UNIVERSITY OF KENTUCKY

MASTER THESIS

Universal Real-Time XYZ rectified Reconstruction for RGB-D Cameras

Author:
Sen Li

Supervisor:
Daniel L. Lau

June 1, 2016

Contents

Chapter 1

Introduction

3D reconstruction aims to reproduce the 3D profile of real objects as accurate as possible, which require accurate world $X/Y/Z$ (noted as X^w $/Y^w$ $/Z^w$ henceforth) coordinate values in three dimensional space for every single point of a profile. Ever since the Kinect brought low-cost depth cameras into consumer market, with PrimeSense 3D sensing technology as the core depth determination principle for its first generation, great interest has been invigorated into RGB-D sensors. Camera calibration is a necessary part in 3D reconstruction in order to extract metric information from 3D images, i.e., to determine a translation from Z^w to X^w $/Y^w$ for every pixel based on its row and column. In the mean time, optical and perspective distortion become a problem that stops from getting a good view. On most wide angle prime lenses and many zoom lenses with relatively short focal lengths, especially cheap low quality lenses, barrel distortion would typically be present.

In this research, a more accurate novel method with precise calibration system is brought in for real-time rectification and 3D reconstruction of universal RGB-D cameras.

1.1 3D Reconstruction

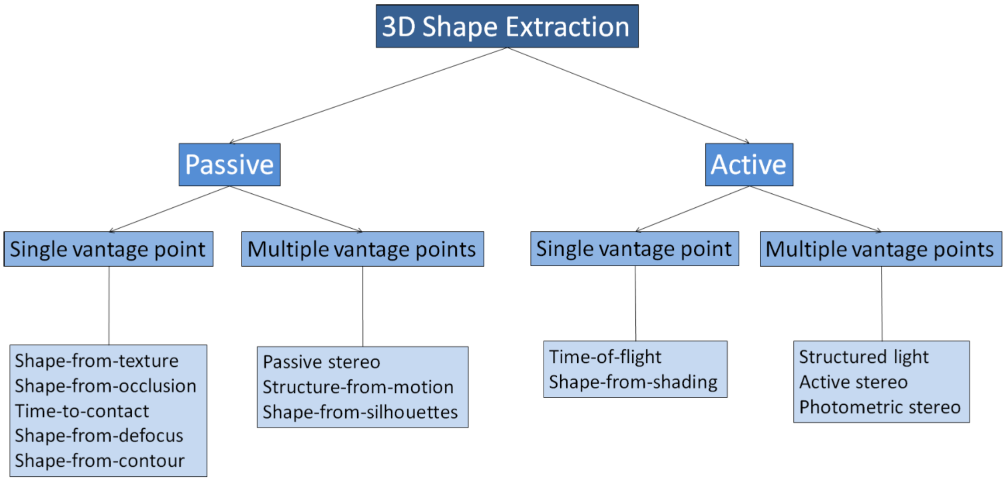


FIGURE 1.1: 3D profile acquisition Taxonomy

Three dimensional (3D) profile measurement technologies have been developed by various means, as summarized by Curless and Seitz [1], among which the non-contact optical methods are widely applied into reality as consumer RGB-D camera. Traditionally, with Pinhole camera model, as the basics of camera calibration, to supply the translation from Z^w to X^w /Y^w , the core procedure of 3D Reconstruction falls on the determination of per-pixel depth to serve as Z^w .

Within the non-contact optical category, as well as 3D reconstruction using multiple images, there are two levels of distinctions[2], as shown in the 3D profile acquisition taxonomy diagram is given in Figure 1.1.

The first distinction: active methods and passive methods. Their classifications are decided by the control of light sources. Active methods need special light sources control as part of the strategy to get 3D information, while on the other hand, passive techniques could work with whichever reasonable available ambient light. With a special known illumination offering more information to simplify some of the steps for 3D information acquiring process, active methods tend to be computationally less demanding. Both of the famous consumer PrimeSense and KinectV2 3D cameras, which are calibrated by the new proposed approach, are using active methods.

The second distinction: single-vantage points methods and multi-vantage points methods. The second distinction is determined by the number of vantage points. With a single vantage system, reconstruction is done based on single view point. In the case that there are multiple viewing or illumination components, all of them would be positioned very close to each other so that they could ideally coincide. For multi-vantage points methods, several viewpoints, with possible controlled illumination source positions, are involved. As contrast with the single-vantage points methods, the multi-vantage systems need the different components to be positioned far enough from each other.

Among the above non-contact optical methods, structured-light and time-of-flight methods are of the most practical importance. As will be discussed shortly, the PrimeSense technology and SeikowaveLCG camera use Stuctured light methods, and the KniectV2 camera uses Time-of Flight.

Structured Light

Structured light (SL) based techniques are famous for its fast speed. It is composed of one camera and one light pattern projector[3]. The projector projects a series of special known patterns onto a target, and the camera captures the corresponding images, which contain special information corresponded to the patterns from the projector. A decoding algorithm would be used to extract world coordinate information of the target object from the captured images, by analyzing the relationship among the camera, the projector and the target object using triangulation.

Being after accuracy, the most important issue for structured light method comes to the question of, how to design the projected patterns. In other words, how to design the coding algorithm and its corresponding decoding strategy will decide the final quality of the reconstructed 3D profile. Various classified SL pattern strategies have been proposed, and are still being studied.

1.1.1 PrimeSense Structured Light

The PrimeSense 3D camera uses an infrared projector to project an infrared speckle pattern onto a target , as shown in figure 1.2,

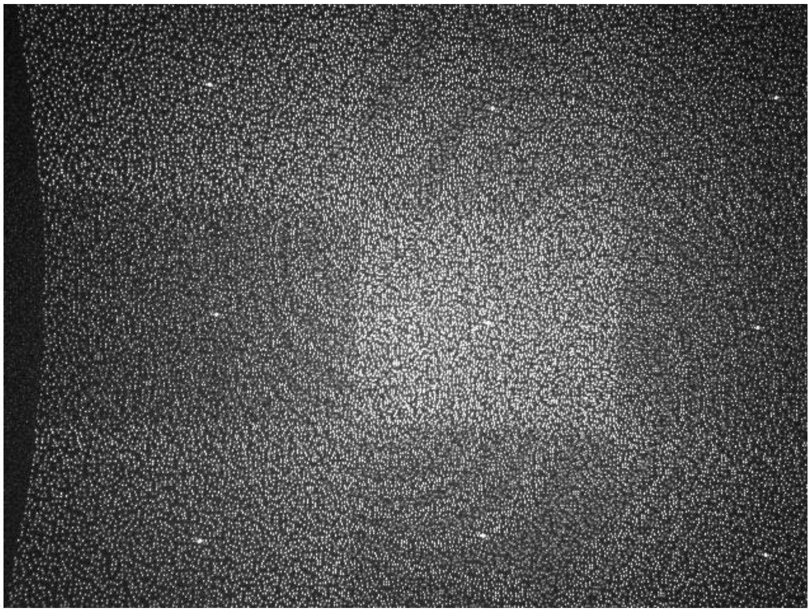


FIGURE 1.2: PrimeSense SL Infrared Pattern

and an infrared camera to capture images of the target. By comparing part by part to reference patterns, that were captured previously at known depths and stored in the device, the per-pixel depth could be looked up based on the reference pattern that the projected pattern matches best.

After the per-pixel depth data determined from the infrared sensor, the next step would be to correlate to a calibrated RGB data, which will generate a popular unified representation of target's profile: point cloud, a collection of points with XYZ 3D coordinates and RGB color data. What's more, the surface normals of the target's profile are also stored in every single point of the point cloud data.

1.1.2 SeikowaveLCG SL Phase Measuring Profilometry

SeikowaveLCG 3D camera consists of a Charge-Coupled Device (CCD) camera and a Digital Micro-mirror Device (DMD) projector. 2D image pattern strategies are always preferred for a fast scan if a is involved. [4][5] And the multi-shot pattern Phase Measuring Profilometry (PMP) strategy was used for its properties of robust and accuracy. With PMP information encoded in the structured light pattern projected onto the target, the CCD camera could capture a series of images that contains PMP informatio. Triangulation analyzing could be used to extract the 3D world coordinates for each points of the target profile, by a determination of the relationships among CCD camera, DMD projector, and the target object. A system configuration of PMP application is given in figure 1.3.

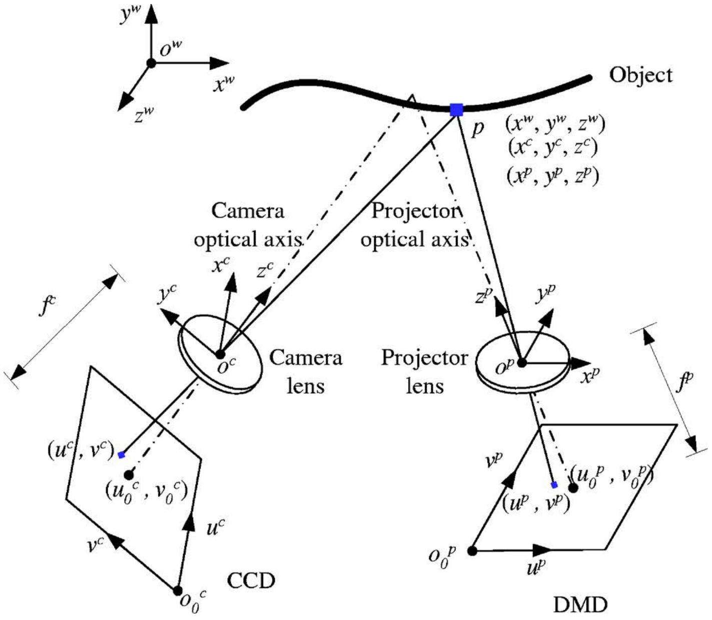
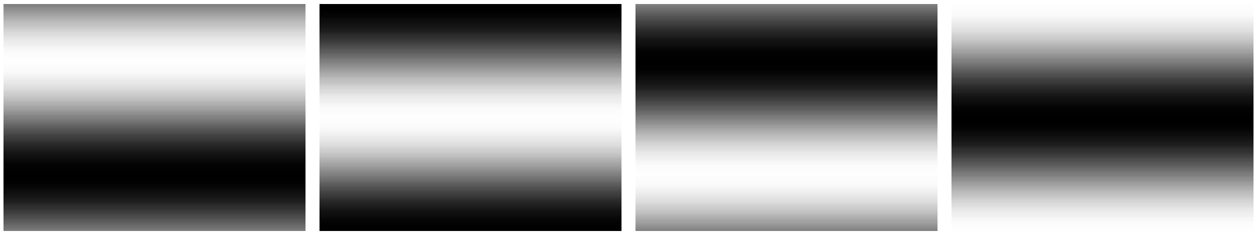


FIGURE 1.3: SL PMP based Configuration Diagram

PMP method uses either vertical or horizontal sinusoid patterns, which could be described as:

$$I_n^p(x^p, y^p) = A^p(x^p, y^p) + B^p(x^p, y^p)Cos(2\pi f y^p - \frac{2\pi n}{N}) \qquad (1.1)$$

where (x^p, y^p) denotes the coordinates of every single pixel in the projector, I_n^p denotes the intensity of the corresponding pixels, A^p and B^p are constants, f is the frequency of sine wave. The subscript n represents the index of phase shift, while capital N is the total number of phase shift.



(A) 1 (B) 2 (C) 3 (D) 4

FIGURE 1.4: PMP base frequency patterns

Figure 1.4 shows a group of sine wave patterns, where the number of total phase shift $N = 4$ and frequency $f = 1$. From viewpoint of the camera, the sinusoid patterns is distorted by the target surface topology, so that the captured images could be expressed as

$$I_n^c(x^c, y^c) = A^c(x^c, y^c) + B^c(x^c, y^c) Cos[\phi(x^c, y^c) - \frac{2\pi n}{N}] \tag{1.2}$$

where(x^c, y^c) denotes the coordinates of every single pixel in the camera, and the term $\phi(x^c, y^c)$ represents the corresponding phase value, which could be computed as follows [6]

$$\phi(x^c, y^c) = arctan\left[\frac{\sum_{n=1}^{N} I(x^c, y^c) Sin(2\pi n/N)}{\sum_{n=1}^{N} I(x^c, y^c) Cos(2\pi n/N)}\right] \tag{1.3}$$

After the camera term $\phi(x^c, y^c)$ for every single pixel is computed, the corresponding projector coordinate y^p could be derived through equation

$$y^p = \phi(x^c, y^c)/(2\pi f) \tag{1.4}$$

With the knowledge of y^p, the perspective information between camera and projector is the last step to go for applying triangulation analysis to extract world coordinates. Based on pinhole camera model, the perspective matrices for both of the CCD camera and DMD projector, as will be derived later in section 1.2.1 equation 1.20, are written as [7]

$$M^c = \begin{bmatrix} m_{11}^c & m_{12}^c & m_{13}^c & m_{14}^c \\ m_{21}^c & m_{22}^c & m_{23}^c & m_{24}^c \\ m_{31}^c & m_{32}^c & m_{33}^c & m_{34}^c \end{bmatrix} \tag{1.5}$$

and

$$M^p = \begin{bmatrix} m_{11}^p & m_{12}^p & m_{13}^p & m_{14}^p \\ m_{21}^p & m_{22}^p & m_{23}^p & m_{24}^p \\ m_{31}^p & m_{32}^p & m_{33}^p & m_{34}^p \end{bmatrix} \tag{1.6}$$

The mapping from 3D world coordinates to 2D camera coordinates are given by

$$x^c = \frac{m_{11}^c X^w + m_{12}^c Y^w + m_{13}^c Z^w + m_{14}^c}{m_{31}^c X^w + m_{32}^c Y^w + m_{33}^c Z^w + m_{34}^c} \tag{1.7}$$

$$y^c = \frac{m_{21}^c X^w + m_{22}^c Y^w + m_{23}^c Z^w + m_{24}^c}{m_{31}^c X^w + m_{32}^c Y^w + m_{33}^c Z^w + m_{34}^c} \tag{1.8}$$

Likewise, the translation from 3D world coordinates to 2D projector coordinates are given by

$$x^p = \frac{m_{11}^p X^w + m_{12}^p Y^w + m_{13}^p Z^w + m_{14}^p}{m_{31}^p X^w + m_{32}^p Y^w + m_{33}^p Z^w + m_{34}^p} \tag{1.9}$$

$$y^p = \frac{m_{21}^p X^w + m_{22}^p Y^w + m_{23}^p Z^w + m_{24}^p}{m_{31}^p X^w + m_{32}^p Y^w + m_{33}^p Z^w + m_{34}^p} \tag{1.10}$$

Since three out of four equations 1.7 ~ 1.10 are enough to solve X^w, Y^w, and Z^w, and y^p is already calculated, the 3D world coordinates $X^w/Y^w/Z^w$ could be derived from Eqs 1.7, 1.8, and 1.10

$$\begin{bmatrix} X^w \\ Y^w \\ Z^w \end{bmatrix} = \begin{bmatrix} m^c_{11} - m^c_{31}x^c, & m^c_{12} - m^c_{32}x^c, & m^c_{13} - m^c_{33}x^c \\ m^c_{21} - m^c_{31}y^c, & m^c_{22} - m^c_{32}y^c, & m^c_{23} - m^c_{33}y^c \\ m^p_{21} - m^p_{31}y^p, & m^p_{22} - m^p_{32}y^p, & m^p_{23} - m^p_{33}y^p \end{bmatrix}^{-1} \begin{bmatrix} m^c_{34}y^c - m^c_{14} \\ m^c_{34}y^c - m^c_{24} \\ m^p_{34}y^p - m^p_{24} \end{bmatrix} \tag{1.11}$$

1.1.3 Time of Flight (KinectV2)

Based on known speed of light, Time-of-Flight (ToF) camera resolves distance by measuring the time cost of a special light signal traveling between the camera and target for every single point. KinectV2 is one of the practical consumer 3D camera that applied the technology of ToF. Using the active modulated infrared source light together with a low-cost CMOS pixel array, KinectV2 realize its an attractive solution that owns compact construction, high accuracy and up to 30 fps frame-rate.

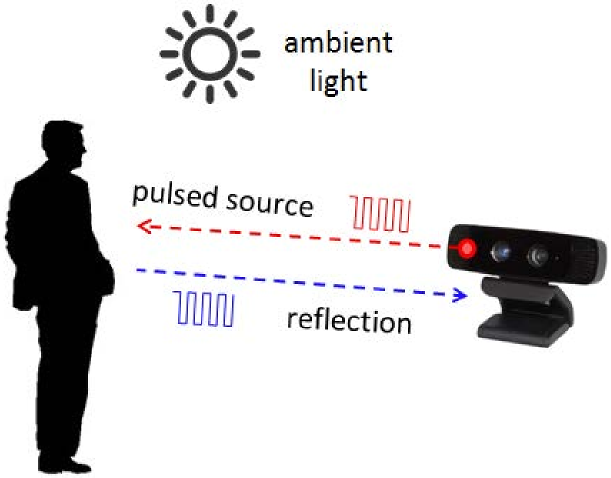


FIGURE 1.5: 3D time-of-flight camera operation

The variable that ToF camera measures is the phase shift between the illumination and reflection, which will be translated to distance [8]. To detect the phase shifts, light source is pulsed or modulated by a continuous wave, typically a sinusoid or square wave. As figure 1.5 shows, the ToF camera illumination is typically from a LED or a solid-state laser operating in the near-infrared range invisible to human eyes. A camera working in the same spectrum captures the reflected light and converts photonic energy to electrical signal, which contains distance (depth) information.

The distance measured for every single pixel is saved into a 2D addressable array, which results in a depth map. KinectV2 has a depth map of 512 * 424 unsigned short data collections, which could be finally rendered, together with corresponded RGB stream, into a tree dimensional space point cloud.

1.2 Traditional RGB-D camera calibration

Traditionally, camera calibration consists of two parts: ideal pinhole camera model calibration, and lens distortion correction. The pinhole camera model works as an simple algorithm in 3D computer vision to describe a mapping from the 3D world coordinates to camera image row and column, thereby giving a translation method from Z^w to X^w and Y^w for every single pixel. It works decently only for ideal pinhole cameras that have no lens, whereas real cameras need extra modifications and supplementations. In order to accurately solve the non-linear distortions problem for a real camera, the second part radial and tangential lens distortion must be considered.

1.2.1 Pinhole Camera Model

Figure 1.6 shows the basic diagram of a pinhole camera model with a reflected image plane for friendly intuition [9].From this model, the mapping between 3D space world coordinate and the image plane row and column could be separated into two parts of transformations. The first part is the transformation between world coordinates system $X/Y/Z$ and camera coordinate system $U/V/W$, which forms a 4-by-4 perspective transformation matrix (**extrinsic calibration**) that works for 3D rotation and translation. And the second part is the mapping between 3D camera coordinates system $U/V/W$ and 2D image plane coordinates u/v, which forms a 3-by-3 perspective transformation matrix (**intrinsic calibration**) that works for not only the rescaling between camera coordinates and virtual ideal image coordinates, but also for translating and skewing between the virtual ideal image plane and real image plane.

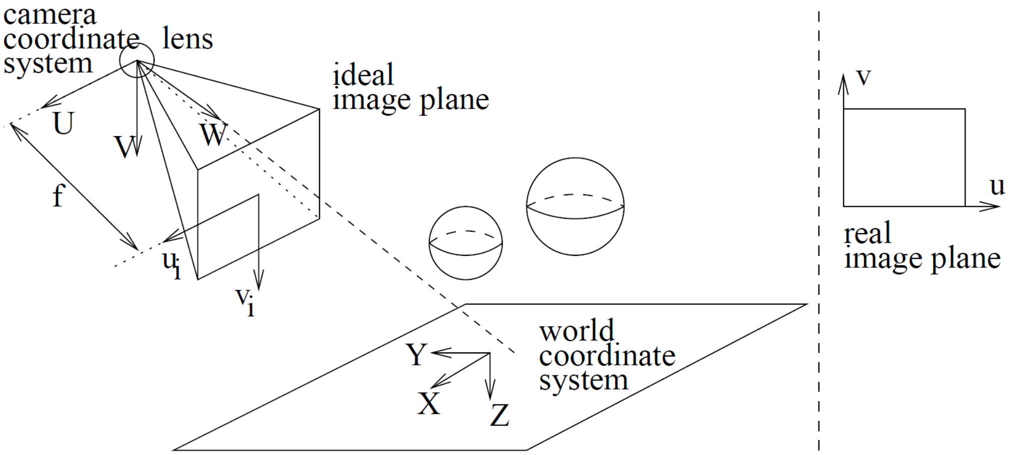


FIGURE 1.6: The pinhole camera model

Extrinsic Calibration

Without any camera parameters, the extrinsic calibration formula could be written, through homogeneous coordinates, as

$$\begin{bmatrix} U \\ V \\ W \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} X^w \\ Y^w \\ Z^w \\ 1 \end{bmatrix} \tag{1.12}$$

or for simplicity,

$$\begin{bmatrix} U \\ V \\ W \end{bmatrix} = \begin{bmatrix} R & T \end{bmatrix} \cdot \begin{bmatrix} X^w \\ Y^w \\ Z^w \end{bmatrix} \tag{1.13}$$

where $(U,V,W)^T$ are the camera coordinates, and the transformation matrix component $[R\ T]$, which is part of the 4-by-4 perspective matrix, is modelling rotation and translation , written as

$$\begin{bmatrix} R & T \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \tag{1.14}$$

Intrinsic Calibration

Intrinsic Calibration could be separated into two sections. The first section is to rescale from camera coordinates to virtual ideal image coordinates. For a easier integration of two sections, the first section's formula is given through both of 2D coordinates to homogeneous coordinates.

2D coordinates:

$$W \begin{bmatrix} u_i \\ v_i \end{bmatrix} = f \begin{bmatrix} U \\ V \end{bmatrix} \tag{1.15}$$

Homogeneous coordinates:

$$W \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} fU \\ fV \\ W \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} U \\ V \\ W \end{bmatrix} \tag{1.16}$$

The second section is for translating and skewing between the virtual ideal image plane u_i/v_i and real image plane u_r/v_r,

$$\begin{bmatrix} u_r \\ v_r \\ 1 \end{bmatrix} = \begin{bmatrix} s_u & s_\theta & u_0 \\ 0 & s_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} \tag{1.17}$$

where (u_0, v_0) denotes the optical center (or principal point), $[s_u, s_v]$ are skew coefficients in pixels along u and v axes , and s_θ is an skewed angle generated by s_u and s_v. To combine equation 1.16 and equation 1.17, we could get

$$W \begin{bmatrix} u_r \\ v_r \\ 1 \end{bmatrix} = \begin{bmatrix} s_u & s_\theta & u_0 \\ 0 & s_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} U \\ V \\ W \end{bmatrix} = \begin{bmatrix} f_u & s & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} U \\ V \\ W \end{bmatrix} \tag{1.18}$$

where $[f_u, f_v]$ denote focal lengths in pixels along u and v after skewing, and s is a new skew coefficient after combination.

Generic Perspective Matrix of the Pinhole Camera Model

After both of the extrinsic and intrinsic transformation matrices have been derived, the generalization formula of the pinhole camera model could be derived, combining equation 1.13 and 1.18, as

$$k\begin{bmatrix} u_r \\ v_r \\ 1 \end{bmatrix} = \begin{bmatrix} f_u & s & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} R & T \end{bmatrix} \cdot \begin{bmatrix} X^w \\ Y^w \\ Z^w \end{bmatrix} = C \cdot \begin{bmatrix} X^w \\ Y^w \\ Z^w \end{bmatrix} \tag{1.19}$$

where W on the left side has been replaced by k to be a more general proportion coefficient, and a combined matrix can be expressed as

$$C = \begin{bmatrix} f_u & s & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} R & T \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{bmatrix} \tag{1.20}$$

The final 3-by-4 matrix C is considered as the generic perspective transformation matrix of a pinhole camera model, which gives a mapping between the 3D world coordinates and 2D real image coordinates.

1.2.2 Lens distortion

Lens distortion could be classified into two groups [10] : radial distortion, and tangential distortion.

Imperfect lens shape causes light rays bending more near the edges of a lens than they do at its optical center. The smaller the lens, the greater the distortion. Barrel distortions happen commonly on wide angle lenses, where the field of view of the lens is much wider than the size of the image sensor.

Improper lens assembly will lead to tangential distortion, which occurs when the lens and the image plane are not parallel.

The lens distortion can be expressed as power series in radial distance $r = \sqrt{x^2 + y^2}$:

$$\begin{aligned} x_{distorted} &= x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + [p_1(r^2 + 2x^2) + 2p_2 xy] \\ y_{distorted} &= y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + [p_2(r^2 + 2y^2) + 2p_1 xy] \end{aligned} \tag{1.21}$$

where higher order parameters are omitted for being negligible; $(x_{distorted}, y_{distorted})$ denote the distorted points, (x, y) denote the undistorted pixel locations, k_i's are coefficients of radial distortion, and p_j's are coefficients of tangential distortion.

1.3 Contributions of this thesis

For RGB-D cameras, RGB steam and Depth steam are two steams that independent but correlated with each other. With respect to every X^w /Y^w correlated single pixel-pair, Depth steam offers the additional voxel world coordinates Z^w , while RGB steam offers the additive color property.

As described in section 1.2, the traditional way to reconstruct 3D point cloud not only has the time-cost problem, but also makes Z^w accuracy unguaranteed. First of all, the lens distortions correction is separated from pinhole camera model calibration. Even though same pixel-pair's of world coordinates and image plan coordinates could be reused to solve radial dominated lens distortions, the calculation of the separated step brings a second-time translation cost for every single pixel of every frame. This is not a good way to do real-time reconstruction. What's worse, the depth resolution deteriorates notably with depth in practical [11], and noises among depth data vary randomly, camera by camera and pixel by pixel; which means a rough point-cloud plane full of bumps and hollows will be reconstructed even though the camera is observing a wall. Like figure 1.8 shows

Figure 1.8: side face of a sags and crests RGB plane

In General, a rectification of 3D coordinates X^w /Y^w /Z^w for every single pixel is needed to get a better view of a target's 3D profile. In this thesis, instead of using model based traditional method, an accurate data based X^w /Y^w /Z^w rectified real-time reconstruction method is proposed. In Chapter 2, X^w and Y^w are rectified separately through a fourth order surface fitting translation from image plane row and column to directly solve the lens distortion problem in equation 1.21, considering that the parameters in that equation with power level larger than 4 are negligible. Then, Z^w values are totally supported from external BLE optical-flow sensor, which accurately tracks camera movements along Z-axis. Finally, a XYZWRGB-D model based lookup table will be generated for real-time reconstruction. Whole calibration system will be introduced in detail in Chapter 3. Results will be shown in Chapter 4.

1.4 Summation

Chapter 2

Real-Time 3D Rectification

2.1 X_w /Y_w Rectifications

2.1.1 simple explaination of Xw/Yw extraction (details in section 3.2)
2.1.2 coordiantes pairs of clusters' center
2.1.3 High order polynomial surface fitting 2.1.4 Mathematical tools a. singular value decomposition (SVD)
b. least square solution with pseudo-inverse method

2.2 Z_w Rectification

2.2.1 Accumulated Z-axis data calibration
2.2.2 Zw polynomial fitting

2.3 XYZWRGB-D Model Based Lookup Table

2.3.1 D based Z polynomial equation
2.3.2 per-pixel beam equation in 3D space

2.4 Real-time analysis

from Kai's SL based beam expansion to data based per-pixel beam in 3D space

2.5 3D Reconstruction differences for PrimeSense, KinectV2, and Prosilica

shader comparison

Chpater 3: System for RGBD Camera Calibration(experiment, collecting data)
3.1 Rail System
3.2 Round Grid for Radial X/Y Distortion Correction (DIP process to extract cluster's center)
3.3 Optical-Flow Sensor for Z Tracking
3.4 Bluetooth Low Energy for wireless communication

Chapter 4: results for 3 types of RGB-D cameras.

Chapter 5: conclusion