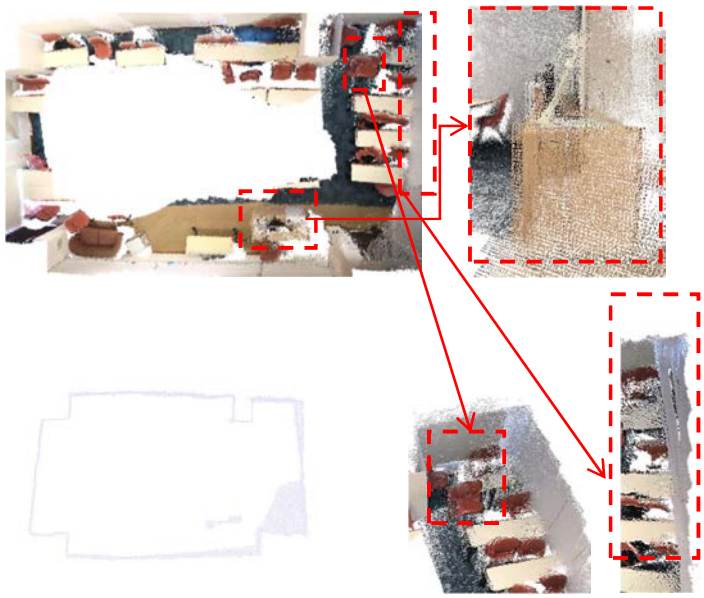


Figure 6.9: Classroom model constructed by visual RGB-D SLAM.

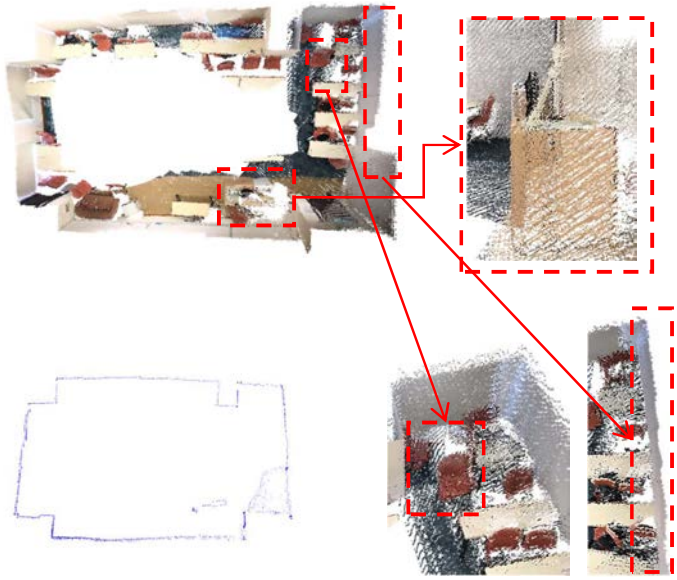


Figure 6.10: Classroom model constructed by FC RGB-D SLAM.

Table 6.1: Error difference between the FC RGB-D SLAM and visual RGB-D SLAM methods (meters).

Distance	Laser	Proposed	SLAM	Absolute Relative Error (FC RGB-D SLAM%)	Absolute Relative Error (Visual RGB-D SLAM %)
D1	0.360	0.352	0.373	2.222	3.611
D2	0.450	0.444	0.522	1.333	16.000
D3	0.550	0.548	0.502	0.364	8.727
D4	0.680	0.674	0.774	0.882	13.824
D5	0.800	0.797	0.791	0.375	1.125
D6	0.900	0.887	0.861	1.444	4.333
D7	0.940	0.918	0.908	2.340	3.404
D8	1.350	1.317	1.300	2.444	3.704
D9	6.660	6.736	6.838	1.141	2.673
D10	11.400	11.430	11.556	0.263	1.368
Mean				**1.281**	**5.877**
STD				**0.792**	**4.953**

For the classroom, 131 RGB-D frames were collected. The scanning distance ranges from one to five meters. Both visual RGB-D SLAM and FC RGB-D SLAM are applied to the data set to reconstruct the 3D model. Figure 6.9 and Figure 6.10 show the reconstructed 3D models for the classroom using visual RGB-D SLAM and FC RGB-D SLAM, respectively. In each figure, the whole model appears on the upper left corner and the projected wall appears on the lower left corner, while the remaining sub figures are three zoomed-in photos which emphasize the difference between the FC RGB-D SLAM and visual RGB-D SLAM methods. The projected wall from the proposed method was reconstructed precisely. The edges for both the chair and the lectern can be distinctly distinguished from the FC RGB-D SLAM results.

For further quantitative assessments, after incorporating the length and width of the classroom, eight predefined distances are measured using a total station. A total of ten measured distances are used to conduct the comparison between FC RGB-D SLAM and visual RGB-D SLAM. Table 6.1 shows the error statistics for each model. FC

RGB-D SLAM achieved a 1.30% relative error rate compared with to 5.90% for visual RGB-D SLAM.

6.4 summary

In this chapter, a new FC RGB-D SLAM system is proposed to overcome such problems as the lost tracking and drift of current RGB-D SLAM systems. The system overcomes those problems by adopting both 2D and 3D features to register successive RGB-D frames. In case of brittle RGB-D frames, the system merges both 3D and 2D features to track the camera; as a result, lost tracking is overcome. For normal tracking, the system uses point, line, and plane features to compute the camera pose; thus, the drift problem is minimized. The newly proposed SLAM system uses the existing constraints between 3D features to globally refine the reconstructed 3D model. If any loop closure is detected, FC RGB-D SLAM system adopts a graph-based optimization technique to distribute the closure error to each node and edge of the graph. Then the FC RGB-D SLAM system is compared to the current RGB-D SLAM systems to evaluate its performance. The results demonstrated the advantages of the FC RGB-D SLAM system in terms of both tracking stability and model accuracy.

Chapter 7: Conclusions and future works

7.1 Conclusions

RGB-D cameras capture both RGB and depth images at video recording speed rate. Numerous modeling and navigation applications can adopt such sensors to reduce costs or to integrate with other systems to improve system integrity. For instance, RGB-D cameras are adapted to the mobile mapping system in indoor environments to help smooth cheap IMU data. They are also used in the indoor navigation platform (Tango) produced by Google. Thanks to their low cost and high mobility, RGB-D cameras have become the most popular cameras used in 3D indoor modelling and navigation applications. Unlike outdoor environments, indoor environments have a lot of constraints on features and mapping system requirements (e.g., mobility, processing time, and complexity). Although RGB-D cameras have a great potential to replace the time consuming and expensive traditional ways like laser scanners to reconstructing 3D indoor models, they still suffer from problems that restrict the working range and the scanning space. Some problems are related to the sensors themselves and others are related to the scanning procedures.

Firstly, the main problems related to the RGB-D sensors are limited operational range and depth precision. Normally, RGB-D cameras produce a low resolution RGB image accompanied with depth image for ranges up to nine meters. As a commercial sensor, RGB-D camera lenses suffer from distortion and geometrical inaccuracy of manufacturer's parameters (e.g., focal point location). Those problems can be overcome by adopting a suitable calibration method. Thus, the sensor can function with utmost precision. In this study, we introduce a novel distortion model to compensate for both IR camera and IR projector distortion effects. Also, a new

calibration procedure is proposed to thoroughly calibrate RGB-D cameras. The calibration method has been implemented in MATLAB; thus, an automatic calibration toolbox has been produced to fully calibrate RGB-D cameras. The calibration method rigorously calibrates all the RGB-D camera's geometric parameters and precisely calibrates the full depth range. The method has been examined through three experiments: the 3D models of a plane surface, a right-angle scene, and a room reconstructed using both calibrated and uncalibrated cameras. For the plane surface, two different sensors are examined and revealed that the calibration method gives a relative accuracy of 0.49% and 0.72% compared with 1.26% and 1.95% of the uncalibrated depth for both sensors from manufacturers. Moreover, the calibration method extends the working range from three to nine meters. Calibrated depth can estimate the angle features captured from a five-meter distance with a precision of 0.37 degrees compared to 7.17 degrees using uncalibrated depth. The results of the reconstructed 3D model of the indoor environment demonstrated qualitatively and quantitatively the usefulness of the proposed calibration. The model of an office can be reconstructed with 1.5-cm precision compared to 7.5-cm precising using the default depth.

Secondly, the problems related to the scanning procedure and indoor scene properties have significant effects on quality of 3D models of indoor environments. These effects can be classified into two categories: the scene structure and the SLAM procedure which handles the RGB-D frames to construct the 3D models. Regarding the scene structure, many indoor environments lack distinguishing point features; thus, depending only on these features can create imperfect SLAM results. Imperfect SLAM performance is presented as lost tracking or severe drift in the resulting 3D model.

Extracting additional features besides points from indoor scenes and using the indoor constraints (e.g., parallelism, perpendicularity) in the SLAM algorithm can overcome both lost tracking and drift of current RGB-D SLAM systems. In this study, a novel method is proposed to extract and describe line and plane features using both RGB and depth images; thus, the depth noise and bias effects on descriptor is minimized. Using those 3D features besides point features in current RGB-D SLAM algorithms can achieve a significant improvement in the alignment state of the RGB-D frames. In cases where too few distant point features exist, the current RGB-D SLAM easily fails; however, when the proposed tracking algorithm uses those point features to describe the other existing 3D features in the scene, the camera continues tracking. A novel RGB-D SLAM method (Fully constrained (FC) RGB-D SLAM) has been proposed to reconstruct indoor 3D models from RGB-D frames. The system uses all features existing in the indoor environments (e.g., points, lines, planes) to compute the relative transformation between successive RGB-D frames. Moreover, this system adopts advantageous constraints in indoor scenes to refine the final reconstructed model. FC RGB-D SLAM uses the tracking information to automatically detect the changes in indoor environments, thus the indoor constraints are automatically extracted and refined by the system in the post-processing stage. The constraint conditions are extracted based on the planar features, then the system optimizes those conditions in the global optimization stage. In case loop closure error is detected, then the system uses the graph-based optimization technique to correct the closure error. The system is implemented in MATLAB, two cores are added to the visual RGB-D SLAM: the tracking algorithm with 3D features detection and description, and the global refinement algorithm. The system has been tested using different indoor environments. For instance, applied to a corridor measuring 58m in length, the system can reconstruct the 3D model with up to 0.20m error, which is much better than the 1.00m error and

1.20m error of visual RGB-D SLAM and sensor default tracking algorithms, respectively. For a big room measuring 11.4x6.7m, FC REG-D SLAM achieved 1.30% relative error, which is much better than the 5.90% relative error of visual RGB-D SLAM. Also, the model reconstructed from FC RGB-D SLAM has clearly distinguished objects compared to the model reconstructed using visual RGB-D SLAM.

7.2 Recommendations and future work

To some extent, two major and promising research directions can be followed and they to further contribute to this field. The first direction is related to integration between different kinds of RGB-D cameras. Structured light (SL), Time of flight (ToF), and Stereo cameras (SC) are three basic concepts for RGB-D camera manufacturers. Integration between those concepts can enhance system performance by overcoming depth bias and noise from SL and SC, and the multipath and edges imperfection of ToF, and the limited working range of both SL and ToF. Using integrated camera system can be useful for producing 3D videos, which can be used for automation processes in the construction industry. Calibration and system design are both major research points for this direction.

The other direction is related to the SLAM system. Instead of using the point based visual SLAM, the new research prospects use the photogrammetric principle assisted with the power of computer vision algorithms, especially in the areas of object recognition and tracking. Converting from feature-based SLAM to sematic SLAM seems to be the future of RGB-D SLAM. Object recognition and description as well as pose estimation based on objects are the basic research points for this direction.

References

Abdel-Aziz, Y., & Karara, H. M. (1971). *Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry.* Paper presented at the Proceedings of the Symposium on Close-Range Photogrammetry, Falls Church, VA (pp. 1-19). American Society of Photogrammetry

Ahmed, M. T., Mohamad, M., Marshall, J. A., & Greenspan, M. (2015, June). *Registration of Noisy Point Clouds Using Virtual Interest Points.* Paper presented at the 2015 12th Conference on Computer and Robot Vision (pp. 31-38). IEEE.

Andújar, D., Dorado, J., Bengochea-Guevara, J., Conesa-Muñoz, J., Fernández-Quintanilla, C., & Ribeiro, Á. (2017). Influence of Wind Speed on RGB-D Images in Tree Plantations. *Sensors, 17*(4), 914.

Asus. (2017, January 1). Xtion PRO LIVE. Retrieved from https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/

Basso, F., Pretto, A., & Menegatti, E. (2014, May). *Unsupervised intrinsic and extrinsic calibration of a camera-depth sensor couple.* Paper presented at the Robotics and Automation (ICRA), 2014 IEEE International Conference on (pp. 6244-6249) Hong Kong, China, IEEE.

Bay, H. (2006). *From wide-baseline point and line correspondences to 3D.* Doctoral dissertation, ETH Zurich.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding, 110*(3), 346-359. doi:http://dx.doi.org/10.1016/j.cviu.2007.09.014

Bell, M., & Gausebeck, D. (2014). Capturing and aligning multiple 3-dimensional scenes: Google Patents.

Bell, M., Gausebeck, D., & Beebe, M. (2016). Capturing and aligning three-dimensional scenes: Google Patents.

Besl, P. J., & McKay, N. D. (1992a). A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 14*(2), 239-256.

Besl, P. J., & McKay, N. D. (1992b). *Method for registration of 3-D shapes.* Paper presented at the Robotics-DL tentative (pp. 586-606), International Society for Optics and Photonics.

Bose, L., & Richards, A. (2016, May). *Fast depth edge detection and edge based RGB-D SLAM.* Paper presented at the Robotics and Automation (ICRA), 2016 IEEE International Conference on (pp. 1323-1330), Stockholm, Sweden, IEEE.

Bouguet, J.-Y. (2000). Matlab camera calibration toolbox. *Caltech Technical Report*.

Camplani, M., Mantecon, T., & Salgado, L. (2013). Depth-color fusion strategy for 3-D scene modeling with Kinect. *IEEE Trans Cybern, 43*(6), 1560-1571. doi:10.1109/TCYB.2013.2271112

Canny, J. (1987). A computational approach to edge detection. *In Readings in Computer Vision* (pp. 184-203), Elsevier.

Chen, J., Bautembach, D., & Izadi, S. (2013). Scalable real-time volumetric surface reconstruction. *ACM Transactions on Graphics (TOG), 32*(4), 113.

Chow, J. C., Lichti, D. D., Hol, J. D., Bellusci, G., & Luinge, H. (2014). Imu and multiple RGB-D camera fusion for assisting indoor stop-and-go 3D terrestrial laser scanning. *Robotics, 3*(3), 247-280.

Chow, J. C. K., & Lichti, D. D. (2013). Photogrammetric Bundle Adjustment With Self-Calibration of the PrimeSense 3D Camera Technology: Microsoft Kinect. *IEEE Access, 1*, 465-474. doi:10.1109/access.2013.2271860

Concha, A., & Civera, J. (2017). RGBDTAM: A Cost-Effective and Accurate RGB-D Tracking and Mapping System. *arXiv preprint arXiv:1703.00754*.

Corke, P. (2011). *Robotics, vision and control: fundamental algorithms in MATLAB* (Vol. 73): Springer.

Cornelis, N., & Van Gool, L. (2008, June). *Fast scale invariant feature detection and matching on programmable graphics hardware.* Paper presented at the Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on (pp. 1-8), IEEE.

Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., & Theobalt, C. (2017). BundleFusion: Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Reintegration. *ACM Transactions on Graphics (TOG), 36*(3), 24.

Darwish, W., Chen, W., Tang, S., & Li, W. (2016, April). *Full Parameter Calibration for Low Cost Depth Sensors*. Paper presented at the Melaha 2016 International Conference and Exhibition, Cairo, Egypt. Retrieved from http://melaha2016.ainegypt.org/event/Sessions.aspx?event_id=13&sid=32

Darwish, W., Li, W., Li, Y., Tang, S., & Chen, W. (2017a, December). *Constrained RGBD SLAM for Robust 3D Model Reconstruction of Indoor Environment.* Paper presented at the International Symposium on GNSS ISGNSS 2017, Hong Kong.

Darwish, W., Li, W., Tang, S., & Chen, W. (2017b, June). *Coarse to Fine Global RGB-D Frames Registration For Precise Indoor 3D Model Reconstruction.* Paper presented at the Localization and GNSS (ICL-GNSS), 2017 International Conference on (pp. 1-5), IEEE.

Darwish, W., Tang, S., Li, W., & Chen, W. (2017c). A New Calibration Method for Commercial RGB-D Sensors. *Sensors, 17*(6), 1204.

Díez, Y., Roure, F., Lladó, X., & Salvi, J. (2015). A Qualitative Review on 3D Coarse Registration Methods. *ACM Computing Surveys, 47*(3), 1-36. doi:10.1145/2692160

dos Santos, D. R., Basso, M. A., Khoshelham, K., de Oliveira, E., Pavan, N. L., & Vosselman, G. (2016). Mapping Indoor Spaces by Adaptive Coarse-to-Fine Registration of RGB-D Data. *IEEE Geoscience and Remote Sensing Letters, 13*(2), 262-266. doi:10.1109/lgrs.2015.2508880

Dryanovski, I., Valenti, R. G., & Xiao, J. (2013, May). *Fast visual odometry and mapping from RGB-D data.* Paper presented at the Robotics and Automation (ICRA), 2013 IEEE International Conference on (pp. 2305-2310), Karlsruhe, Germany, IEEE.

Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., & Burgard, W. (2012, May). *An evaluation of the RGB-D SLAM system.* Paper presented at the 2012 IEEE International Conference on Robotics and Automation (pp. 1691-1696), Saint Paul, MN, USA, IEEE.

Endres, F., Hess, J., Sturm, J., Cremers, D., & Burgard, W. (2014). 3-D Mapping With an RGB-D Camera. *IEEE Transactions on Robotics, 30*(1), 177-187. doi:10.1109/TRO.2013.2279412

Fathi, H., Dai, F., & Lourakis, M. (2015). Automated as-built 3D reconstruction of civil infrastructure using computer vision: Achievements, opportunities, and challenges. *Advanced Engineering Informatics, 29*(2), 149-161. doi:10.1016/j.aei.2015.01.012

Fioraio, N., & Konolige, K. (2011, June). *Realtime visual and point cloud slam.* Paper presented at the Proc. of the RGB-D workshop on advanced reasoning with depth cameras at robotics: Science and Systems Conf. (RSS) (Vol. 27).

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM, 24*(6), 381-395.

Fryer, J. G. (1989). Camera Calibration in Non-Topographic Photogrammetry. In H. M. Karara (Ed.), *Non-Topographic Photogrammetry* (pp. 59-70). Virginia: American Society for Photogrammetry and Remote Sensing.

Fryer, J. G., & Brown, D. C. (1986). Lens distortion for close-range photogrammetry. *Photogrammetric engineering and remote sensing, 52*(1), 51-58.

Gallup, D., Frahm, J.-M., Mordohai, P., & Pollefeys, M. (2008, June). *Variable baseline/resolution stereo.* Paper presented at the Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (pp. 1-8), Anchorage, AK, USA, IEEE.

Geosystems (2016). Retrieved from http://hds.leica-geosystems.com/en/HDS-Laser-Scanners-SW_5570.htm

Gokturk, S. B., Yalcin, H., & Bamji, C. (2004, June). *A Time-Of-Flight Depth Sensor - System Description, Issues and Solutions.* Paper presented at the Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on (pp. 35-44), Washington, DC, USA, IEEE.

google. (2016). Tango. Retrieved from http://get.google.com/tango/

Gui, P., Ye, Q., Chen, H., Zhang, T., & Yang, C. (2014, October). *Accurately calibrate kinect sensor using indoor control field.* Paper presented at the Earth Observation and Remote Sensing Applications (EORSA), 2014 3rd International Workshop on (pp. 9-13), Changsha, China, IEEE.

Gupta, T., & Li, H. (2017, November). *Indoor mapping for smart cities—An affordable approach: U*sing Kinect Sensor and ZED stereo camera. Paper presented at the Indoor Positioning and Indoor Navigation (IPIN), 2017 International Conference on (pp. 1-8), Sapporo, Japan, IEEE.

Haggag, H., Hossny, M., Filippidis, D., Creighton, D., Nahavandi, S., & Puri, V. (2013, December). *Measuring depth accuracy in RGBD cameras.* Paper presented at the Signal Processing and Communication Systems (ICSPCS), 2013 7th International Conference on (pp. 1-7), Carrara, VIC, Australia, IEEE.

Halber, M., & Funkhouser, T. (2017, November). Fine-To-Coarse Global Registration of RGB-D Scans. *Proc. Computer Vision and Pattern Recognition (CVPR),* (pp. 1755-1764), Honolulu, HI, USA , IEEE.

Han, J., Shao, L., Xu, D., & Shotton, J. (2013). Enhanced Computer Vision With Microsoft Kinect Sensor: A Review. *IEEE Transactions on Cybernetics, 43*(5), 1318-1334. doi:10.1109/TCYB.2013.2265378

Harley, I. A. (1967). The non-topographical uses of photogrammetry. *Australian Surveyor, 21*(7), 237-257. doi:10.1080/00050326.1967.10440056

Heikkilä, J., & Silvén, O. (1997, June). *A four-step camera calibration procedure with implicit image correction.* Paper presented at the Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, San Juan (pp. 1106-1112), San Juan, Puerto Rico, USA, IEEE.

Henry, P., Krainin, M., Herbst, E., Ren, X., & Fox, D. (2010). *RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments.* Paper presented at the In the 12th International Symposium on Experimental Robotics (ISER).

Herrera C, D., Kannala, J., & Heikkilä, J. (2011, August). *Accurate and practical calibration of a depth and color camera pair.* Paper presented at the Computer analysis of images and patterns (pp. 437-445), Berlin, Heidelberg, Germany, Springer.

Herrera, D., Kannala, J., & Heikkilä, J. (2012). Joint depth and color camera calibration with distortion correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(10), 2058-2064.

Hirschmuller, H. (2005, June). *Accurate and efficient stereo processing by semi-global matching and mutual information.* Paper presented at the Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 2, pp. 807-814), San Diego, CA, USA, IEEE.

Hsiao, M., Westman, E., Zhang, G., & Kaess, M. (2017, May). *Keyframe-based Dense Planar SLAM.* Paper presented at the Proc. International Conference on Robotics and Automation (ICRA) (pp. 5110-5117), Singapore, Singapore, IEEE.

Hu, G., Huang, S., Zhao, L., Alempijevic, A., & Dissanayake, G. (2012, October). *A robust rgb-d slam algorithm.* Paper presented at the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 1714-1719), Vilamoura, Portugal, IEEE.

Huang, A. S., Bachrach, A., Henry, P., Krainin, M., Maturana, D., Fox, D., et al. (2017). Visual odometry and mapping for autonomous flight using an RGB-D camera *Robotics Research* (pp. 235-252): Springer.

Ilin, A., & Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research, 11*(Jul), 1957-2000.

Jia, W., Yi, W. J., Saniie, J., & Oruklu, E. (2012, May). *3D image reconstruction and human body tracking using stereo vision and Kinect technology.* Paper presented at the 2012 IEEE International Conference on Electro/Information Technology (pp. 1-4), Indianapolis, IN, USA, IEEE.

Jiang, R., Jáuregui, D. V., & White, K. R. (2008). Close-range photogrammetry applications in bridge measurement: Literature review. *Measurement, 41*(8), 823-834.

Jung, J., Lee, J. Y., Jeong, Y., & Kweon, I. S. (2015). Time-of-Flight Sensor Calibration for a Color and Depth Camera Pair. *IEEE Trans Pattern Anal Mach Intell, 37*(7), 1501-1513. doi:10.1109/TPAMI.2014.2363827

Kahn, S., Bockholt, U., Kuijper, A., & Fellner, D. W. (2013). Towards precise real-time 3D difference detection for industrial applications. *Computers in Industry, 64*(9), 1115-1128.

Kerl, C., Sturm, J., & Cremers, D. (2013, November). *Dense visual SLAM for RGB-D cameras.* Paper presented at the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems.

Khoshelham, K. (2011). *Accuracy analysis of kinect depth data.* Paper presented at the ISPRS workshop laser scanning (pp. 2100-2106), Tokyo, Japan, IEEE.

Khoshelham, K., Dos Santos, D., & Vosselman, G. (2013). Generation and weighting of 3D point correspondences for improved registration of RGB-D data. *Proceedings of the ISPRS Annals of the Photogrammetry and Remote Sensing and Spatial Information Sciences, 5*, W2.

Khoshelham, K., & Elberink, S. O. (2012). Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors, 12*(2), 1437-1454.

Kinect. (2010, january 15). *Meet Kinect for Windows*. Retrieved from https://dev.windows.com/en-us/kinect

Koenderink, J. J., & Van Doorn, A. J. (1991). Affine structure from motion. *JOSA A, 8*(2), 377-385.

Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., & Burgard, W. (2011, May). *G²o: A general framework for graph optimization.* Paper presented at the 2011 IEEE International Conference on Robotics and Automation (pp. 3607-3613), Shanghai, China, IEEE.

Kytö, M., Nuutinen, M., & Oittinen, P. (2011, January). *Method for measuring stereo camera depth accuracy based on stereoscopic vision.* Paper presented at the IS&T/SPIE Electronic Imaging

(Vol. 7864, p. 78640I). San Francisco Airport, California, US, International Society for Optics and Photonics.

Lachat, E., Macher, H., Landes, T., & Grussenmeyer, P. (2015). Assessment and calibration of a RGB-D camera (kinect v2 sensor) towards a potential use for close-range 3D modeling. *Remote Sensing, 7*(10), 13070-13097.

Lee, T.-k., Lim, S., Lee, S., An, S., & Oh, S.-y. (2012, October). *Indoor mapping using planes extracted from noisy RGB-D sensors.* Paper presented at the Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on (pp. 1727-1733), Vilamoura, Portugal, IEEE.

Lehtola, V. V., Kaartinen, H., Nüchter, A., Kaijaluoto, R., Kukko, A., Litkey, P., et al. (2017). Comparison of the Selected State-Of-The-Art 3D Indoor Scanning and Point Cloud Generation Methods. *Remote Sensing, 9*(8), 796.

Litomisky, K. (2012). Consumer rgb-d cameras and their applications. *Rapport technique, University of California, 20*.

Liu, W., Fan, Y., Zhong, Z., & Lei, T. (2012, July). *A new method for calibrating depth and color camera pair based on kinect.* Paper presented at the Audio, Language and Image Processing (ICALIP), 2012 International Conference on (pp. 212-217), Shanghai, China, IEEE.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91-110.

Macknojia, R., Chávez-Aragón, A., Payeur, P., & Laganiere, R. (2013, January). *Calibration of a network of kinect sensors for robotic inspection over a large workspace.* Paper presented at the Robot Vision (WORV), 2013 IEEE Workshop on (pp. 184-190), Clearwater Beach, FL, USA, IEEE.

Mallick, T., Das, P. P., & Majumdar, A. K. (2014). Characterizations of Noise in Kinect Depth Images: A Review. *IEEE SENSORS JOURNAL, 14*(6), 1731-1740. doi:10.1109/jsen.2014.2309987

Martinello, M., & Favaro, P. (2011). Single image blind deconvolution with higher-order texture statistics. *Video Processing and Computational Video* (pp. 124-151), Springer, Berlin, Heidelberg.

Martinello, M., & Favaro, P. (2012, July). *Depth estimation from a video sequence with moving and deformable objects.* Paper presented at the Image Processing (IPR 2012), IET Conference on (p. 131), London, UK, IET.

Matterport. (2018, January 1). *Immersive 3D Spaces for real-world applications*. Retrieved from https://matterport.com/

McGlone, J. C. (1989). Analytic Data-Reduction Schemes in Non-Topographic Photogrammetry. In H. M. Karara (Ed.), *Non-Topographic Photogrammetry* (2 ed., pp. 37-55). Falls Church, VA: American Society for Photogrammetry and Remote Sensing.

Melbouci, K., Collette, S. N., Gay-Bellile, V., Ait-Aider, O., Carrier, M., & Dhome, M. (2015, May). *Bundle adjustment revisited for SLAM with RGBD sensors.* Paper presented at the 2015 14th IAPR International Conference on Machine Vision Applications (MVA) (pp. 166-169), Tokyo, Japan, IEEE.

Möller, T., & Hughes, J. F. (1999). Efficiently building a matrix to rotate one vector to another. *Journal of graphics tools, 4*(4), 1-4.

Morvan, Y. (2009). *Acquisition, Compression and Rendering of Depth and Texture for Multi-View Video*. *epixea.com*. Retrieved from http://www.epixea.com/research/multi-view-coding-thesis.html

Mur-Artal, R., & Tardos, J. D. (2017). ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics, PP*(99), 1-8. doi:10.1109/TRO.2017.2705103

Navvis. (2018, January). *Digitizing indoors—NavVis*. Retrieved from http://cn01.navvis.com/cn/

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., et al. (2011, October). *KinectFusion: Real-time dense surface mapping and tracking.* Paper presented at the Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on (pp. 127-136), Basel, Switzerland, IEEE.

Occipital. (2014, November 22). *structure sensor.* Retrieved from http://structure.io/

Omar, T., & Nehdi, M. L. (2016). Data acquisition technologies for construction progress tracking. *Automation in Construction, 70*, 143-155.

Pacella, M., & Colosimo, B. (2013). Different Formulations of Principal Component Analysis for 3D Profiles and Surfaces Modeling. *Procedia CIRP, 12*, 474-479.

Pagliari, D., & Pinto, L. (2015). Calibration of kinect for xbox one and comparison between the two generations of Microsoft sensors. *Sensors, 15*(11), 27569-27589.

Park, J.-H., Shin, Y.-D., Bae, J.-H., & Baeg, M.-H. (2012). Spatial uncertainty model for visual features using a Kinect™ sensor. *Sensors, 12*(7), 8640-8662.

Pătrăucean, V., Armeni, I., Nahangi, M., Yeung, J., Brilakis, I., & Haas, C. (2015). State of research in automatic as-built modelling. *Advanced Engineering Informatics, 29*(2), 162-171.

Pillai, S., Ramalingam, S., & Leonard, J. J. (2016, May). *High-performance and tunable stereo reconstruction.* Paper presented at the Robotics and Automation (ICRA), 2016 IEEE International Conference on (pp. 3188-3195), Stockholm, Sweden, IEEE.

Pilu, M. (1997, June). *A direct method for stereo correspondence based on singular value decomposition.* Paper presented at the Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on (pp. 261-266), San Juan, Puerto Rico, USA, IEEE.

Raposo, C., Barreto, J. P., & Nunes, U. (2013, June). Fast and Accurate Calibration of a Kinect Sensor. In *3D Vision-3DV 2013, 2013 International Conference on* (pp. 342-349), Seattle, WA, USA, IEEE.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011, November). *ORB: An efficient alternative to SIFT or SURF.* Paper presented at the Computer Vision (ICCV), 2011 IEEE international conference on (pp. 2564-2571), Barcelona, Spain, IEEE.

Rusinkiewicz, S., & Levoy, M. (2001, May). *Efficient variants of the ICP algorithm.* Paper presented at the 3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on (pp. 145-152), Quebec City, Quebec, Canada, IEEE.

Shibo, L., & Qing, Z. (2012, August). *A New Approach to Calibrate Range Image and Color Image From Kinect.* Paper presented at the 4th International Conference on Intelligent Human-Machine Systems and Cybernetics (Vol. 2, pp. 252-255), Nanchang, Jiangxi, China, IEEE.

Shpunt, A., Rais, D., & Galezer, N. (2010). Reference image techniques for three-dimensional sensing: Google Patents.

Stachniss, C., Frese, U., & Grisetti, G. (2017, April 2). *OpenSLAM*. Retrieved from http://www.openslam.org/

Stückler, J., & Behnke, S. (2012, September). *Integrating depth and color cues for dense multi-resolution scene mapping using rgb-d cameras.* Paper presented at the Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on (pp. 162-167), Hamburg, Germany, IEEE.

Tang, P., Huber, D., Akinci, B., Lipman, R., & Lytle, A. (2010). Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques. *Automation in Construction, 19*(7), 829-843. doi:10.1016/j.autcon.2010.06.007

Tang, S., Zhu, Q., Chen, W., Darwish, W., Wu, B., Hu, H., et al. (2016). Enhanced RGB-D Mapping Method for Detailed 3D Indoor and Outdoor Modeling. *Sensors, 16*(10), 1589.

Tsai, G. J., Chiang, K. W., Chu, C. H., Chen, Y. L., El-Sheimy, N., & Habib, A. (2015). The Performance Analysis of an Indoor Mobile Mapping System with Rgb-D Sensor. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-1/W4*, 183-188. doi:10.5194/isprsarchives-XL-1-W4-183-2015

Turner, E., Cheng, P., & Zakhor, A. (2015). Fast, Automated, Scalable Generation of Textured 3D Models of Indoor Environments. *IEEE Journal of Selected Topics in Signal Processing, 9*(3), 409-421. doi:10.1109/jstsp.2014.2381153

Turner, E. L. (2015). *3D Modeling of Interior Building Environments and Objects from Noisy Sensor Suites*. (Doctoral dissertation), University of California, Berkeley.

Wang, K., Zhang, G., & Bao, H. (2014a). Robust 3D reconstruction with an RGB-D camera. *IEEE Transactions on Image Processing, 23*(11), 4893-4906.

Wang, Y.-T., Shen, C.-A., & Yang, J.-S. (2014b, September). *Calibrated Kinect sensors for robot simultaneous localization and mapping.* Paper presented at the Methods and Models in Automation and Robotics (MMAR), 2014 19th International Conference On (pp. 560-565), Miedzyzdroje, Poland, IEEE.

Whelan, T., Johannsson, H., Kaess, M., Leonard, J. J., & McDonald, J. (2013, May). *Robust real-time visual odometry for dense RGB-D mapping.* Paper presented at the Robotics and Automation (ICRA), 2013 IEEE International Conference on (pp. 5724-5731) Karlsruhe, Germany, IEEE.

Whelan, T., Kaess, M., Johannsson, H., Fallon, M., Leonard, J. J., & McDonald, J. (2015). Real-time large-scale dense RGB-D SLAM with volumetric fusion. *The International Journal of Robotics Research, 34*(4-5), 598-626. doi:doi:10.1177/0278364914551008

Wu, C. (2011). Visualsfm: A visual structure from motion system (2011). *URL* http://www.*cs. washington. edu/homes/ccwu/vsfm, 14*.

Wu, C. (2013, June). *Towards linear-time incremental structure from motion.* Paper presented at the 3D Vision-3DV 2013, 2013 International conference on (pp. 127-134), Seattle, WA, USA, IEEE.

Yamazoe, H., Habe, H., Mitsugami, I., & Yagi, Y. (2012, November). *Easy depth sensor calibration.* Paper presented at the Pattern Recognition (ICPR), 2012 21st International Conference on (pp. 465-468), Tsukuba, Japan, IEEE.

Zeng, M., Zhao, F., Zheng, J., & Liu, X. (2012). A Memory-Efficient KinectFusion Using Octree. In S.-M. Hu & R. R. Martin (Eds.), *Computational Visual Media: First International Conference, CVM 2012, Beijing, China, November 8-10, 2012. Proceedings* (pp. 234-241). Berlin, Heidelberg: Springer Berlin Heidelberg.

Zhang, C., & Zhang, Z. (2011). *CALIBRATION BETWEEN DEPTH AND COLOR SENSORS FOR COMMODITY DEPTH CAMERAS.* Paper presented at the Multimedia and Expo (ICME), 2011 IEEE International Conference on (pp. 1-6), Barcelona, Spain.

Zhang, C., & Zhang, Z. (2014). Calibration between depth and color sensors for commodity depth cameras *Computer Vision and Machine Learning with RGB-D Sensors* (pp. 47-64): Springer, Cham.

Zhang, H., Liu, Y., & Tan, J. (2015). Loop closing detection in RGB-D SLAM combining appearance and geometric constraints. *Sensors, 15*(6), 14639-14660.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(11), 1330–1334.