2015

Euclidean reconstruction of natural underwater scenes using optic imagery sequence

Han Hu
University of New Hampshire, Durham

EUCLIDEAN RECONSTRUCTION OF NATURAL UNDERWATER SCENES USING OPTIC IMAGERY SEQUENCE

By

Han Hu

B.S. in Remote Sensing Science and Technology, Wuhan University, 2005

M.S. in Photogrammetry and Remote Sensing, Wuhan University, 2009

THESIS

Submitted to the University of New Hampshire

In Partial Fulfillment of

The Requirements for the Degree of

Master of Science

In

Ocean Engineering

Ocean Mapping

September, 2015

This thesis has been examined and approved in partial fulfillment of the requirements for the degree of M.S. in Ocean Engineering Ocean Mapping by:

Thesis Director, Yuri Rzhanov, Research Professor, Ocean Engineering

Philip J. Hatcher, Professor, Computer Science

R. Daniel Bergeron, Professor, Computer Science

On May 26, 2015

Original approval signatures are on file with the University of New Hampshire Graduate School.

ACKNOWLEDGEMENTS

I would like to thank all those people who made this thesis possible.

My first debt of sincere gratitude and thanks must go to my advisor, Dr. Yuri Rzhanov. Throughout the entire three-year period, he offered his unreserved help, guidance and support to me in both work and life. Without him, it is definite that my thesis could not be completed successfully. I could not have imagined having a better advisor than him.

Besides my advisor in Ocean Engineering, I would like to give many thanks to the rest of my thesis committee and my advisors in the Department of Computer Science: Prof. Philip J. Hatcher and Prof. R. Daniel Bergeron as well for their encouragement, insightful comments and all their efforts in making this thesis successful.

I also would like to express my thanks to the Center for Coastal and Ocean Mapping and the Department of Computer Science at the University of New Hampshire for offering me the opportunities and environment to work on exciting projects and study in useful courses from which I learned many advanced skills and knowledge.

Last but not the least, I would like to thank my family: my parents Baiming Hu and Guilan Yao, my wife Fang Yao. They have given me a very carefree environment and unconditional love so that I can concentrate on my work.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

EUCLIDEAN RECONSTRUCTION OF NATURAL UNDERWATER SCENES USING OPTIC IMAGERY SEQUENCE

By

Han Hu

University of New Hampshire, September, 2015

The development of maritime applications require monitoring, studying and preserving of detailed and close observation on the underwater seafloor and objects. Stereo vision offers advanced technologies to build 3D models from 2D still overlapping optic images in a relatively inexpensive way. However, while image stereo matching is a necessary step in 3D reconstruction procedure, even the most robust dense matching techniques are not guaranteed to work for underwater images due to the challenging aquatic environment. In this thesis, in addition to a detailed introduction and research on the key components of building 3D models from optic images, a robust modified quasi-dense matching algorithm based on correspondence propagation and adaptive least square matching for underwater images is proposed and applied to some typical underwater image datasets. The experiments demonstrate the robustness and good performance of the proposed matching approach.

Chapter I – INTRODUCTION

1.1 Background

1.1.1 The Human Vision System

Image-based 3D reconstruction is one of the most significant and traditional research topics in computer vision and photogrammetry. In our daily life, we navigate through and interact with objects in 3D space. Our brain receives and interprets visual signals to locate objects and to coordinate our movements. Human beings use the Human Vision System (HVS) to convert the visual signals into the interpretation and understanding of surroundings. This wonderful and magical system has been both the inspiration and source of the research and realization of the advanced stereo vision systems in image-based 3D reconstruction [12]. The aim of stereo vision is to construct 3D scenes or objects to allow robot and computer to interact with the world. The basic principle of how human beings distinguish the distance to some object is easy to comprehend, as shown in Figure 1.

The application areas would be huge if stereo vision technology can achieve accurate and fast automatic 3D reconstruction. Some simple examples are 3D game entertainment, autonomous driving and vision measurement. The importance of 3D information can be seen by its growth on most industrial and entertainment areas ranging from industrial design to stereo animation and movie. The fast development and evolution of different 3D computer applications have reflected the importance of 3D perceptions in human interactions.

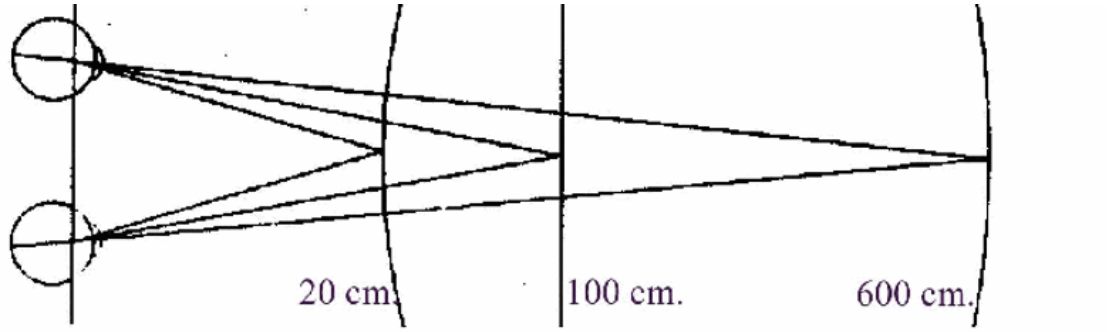


Figure 1. Diagram of Stereo Vision showing curvature at several distances for human vision.

1.1.2 Overview of Stereo Vision

The source images used for 3D reconstruction may come either from a set of closely spaced photographic still images, a hand-held video camcorder, a multi-camera rig with rigidly coupled digital firewire cameras [2], or even from Internet photo collections [3]. Due to its relatively low cost, the image-based 3D reconstruction technique has been widely applied and investigated in many different areas, such as virtual reality, land surveying, simulation, and entertainment. Thus in recent decades, many algorithms and systems have been proposed to contribute to the development of image-based 3D reconstruction. Among all these techniques, the Structure from Motion (SfM) technique in Computer Vision industry has reached a degree of maturity with several commercial offerings, in addition to an extensive research literature [4]. The SfM is an image ranging technique which simultaneously reconstruct the unknown 3D scene structure, camera positions, camera orientations, and even camera calibration parameters, from a set of feature correspondences among images. It has developed from a simple two-frame relative orientation and reconstruction as a basic into more complicated multi-frame approaches. Now in some projects the images can even come from different sources with unfixed intrinsic parameters and can be unorganized and/or unordered [4]. Figure 2 shows the typical steps of underwater 3D reconstruction that have been exploited in previous works using the SfM technique.

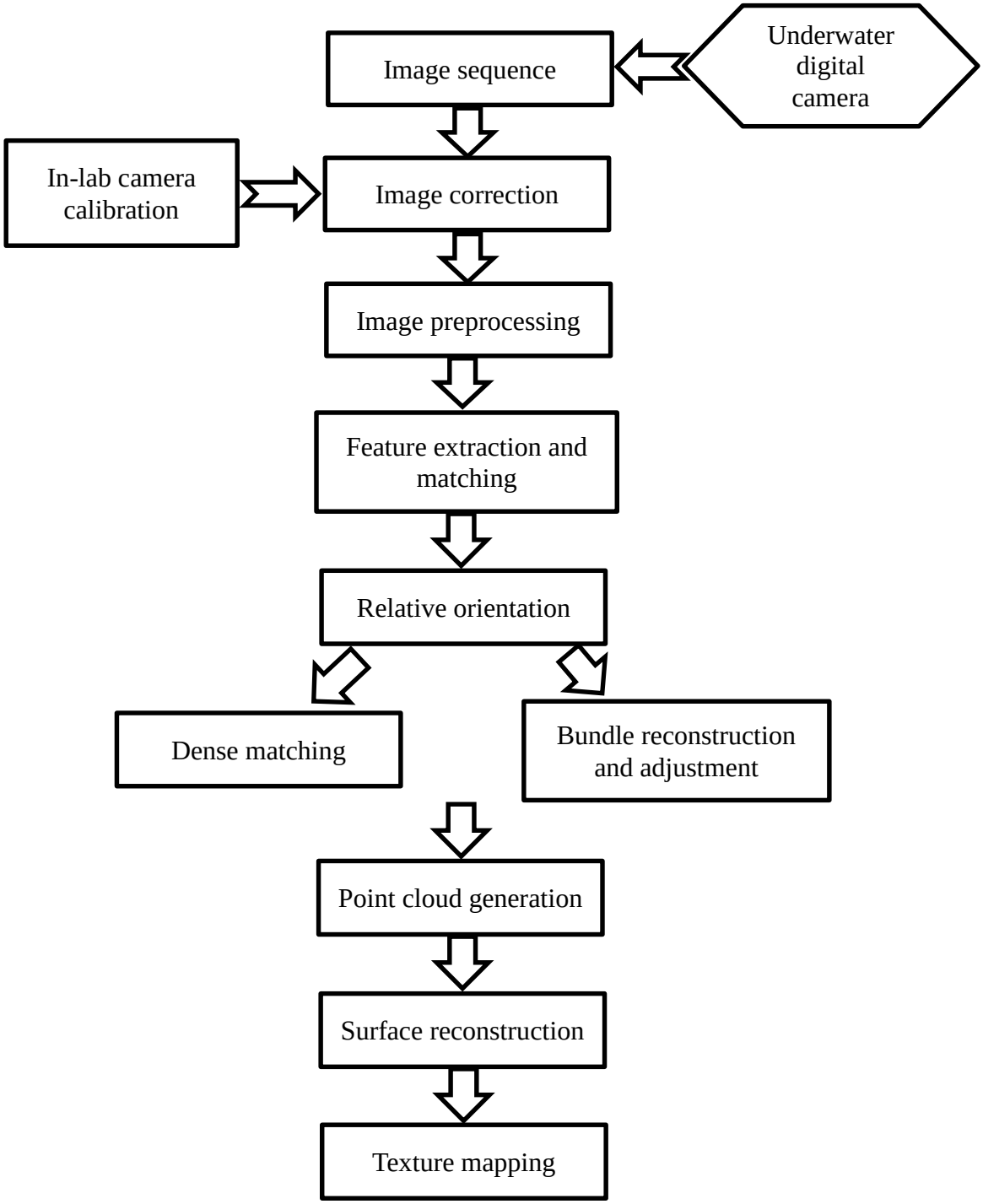


Figure 2. The workflow of 3D reconstruction from image sequence.

The integrated and fundamental theory has lead the SfM technique into the photogrammetry field for a wider research [5]. It has incorporated related techniques from photogrammetry such as bundle adjustment to perform optimal 3D reconstruction from a large set of images.

The development of 3D computer vision algorithms also includes advances in image pre-

processing, feature correspondences, image stereo matching [10], and image-based modeling [3]. Most sparse feature matching algorithms first extract a set of potentially "matchable" image locations, using either interest point operators or edge detectors, and then search for correspondences in other images using a patch-based metric. This sparse image matching strategy is partially due to the limited computational resources because it is almost impossible to match every pixel between two images without any prior knowledge. While the sparse matching algorithms are used to recover the geometry between camera frames, dense matching algorithms focus on determining dense correspondences (pixel to pixel) which are necessary for applications that require detailed surface models. However, due to challenging properties of the medium (water in our case), transferring the traditional standard dense matching methods to other environments seems to be a very challenging task [28]. Image rectification [11] is usually a necessary step in order to perform dense stereo matching. Most dense stereo matching algorithms require either the rectified image pair or the epipolar geometry as input.

1.2 Purpose of Study

1.2.1 Problems of 3D Reconstruction in Underwater Environment

Typically there are two main approaches for 3D reconstruction, namely, laser scanning and image-based approaches [1]. Laser scanners are robust, their measurements are dense and accurate, but they are also costly and have certain restrictions on the size and the surface properties of the objects. The usage of laser scanners is limited by the distance that the laser can propagate in the aquatic environment as well. It is also known that the laser propagation path would be bent due to the changing density at different water depth causing deterioration of measurement accuracy and this system error is hard to calibrate. In addition, the calibration process for a laser device is usually more difficult and complicated than the calibration of a digital camera, especially in an aquatic environment. The reflection from suspended particles

and organisms in water also makes the laser scanned data noisy and harder to process. In recent years, optic sensors have been introduced into underwater vehicles or operated by divers to survey the seafloor or underwater objects. Comparing with the laser scanners, digital cameras have some significant advantages. With the advance of micro and nano electronics in the past few years, digital cameras have become significantly cheaper. They are able to produce high quality images with rich texture, which are important factors to the final reconstruction product. What is more, they are easier for the divers to operate than laser scanners and it is simpler to integrate them into underwater vehicles. In addition, image distortion parameters can be easily calibrated in the controlled environment to compensate for the imperfect linear propagation of underwater optic image rays. In other words, most drawbacks and limitations of underwater imaging system can be compensated by some *in situ* or post-processing strategies. The ease of image-based 3D reconstruction has attracted oceanographers' interest. It is natural for oceanographers to investigate the advanced image-based reconstruction theory and algorithms in order to build 3D models for underwater objects or the seafloor. Underwater objects and structures like black smokers, ship wrecks, or coral reefs, which can only be observed by diving personally or operating a submersible, are difficult to study. However, divers or manned or unmanned underwater vehicles can be equipped with cameras which provide visual image sequences of underwater scenes or objects. The image-based techniques from the computer vision field can be utilized to compute 3D reconstruction and the product can be used for volumetric measurements, documentation or even presented to the general public [8]. Due to the totally different imaging environment, the underwater images have their specific properties and introduce a more challenging job in order to build 3D underwater models.

Summarizing, there are several limitations when processing underwater images. (1) The scene illumination is usually non-uniform. Typically the underwater image is acquired with either ambient or artificial illumination. Both cases lead to time-varying illumination patterns. In the

former case it is due to water column and surface perturbations, and in the latter case - motion of the light source. (Note that this thesis considers imagery acquired by a single moving underwater camera, i.e., classical SfM case.) (2) Light attenuation is range-dependent. Because of those environmental factors, underwater images exhibit different properties compared with the images taken in air. Thus, matching of underwater images becomes difficult in the following aspects: a) underwater images have much fewer salient points, like corners (even man-made objects are often covered with sediments and vegetation), b) the light-attenuating medium (water) leads to a violation of the brightness constancy constraint, c) wavelength-dependent attenuation of light does not allow reliance on color constancy, and d) the presence of suspended particles introduces noise in optical measurements.

1.2.2 Research Purpose

Accurate 3D models of underwater environments will enable us to provide ocean scientists with a tool for making quantitative measurements of submerged structures in addition to sampling and physicochemical measurements [28]. Therefore the need for instrumentation and methodology that enables 3D reconstruction of underwater scenes is a high priority for the scientific community [36].

In this thesis, the aim is to create a system that allows research institutes and organizations to do 3D reconstruction from a sequence of underwater optic images with overlap. The system requires image sequences that do not have too large viewpoint transformations. The image should have sufficient overlap in order to reconstruct the location and orientation for each image viewpoint. In addition, it is advantageous when good illumination conditions are provided, thus the images display large rich texture areas. A robust underwater vision system is capable of managing these challenges in majority of environments and should allow successful image-based reconstruction.

1.3 Related Work

Until recently, cameras have been used in underwater environments for several different purposes, such as monitoring marine habitats, tracking fish populations, reconstructing archaeological sites, and inspecting industrial equipment [38]. The increasingly common practice of deploying cameras aboard underwater remotely-operated vehicles (ROVs) and observation platforms and the mature 3D reconstruction algorithms from computer science have motivated oceanologists to start paying attention to the research of image-based underwater 3D reconstruction. Even though many methods proposed in Computer Vision and Photogrammetry are highly effective in air, they perform poorly in underwater settings [40]. Many attempts have been made to reconstruct underwater objects or seafloors using optic image sequence [6, 7, 8]. A general framework and early techniques for 3D underwater mapping have been introduced in [29]. The method proposed in [30] was used to reconstruct submerged coral reefs in a scientific survey. However, these earlier methods usually resulted in a poor density reconstruction, which reduces the possibility of identification of the detailed structure. Camera trajectory optimization for the reconstruction for a large-scale seafloor mapping with a large number of images was introduced in [28]. However, it still ended with a sparse-textured mesh, which is not sufficient for detailed observation. In [34] Brandou used underwater cameras attached to a stereo rig to capture images. The camera trajectory was preplanned and programmed to traverse a small scene of interest that had to be reconstructed in 3D with known extrinsic parameters. The intrinsic parameters were determined by calibration *in situ* using a planar checkerboard imaged in an image pair. Dense reconstruction was finally performed on the rectified image pair using the graph-cut algorithm. Conditions in this work were different from our work where a single freely moving camera was used and hence external parameters are not known *a priori*. In addition, prior knowledge is required in

order to make and preprogram the camera trajectory for the camera rig. In a same way, [46] combines a 3 DOF inertial sensor and a calibrated stereo rig to estimate the pose of the cameras and create a final 3D dense map. Yau in [38] emphasized the performance of the camera calibration with water refraction taken into account, and proposed a new refraction calibration model and device in order to obtain high reconstruction accuracy. It exploits the dispersion of light and adapts existing reconstruction algorithms for using the proposed calibration method to obtain a complete process for underwater 3D reconstruction. The modification to the dense reconstruction algorithm Patch-based Multi-View Stereo (PMVS) is made to complete dense reconstruction [39]. It uses an eight-camera rig and a calibrated camera pose is required for each image. While offering accurate results, this method has notable limitations. First, the calibration device is rather bulky and can only be used in a controlled environment. Second, precise values of the refractive index are needed for each wavelength. [36] has obtained dense reconstruction for submerged structures, but this method only works in a specific environment where the stereo system was mounted on a controlled manipulator arm, so the camera rotation and translation were known. However, in large-scale underwater mapping applications where a freely moving camera is used, this restriction is prohibitive. Cavan in [41] presents a 3D reconstruction pipeline that is capable of generating photo-realistic 3D models from underwater video acquired by an uncalibrated camera. Although an autocalibration method is presented in his method to transform the projective reconstruction based on trifocal-tensor (3-view geometry) to metric reconstruction, this method is proved to work only on short videos. Overlap of stereo images for video sequences is guaranteed to be larger than that for still images taken by a freely moving single camera.

1.4 Data Acquisition and Description

The optic underwater image sequence for this work was acquired by divers from the Parks

Canada Agency in 2009 using an underwater digital camera. The image sequence is taken of airplane debris on the seafloor near Longue-Pointe-de-Mingan, a small fishing village on Quebec's north shore. The plane - a PBY Catalina -failed at its second take-off attempt and slipped beneath the water on Nov. 2, 1942. The whole dataset consists of 3 separate image sequences. The airplane was imaged by divers from the top, north and south side respectively with the top view including 47 images, north view 42 images and south view 88 images. The original image dimension is 4288 by 2848. The camera intrinsic parameters including focal length and distortion parameters were calibrated underwater in the lab conditions. Figure 4 gives an image overview of a subset of the top side image sequence. In the experiments in this thesis, the original images are sub-sampled into a quarter of the original image size for the sake of efficiency.

Figure 3. Site photos during the acquisition of the underwater image dataset.

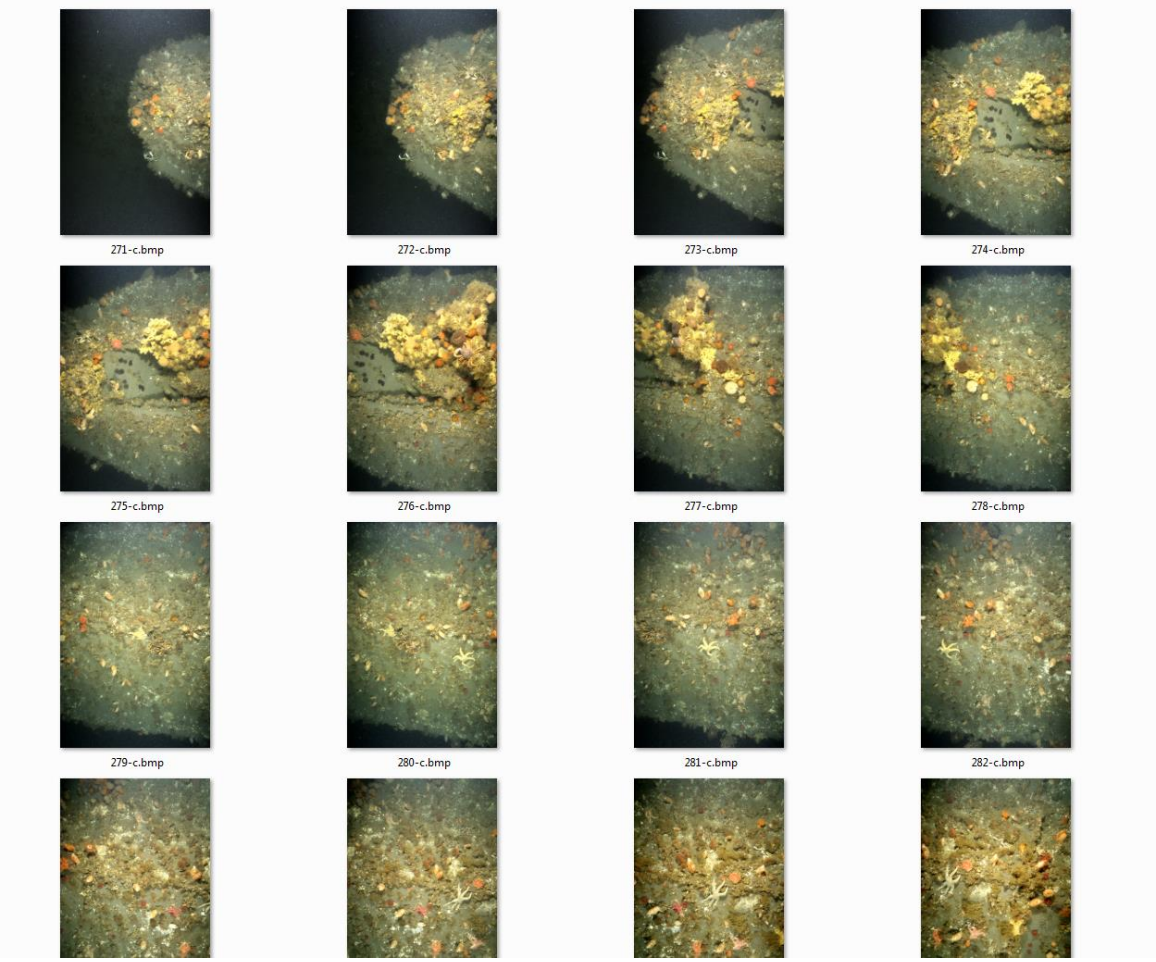


Figure 4. Subset image sequence thumbnails from top side.

1.5 Thesis Outline

The outline of this thesis is as follows. Chapter II introduces the concepts and geometric relationships used in 3D reconstruction of computer vision, including single-view geometry and two-view epipolar geometry. Chapter III discusses image matching techniques and gives an overview of successful sparse image matching and stereo image matching approaches which have been accepted in computer vision. The proposed quasi-dense matching technique is described and discussed in Chapter IV. The workflow and methods for concatenating underwater image sequences and the final bundle adjustment optimization is described in Chapter V. Experimental results and discussion are presented in Chapter VI. Finally, conclusions are made and future work is summarized in chapter VII.

Chapter II - GEOMETRY IN COMPUTER VISION

2.1 Single-View Geometry

A camera in computer vision parlance is a mapping between the 3D world (object space) and a 2D image [13]. Even though there are some different camera models, this work considers only the traditional pinhole camera. In this section only the notations and principles of the central projection (pinhole) camera model are described. Figure 5 depicts the geometry of pinhole cameras. In Figure 5, C denotes the image center and p the image principal point. The distance from the image center C to the principal point is the camera focal length f. The plane $\mathrm{Z} = \mathrm{f}$ is the image plane, and this image plane is perpendicular to the principal optic axis Cp.

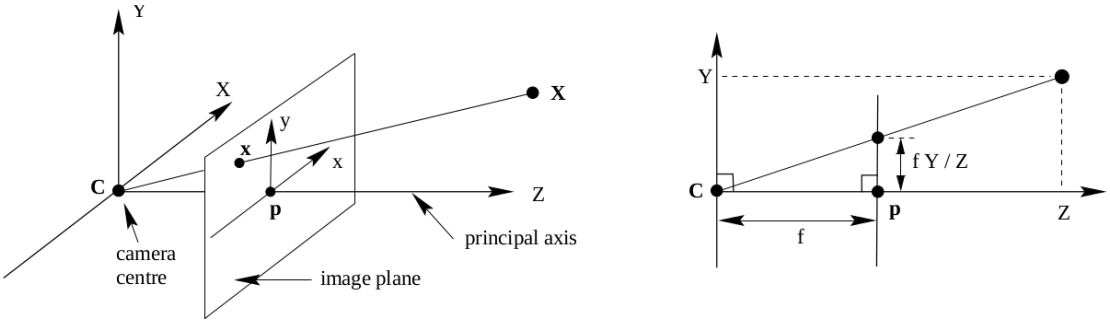


Figure 5. Pinhole camera geometry. C is the camera center and p is the principal point.

The space point $\mathrm{X} = (\mathrm{X}, \mathrm{Y}, \mathrm{Z})^T$ is mapped to image point $\mathrm{x} = (\mathrm{x}, \mathrm{y})^T$. From the right part of Figure 5, it can be noted that $\mathrm{x} = \mathrm{fX}/\mathrm{Z}$ and $\mathrm{y} = \mathrm{fY}/\mathrm{Z}$. In other words, the space point $\mathrm{X} = (\mathrm{X}, \mathrm{Y}, \mathrm{Z})^T$ in 3D space is mapped to the point $(\mathrm{fX}/\mathrm{Z}, \mathrm{fY}/\mathrm{Z})$ on the image plane in 2D space. The projection of a pinhole camera is called the central projection.

The homogeneous or augmented vector is often used in projective geometry. The vector is called a homogeneous vector or an augmented vector if the vector is augmented by one additional scale dimension. Then, if the point coordinates are represented by homogeneous vectors, the central projection can be written in terms of matrix multiplication as

$$\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} fX \\ fY \\ Z \end{pmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{2.1}$$

Now redefine the notation X for the world point represented by the homogeneous 4-vector $(\mathrm{X}, \mathrm{Y}, \mathrm{Z}, 1)^T$, x for the image point represented by a homogeneous 3-vector, and P for the 3×4 homogenous camera projection matrix. Then equation (2.1) can be written as

$$\mathrm{x} = \mathrm{PX} \tag{2.2}$$

The above expression assumes that the origin of the image plane coordinate system is at the principal point. In practice, the left bottom corner of the image is assumed to be the origin as it is shown in Figure 6.

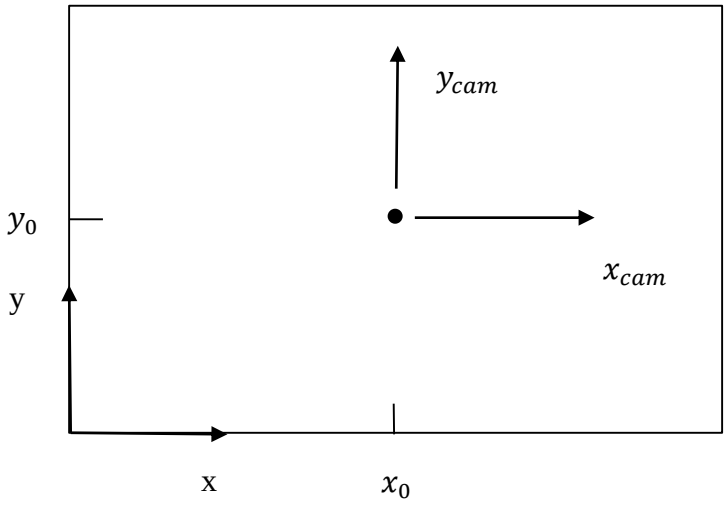


Figure 6. Image (x, y) and camera (x_{cam}, y_{cam}) coordinate systems.

If the left bottom image corner is set as the origin of the image coordinate system, the mapped point is actually $(\mathrm{fX} + \mathrm{Z}p_x, \mathrm{fY} + \mathrm{Z}p_y, \mathrm{Z})$, where (p_x, p_y) is the coordinate of the principal point. Then equation (2.1) should be written as

$$\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} fX + Zp_x \\ fY + Zp_y \\ Z \end{pmatrix} = \begin{bmatrix} f & 0 & p_x & 0 \\ 0 & f & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{2.3}$$

The concise form of equation (2.3) is

$$x = K[I\,|0]X \tag{2.4}$$

where

$$K = \begin{bmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2.5}$$

K is called the camera calibration matrix. It should be noted that the camera calibration matrix may have other forms as can be seen in the literature. A general form of the calibration matrix is

$$K = \begin{bmatrix} \alpha_x & s & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.6}$$

The parameters α_x and α_y may be different for the reason that the digital pixel is not always square. The added parameter s is referred to as the skew parameter. Although this form of calibration has added generality for all pinhole cameras, in this work we will be only using the calibration matrix of form (2.5) since this form is simple but sufficient for the accuracy requirement and the camera is carefully calibrated in the laboratory conditions.

In the above description, the origin of the world coordinate frame is chosen to be the camera center and the Z axis coincides with the optic axis. However, the world coordinate system does not have to be this ideal one. The origin could be set anywhere and the orientation could be in any direction. The actual coordinate frame and the camera coordinate frame are related via rotation and translation. See Figure 7 for illustration. Any 3D point can be transformed between these two coordinate frames – the camera coordinate frame and the world coordinate frame in terms of homogeneous vectors using

$$X_{cam} = \begin{bmatrix} R & -R\tilde{C} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} = \begin{bmatrix} R & -R\tilde{C} \\ 0 & 1 \end{bmatrix} X \tag{2.7}$$

where R is the rotation matrix to rotate from the world coordinate system to the camera coordinate system, $\tilde{C}$ is the coordinate of the camera projection center in the world coordinate

system, X_{cam} is the coordinate with respective to the camera coordinate system with origin located at the camera center and X is the point coordinate with respective to the world coordinate system. Combining with equation (2.4), the general central projection from an arbitrary world coordinate frame to the image projection can be written as

$$\mathrm{x} = \mathrm{KR}[\mathrm{I}|-\tilde{C}]\mathrm{X} \tag{2.8}$$

Then the projection matrix is

$$\mathrm{P} = \mathrm{KR}[\mathrm{I}|-\tilde{C}] = K[R|t] \tag{2.9}$$

where $\mathrm{t} = -\mathrm{R}\tilde{C}$ represents the translation vector from the world coordinate system to the camera coordinate system with the camera center at the origin with respect to the world coordinate frame.

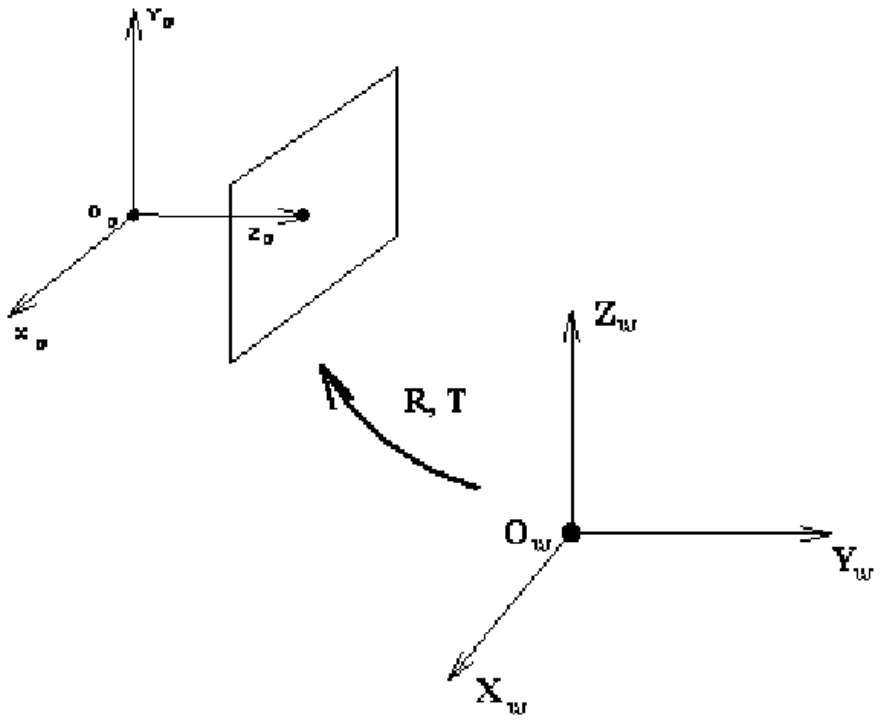


Figure 7. The Euclidean transformation between the world and camera coordinate frames.

2.2 Camera Calibration and Image Correction

In any robot vision system, in order to perform Euclidean reconstruction, the camera must be calibrated to correct for the inherent distortions and misalignments in the imaging system [40].

In the above single-view geometry description, there is an assumption that the space point, image projection and camera optic center are collinear and the principal point is right at the center of image. However, for real cases, due to lenses imperfections, this assumption in general does not hold. The most important deviation of this imperfect camera projection is radial and tangential distortions [13]. Camera calibration consists of finding the parameters internal to the camera that affects the image forming process. In order to obtain Euclidean scene reconstruction, the knowledge of intrinsic camera calibration parameters is essential [35]. The parameters include the position of the principal point (p_x, p_y), the focal length f, the scaling factors for row pixels and column pixels, the skew factor and the radial lens distortion parameters. As we can see from the projection geometry described in the previous section, these parameters define the correspondence between pixel coordinates and point coordinates in 3D space. The calibrated parameters determine the elements in the calibration matrix K.

Most camera calibration algorithms use a calibration pattern to accurately locate points in space. These 3D point coordinates are matched with their corresponding image projections for calculation of camera parameters [49]. In this work, the camera used to take the dataset was calibrated underwater in a tank following the technique described in the Matlab Camera Calibration Toolkit [47]. The typical images taken in a tank before and after correction using calibration parameters are shown in Figure 8. Although the salinity and temperature of the water in the tank were different from that on the real site, lens distortion that remains after the correction appeared to be small enough not to hinder 3D scene reconstruction. However, if higher accuracy is required and the imaging conditions are good enough, it is recommended that the camera be calibrated in the original environment.

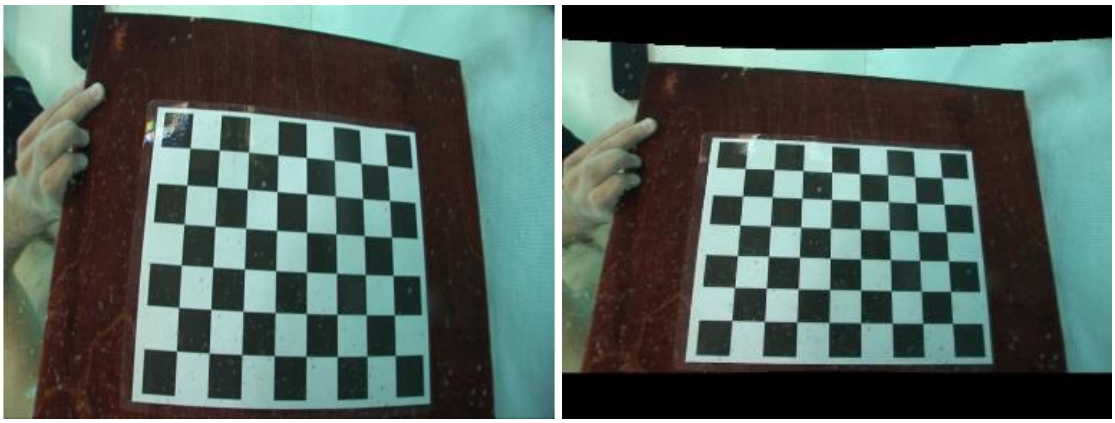


Figure 8. Typical image used for camera calibration before and after correction.

2.3 Two-View Geometry

2.3.1 The Fundamental Matrix

The epipolar geometry is a key concept in computer vision. It defines the relationship between two cameras imaging the same scene. In this section, a brief description will be given on the basic notations and principles of epipolar geometry. Figure 9 illustrates the two-view epipolar geometry. The world point M is projected on two images giving m_1 and m_2. The baseline is defined as the line connecting the camera centers C_1 and C_2. The plane containing the baseline and the world point is called the epipolar plane. The epipolar line is the intersection of the epipolar plane with the image plane. It is denoted by l_1 and l_2 in Figure 9. The most important property of the epipolar line is that the correspondence of m_1 must lie on its corresponding epipolar line l_2. This is easy to see from Figure 9, to each point x in one image, there exists a corresponding epipolar line l' in the other image. Any point x' in the second image matching point x must lie on the epipolar line l'. If the epipolar geometry is known, the search for a correspondence could be limited to a search within the image line.

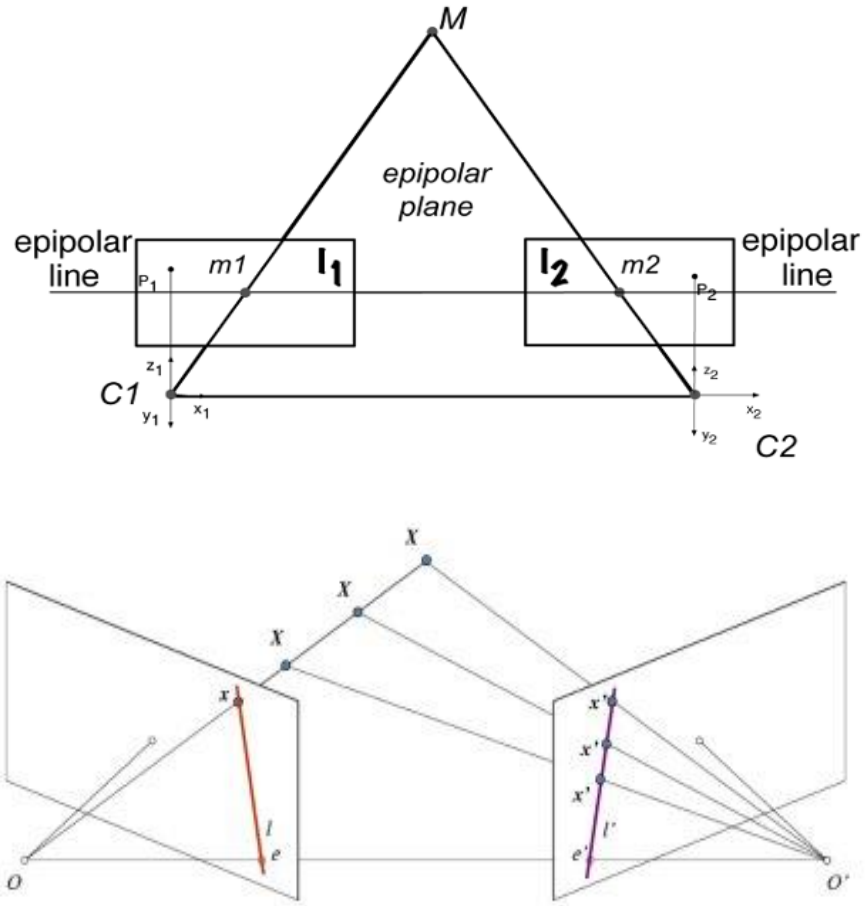


Figure 9. Epipolar geometry.

In order to represent the epipolar geometry algebraically, a fundamental matrix is introduced. The fundamental matrix is a 3 by 3 matrix relating a point from one image to its corresponding epipolar line on another image.

$$\mathrm{F} = \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \tag{2.10}$$

More details about the algebraic derivation of the fundamental matrix can be found in [13]. The property of the fundamental matrix is that it relates the point in one image to the corresponding epipolar line in another image:

$$\mathrm{l}' = \mathrm{Fx} \tag{2.11}$$

where F is the fundamental matrix, x is the point in the first image, and l′ is its corresponding epipolar line in the second image. To get the corresponding epipolar line l for

a given point x' in the second image, the following equation is used:

$$\mathrm{l} = F^T x' \tag{2.12}$$

where F^T is the transpose of fundamental matrix F. And since x lies on l and x' lies on l', we have

$$\begin{aligned} \mathrm{x}'^T \mathrm{Fx} &= 0 \\ x^T F^T \mathrm{x}' &= 0 \end{aligned} \tag{2.13}$$

2.3.2 The Computation of the Fundamental Matrix

The 8-point algorithm is the simplest method of computing the fundamental matrix. The input of this algorithm is a set of greater or equal than 8 point correspondences $\{x_i \leftrightarrow x_i'\}$ and the output is the fundamental matrix. The 8-point algorithm of fundamental matrix is summarized as the following steps.

i. Normalize the input point correspondence: transform the image coordinates according to $\hat{x}_i = \mathrm{T}x_i$ and $\hat{x}_i' = \mathrm{T}x_i'$, where T and T′ are normalizing transformations consisting of a translation and scaling. The suggested normalization is a translation and scaling of each image so that the centroid of the reference points is at the origin of the coordinates and the Root Mean Square (RMS) distance of the points from the origin is equal to $\sqrt{2}$;

ii. Find the fundamental matrix $\hat{F}'$ that corresponds to the normalized matches $\hat{x}_i \leftrightarrow \hat{x}_i'$ by

   (a) Linear Solution: Find linear solution $\hat{F}$ using matrix SVD decomposition;

   (b) Singularity enforcement: Replace $\hat{F}$ by $\hat{F}'$ such that det $\hat{F}'$ = 0 using the SVD;

iii. Denormalize the fundamental matrix solution $\hat{F}'$ from step ii using $\mathrm{F} = T'^T \hat{F}' T$.

Step ii(a) is the essence of this normalized 8-point algorithm and it can be performed using equation (2.13). Writing $\mathrm{x} = (x, y, 1)^T$ and $\mathrm{x}' = (x', y', 1)^T$, the equation (2.13) can be written linearly as

$$\mathrm{x}'\mathrm{x}f_{11} + x'yf_{12} + x'^{f_{13}} + y'xf_{21} + y'yf_{22} + y'^{f_{23}} + xf_{31} + yf_{32} + f_{33} = 0 \qquad (2.14)$$

Equation (2.14) can be written as vector inner product if the fundamental matrix F is written as a 9-entry row-major vector. We denote this vector by f.

$$(\mathrm{x}'\mathrm{x}, \mathrm{x}'\mathrm{y}, \mathrm{x}', \mathrm{y}'\mathrm{x}, \mathrm{y}'\mathrm{y}, \mathrm{y}', \mathrm{x}, \mathrm{y}, 1)\mathrm{f} = 0 \qquad (2.15)$$

Given a set of n point correspondences, a set of linear equations can be obtained of the form.

$$\mathrm{Af} = \begin{bmatrix} x'_1x_1 & x'_1y_1 & x'_1 & y'_1x_1 & y'_1y_1 & y'_1 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_1x_1 & x'_1y_1 & x'_1 & y'_1x_1 & y'_1y_1 & y'_1 & x_1 & y_1 & y_1 \end{bmatrix} f = 0 \qquad (2.16)$$

Since this is a homogeneous set of equations, f can only be determined up to scale. f is the last column of V in the Singular Value Decomposition (SVD) decomposition of $\mathrm{A} = \mathrm{UD}V^T$ and in this way, the solution f is the vector that minimizes ||Af|| subject to the condition ||f|| = 1 [13].

After applying sparse feature extraction and matching, a set of potential image correspondences with false matches are given. Nowadays it has come to be a standard process in computer vision to compute the fundamental matrix given a set of potential image correspondences since the estimation of fundamental matrix is essential to many tasks in 3D computer vision [14]. A standard iterative automatic computation process utilizing the RANdom Sample Consensus (RANSAC) [17] algorithm has been mostly used given a set of point correspondences with a certain ratio of outliers [15, 16]. The following are the basic steps of the automatic computation process of the fundamental matrix.

i. Randomly choose 8 point correspondences from the input set of point correspondences;
ii. Compute the fundamental matrix out from the selected 8 point correspondences using the linear 8-point algorithm;
iii. Evaluate other point correspondences;
iv. Repeat steps i to iii until the best fundamental matrix with the maximum number of good correspondences (inliers) has been found;

v. Reevaluate all the point correspondences using the estimated fundamental matrix from previous steps and determine all the final good correspondences. This is called guided matching.

2.3.3 Retrieving Projection Matrix from the Fundamental Matrix

When the camera intrinsic parameters are known, the essential matrix E [62] can be obtained from the fundamental matrix using equation (2.17).

$$\mathrm{E} = K'^{T} F K \tag{2.17}$$

The essential matrix is the specialization of the fundamental matrix to the case of normalized image coordinates where the camera calibration parameters are known [13]. Comparing to the fundamental matrix, the essential matrix has fewer degrees of freedom - only 5. Both rotation and translation have 3 degrees of freedom, but there is a scale ambiguity which makes the total degrees of freedom to be 5. The most important property of the essential matrix is the capability of direct camera projection matrix extraction from the essential matrix. The camera projection matrices may be extracted from the essential matrix up to a scale and a four-fold ambiguity while there is a projective ambiguity for the fundamental matrix. Given the essential matrix, the projection matrix can be solved in the following steps.

Set the first projection matrix to $\mathrm{P} = \mathrm{K}[\mathrm{I} \mid 0]$ where K is the camera calibration matrix.

Suppose that the SVD of E is $\mathrm{U}\mathrm{diag}(1,1,0)V^{T}$, the second camera projection matrix P' is one of the following matrices

$$K[\mathrm{UW}V^{T}|+u_3] \text{ or } \mathrm{K}[\mathrm{UW}V^{T}|-u_3] \text{ or } \mathrm{K}[\mathrm{U}W^{T}V^{T}|+u_3] \text{ or } \mathrm{K}[\mathrm{U}W^{T}V^{T}|-u_3]$$

u_3 is the last column of U and

$$\mathrm{W} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.18}$$

The four choices of the second camera projection matrix must be tested to determine the correct one with one point correspondence by triangulating the point match. There is only one matrix

among these four choices that makes the triangulated points be in front of both cameras.

Chapter III - IMAGE MATCHING

3.1 Sparse Matching

Sparse matching plays an important role in many computer vision and image processing tasks, such as image mosaic, image recognition, and the determination of the camera pose. Sparse matching aims at establishing correspondences between the same objects which appear in different images based only on the information the images provide. The matching primitives are usually point features, lines and regions, among which point features have been mostly and widely used and investigated. The state-of-the-art image matching algorithms usually consist of two main parts: the detector and the descriptor [18]. The detector searches point features such as corners and edges in the compared images and template images. The descriptor associates each point correspondence with some sort of description of point surroundings. The good descriptor should be invariant to environment changes such as image rotation, scale difference or the illumination condition. The correspondences are established by the comparison and matching between each descriptor pair.

Since sparse matching is usually the first step, the performance of many applications rely on the existence of stable, representative features in the image, driving research and yielding a plethora of approaches to this problem. In other words, when a scene or an object must be reconstructed in 3D, detection and matching points in the image are the most crucial parts for the model accuracy. The 3D model would be of low quality or even completely wrong if the feature extraction and matching steps introduce errors [35]. It has been proved that some local features are robust to occlusion, background clutter and other content changes [19]. The ideal keypoint detector finds salient image regions that can be repeatedly detected despite change of view point; more generally it is robust to as many image transformations as possible. The ideal

descriptor should be able to capture the most distinctive and characteristic information enclosed in the feature regions and should also be invariant to all possible image transformations, thus the features can be matched under different imaging conditions [25]. In recent years, many image feature detectors have been proposed and they are able to reach a certain degree of invariance. Among popular feature detectors are Moravec [44] and Harris [20] point detectors, SIFT[26], SURF [21], KAZE [22], FAST [23, 24], BRIEF [27], BRISK [25], etc. The majority of feature detection algorithms work by computing a corner response function across the image. Pixels at which response exceeds a predefined threshold value (and are local maxima) are then retained [24].

In this thesis, we chose the SIFT detector to be our primary sparse matching method. Lowe's approach [26] is widely accepted as one of the highest quality options in sparse image matching [25] both in performance and computational cost and it has already been integrated into many commercial and public domain development packages such as OpenCV (http://opencv.org). The SIFT features have a high degree of invariance to image scale and rotation. They are also robust to changes in illumination, noise, occlusion and minor changes in viewpoint. The algorithm is performed in the following four stages:

i. Extrema detection in scale space;
ii. Refining keypoints location;
iii. Keypoint orientation assignment;
iv. Generation of keypoint descriptor.

The SIFT detector achieves scale invariance by convolving the image with a Difference of Gaussian (DOG) kernel at multiple scales, retaining locations which have maxima both in scale and space. Let $I(x, y)$ denote the input image. A scale-space of an image is defined as a function $L(x, y, \sigma)$, where σ is the scale factor. A Gaussian scale-space for an input image I, is defined by

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{3.1}$$

, where $*$ represents the convolution operation in both x- and y- direction and the Gaussian filter is defined as

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{3.2}$$

Laplacian of Gaussians as described in equation (3.1) has been proven to be extremely useful because it is stable computationally and gives important information about the scale of regions of an image. But it is computationally expensive, so in practice, Difference of Gaussian (DOG)

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{3.3}$$

is often used as an approximation to Laplacian of Gaussians and it is illustrated in Figure 10.

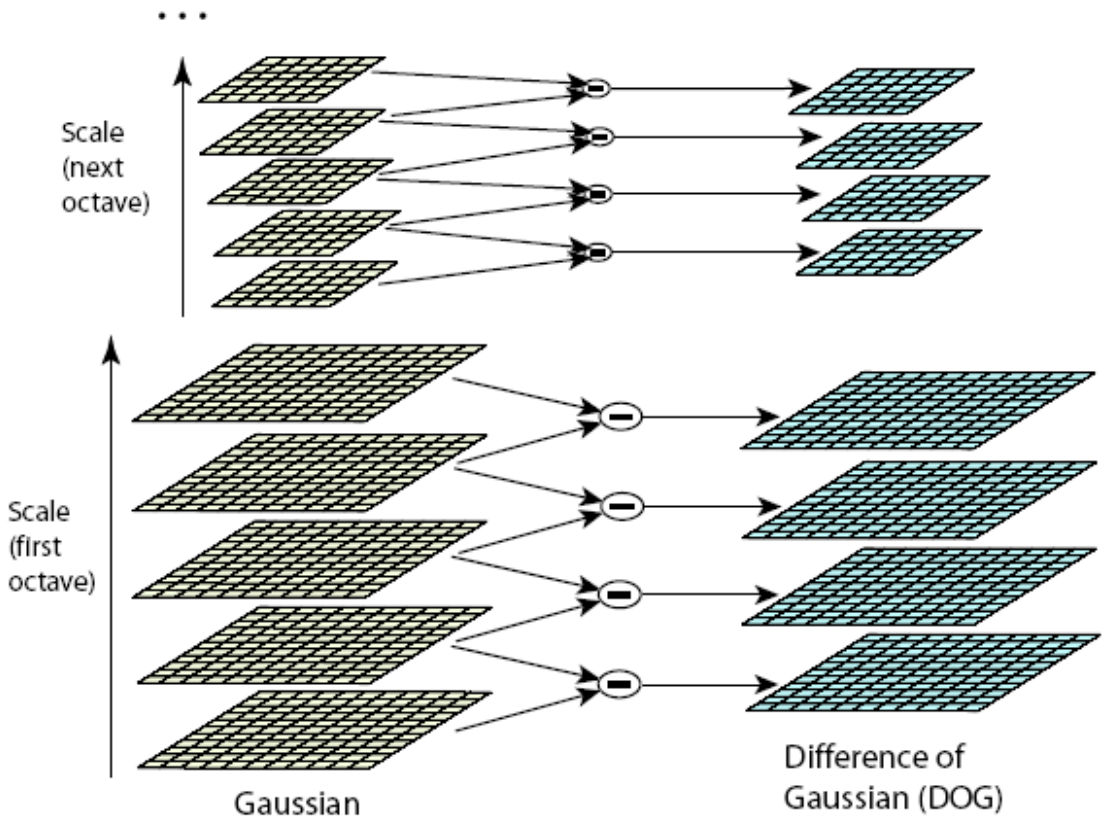


Figure 10. The Difference of Gaussian (DOG) space of image.

The key SIFT feature points are then located at the extreme of the DOG. To enhance the accuracy of localization, the zero crossing of the Taylor series expansion for the DOG is

estimated

$$\mathrm{D(x)} = D_0 + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x \tag{3.4}$$

The accurate key point location and scale is solved using the derivative of equation (3.4) when it is equal to zero.

Each key point is then assigned one or more orientations based on the local image gradient directions which achieves invariance to rotation. For an image sample $\mathrm{L(x, y)}$ at scale σ, the gradient magnitude $\mathrm{m(x, y)}$, and orientation, $\theta(\mathrm{x, y})$, are calculated using pixel differences:

$$\mathrm{m(x, y)} = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \tag{3.5}$$

$$\theta(\mathrm{x, y}) = tan^{-1}\left(\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)}\right) \tag{3.6}$$

The magnitude and orientation calculations for the gradient are done for every pixel in a neighboring region around the key point in the Gaussian-blurred image L. An orientation histogram with 36 bins is formed, with each bin covering 10 degrees. In the histogram, the orientations corresponding to the highest peak and local peaks that is within 80% of the highest peaks are assigned to the key point. Now each key point has a location, scale and orientation. Then a distinctive, scale and rotation invariant descriptor is generated for each key point by calculating gradient histograms around the key point. Typically, a SIFT descriptor is of length 128 (8 orientation bins and 4 by 4 cells for voting) [1]. The Euclidean distance between the descriptor vectors is used as a similarity measurement between features. Figure 11 demonstrates how the SIFT algorithm is invariant to image rotation and scale.

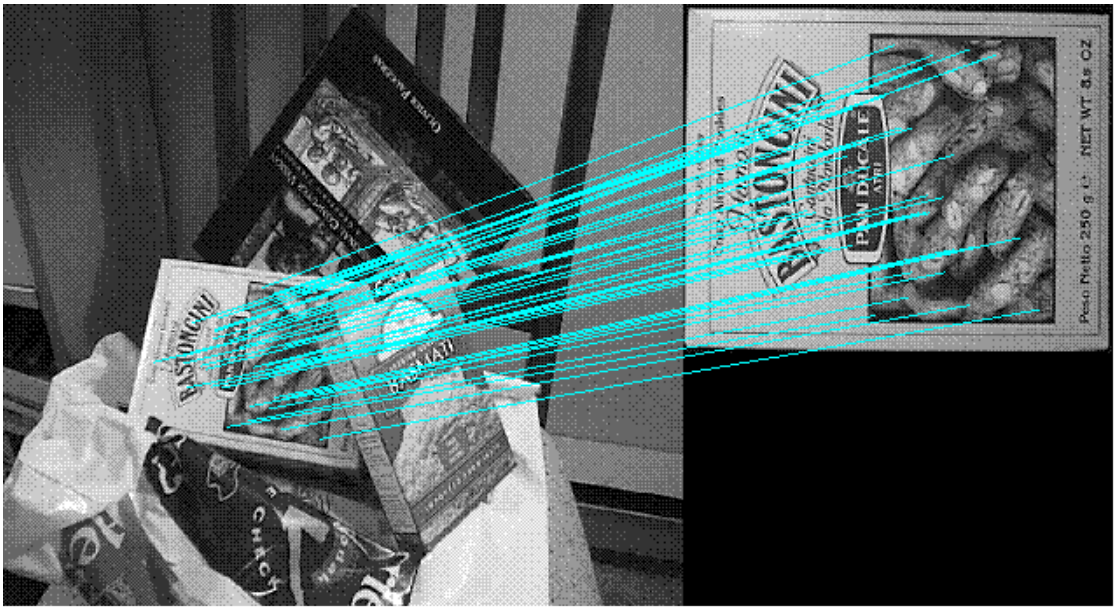


Figure 11. SIFT matching performed on two example images.

We apply the robust method for the fundamental matrix estimation described in section II to remove match outliers and acquire the final sparse matching. The initial set of correspondences are obtained by evaluating the ratio between the first and the second closest descriptor for each feature point [26]. The ratio must be larger than a certain threshold. Figure 12 gives the sparse matches of an example stereo image pair from our dataset.

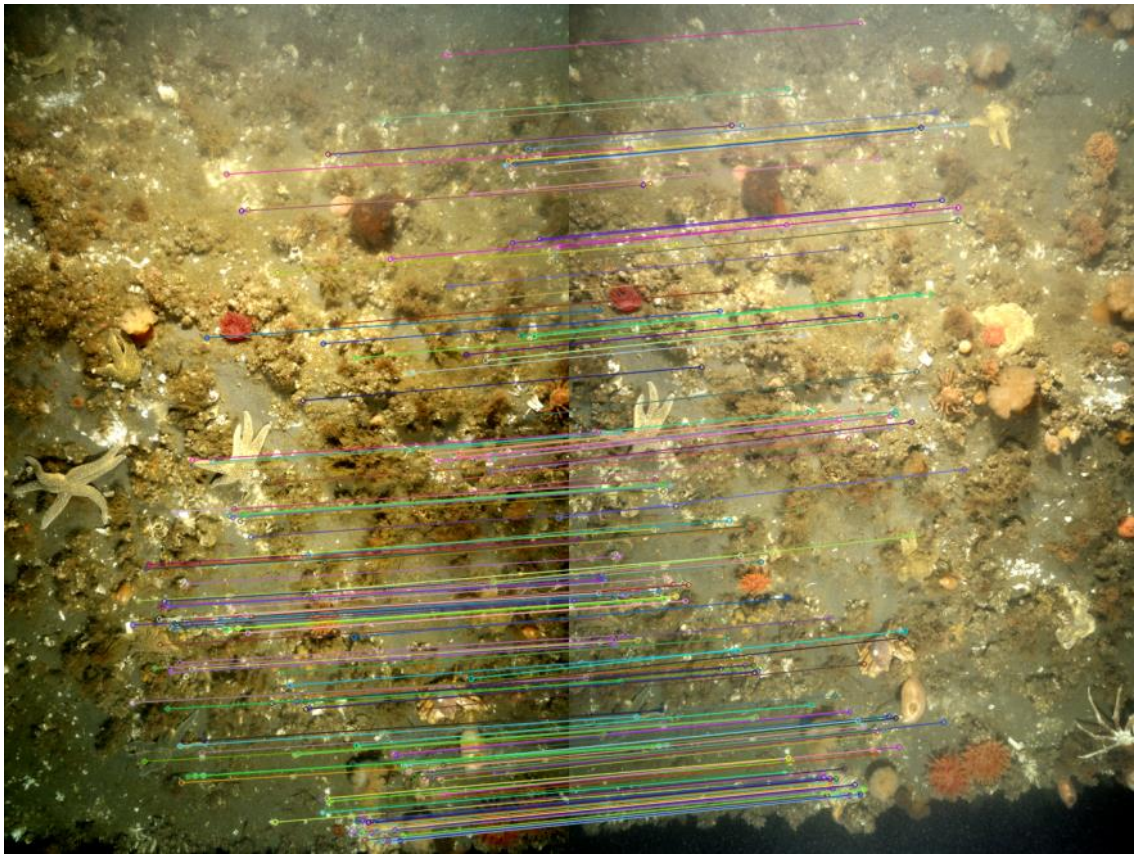


Figure 12. SIFT matches of two example images from our dataset.

3.2 Stereo Matching

3.2.1 Image Rectification

Image rectification is the process of applying 2-dimensional projective transforms, or homographies, to a pair of images whose epipolar geometry is known so that epipolar lines in the original images map to horizontally aligned lines in the transformed images [48]. By rectifying the images, the corresponding epipolar lines coincide. Both computational complexity and the possibility of false matches are greatly reduced.

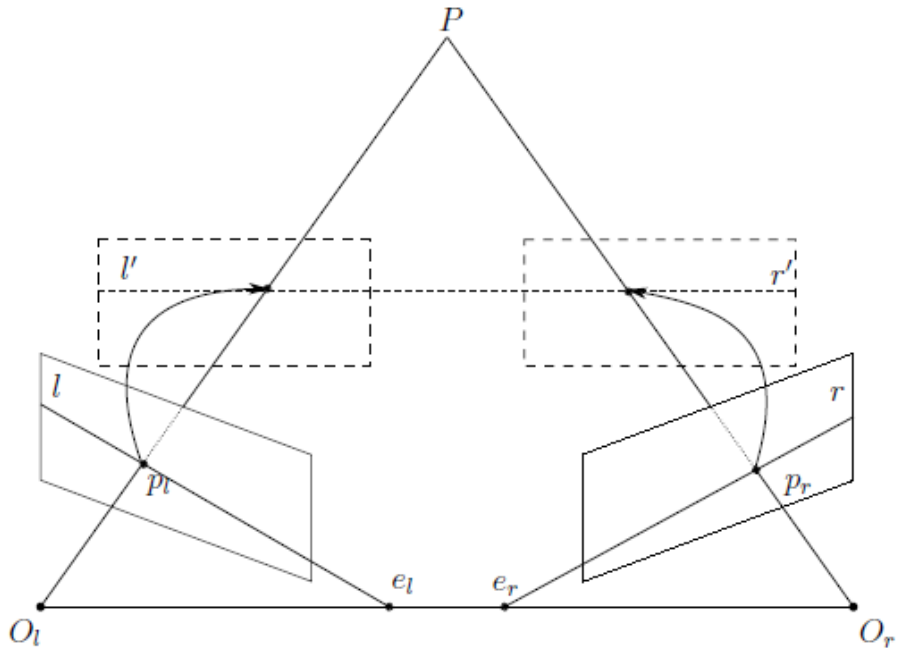


Figure 13. Linear planar rectification process.

Rectification aligns the corresponding epipolar lines to be on the same image scan lines so that the dense correspondence search only happens on the same horizontal line in the other image. The rectification is carried out based on the epipolar geometry that has been computed in the sparse matching step. Typically the methods of image rectification could be grouped into linear planar rectification [48] and non-linear polar rectification [11].

The planar rectification is accomplished by applying a homography to each image that maps the epipole to a predetermined point (Figure 13). The predetermined epipole should be $\mathrm{i} = [1\ 0\ 0]^T$ (a point at infinite) thus all the epipolar lines are parallel with the horizontal axis. The fundamental matrix for this canonical case is

$$\bar{F} = [i]_\times \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \tag{3.7}$$

The image rectification is to find a transformation applied to images I and I' such that the fundamental matrix for these two images have the form of Equation (3.7). Let H and H'

denote the homographies to be applied to images I and I′ respectively, and consider a pair of rectified image point $\bar{m}$ and $\bar{m}'$, we have

$$\bar{m} = \mathrm{Hm} \text{ and } \bar{m}' = H'm' \tag{3.8}$$

where m and m′ denote the original image points. Then we have

$$\bar{m}'\bar{F}\bar{m} = 0 \tag{3.9}$$

$$m'^{T}H'\bar{F}Hm = 0 \tag{3.10}$$

Thus the fundamental matrix for the stereo image pair can be denoted as

$$\mathrm{F} = \mathrm{H}'^{T}[i]_{\times}H \tag{3.11}$$

The homographies H and H′ that satisfy Equation (3.11) are not unique. Charles Loop and Zhengyou Zhang [68] gave a solution for finding a pair of homographies that minimize image distortion. Their proposed method can be summarized into the following steps.

1. Find initial correspondence;
2. Compute the fundamental matrix;
3. Compute a projective transformation H′ that maps the epipole e′ to infinity $(1,0,0)^{T}$;
4. Find the matching projective transformation H that minimizes;

$$\sum_{i} d(Hx_i, H'x_i')^2$$

5. Warp the first image according to H and the second image according to H′.

The advantage of such an approach is that only one transformation matrix needs to be found for each view, thus resulting in high speed and simplicity of algorithms. Another advantage is that rectified images still comply with the perspective camera model. However, a significant drawback of the linear algorithm is that they fail to perform when one or both epipoles are inside the images. Nevertheless, a non-linear algorithm is proposed to perform under this configuration.

The non-linear polar rectification is line-based and uses polar transformation in order to make

corresponding epipolar lines become the same as the image scan lines. This method represents each pair of epipolar lines using the polar equations under the polar coordinate system and then transforms the epipolar lines to the horizontal or vertical image scan lines (Figure 14).

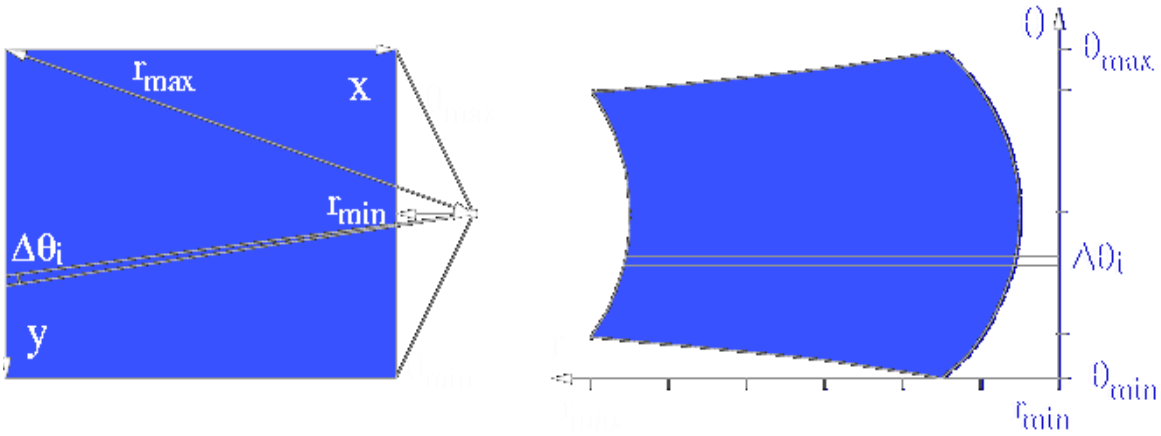


Figure 14 Non-linear polar rectification process.

Hence, the method can handle the case where epipoles are located inside the image which often happens in image acquisition by moving robots as the camera is moving toward or further away from the structure target.

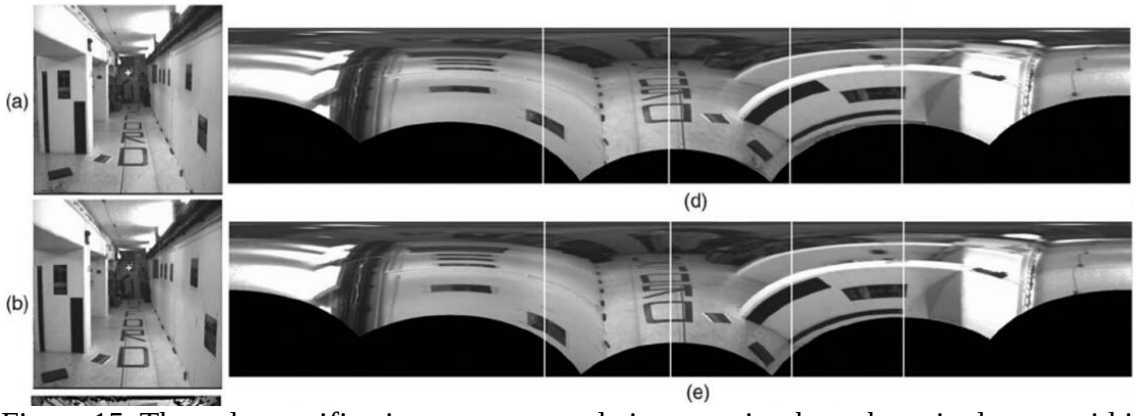


Figure 15. The polar rectification on an example image pair where the epipoles are within the image area.

The polar image rectification method is efficient for underwater environment since the epipole is often inside the image which is caused by moving camera along the viewing direction. Figure 16 presents a pair of underwater rectified images using the polar rectification method. Figure 17 presents the rectified image for the same example image pair using linear planar rectification.

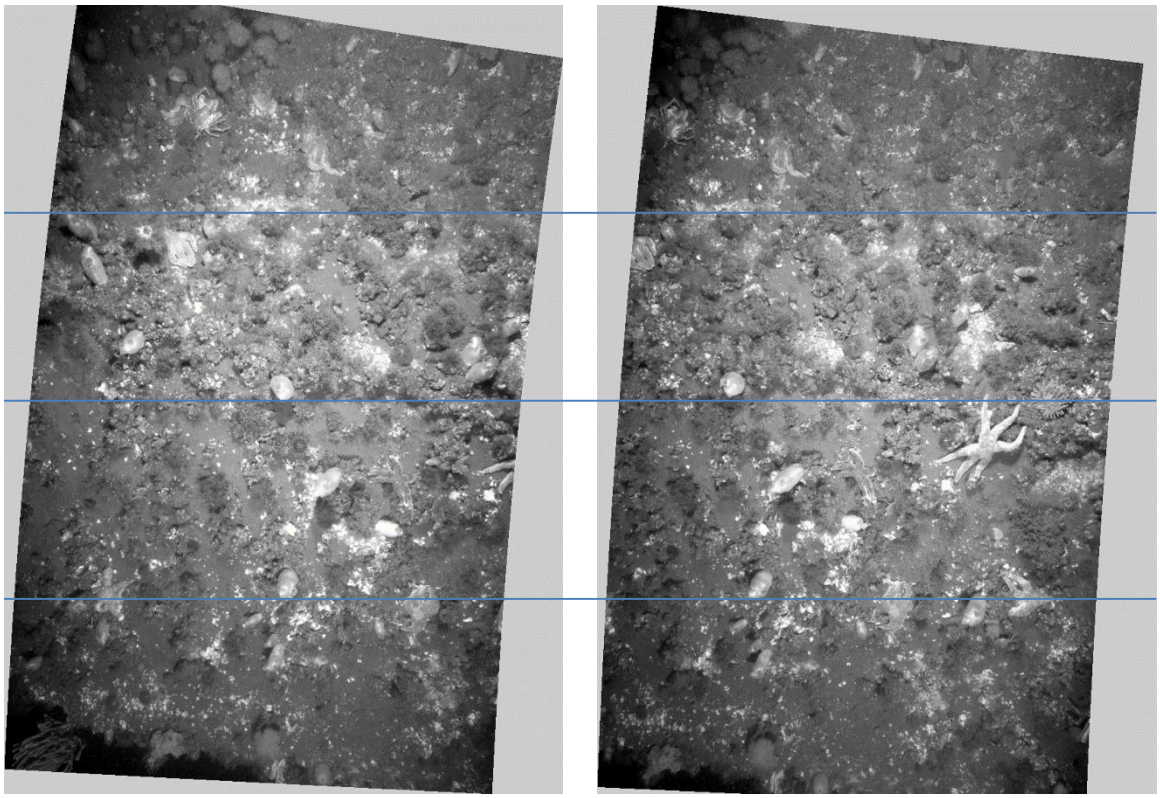


Figure 16. Image Polar rectification on an example stereo image pair from our image dataset. The images are warped so that epipolar lines are oriented on the same horizontal line.

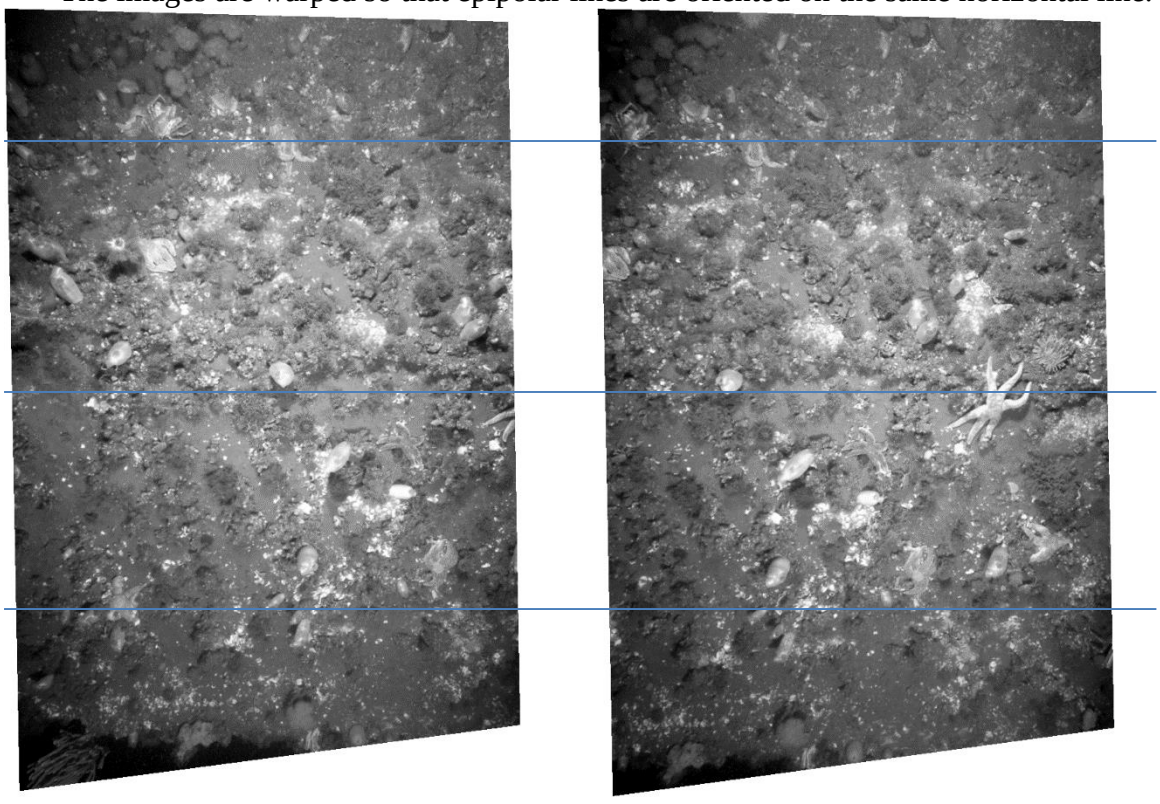


Figure 17. Linear Planar Rectification on an example stereo image pair from our image dataset. The homography transformation is applied for each image.

3.2.2 Stereo Matching Overview

Dense matching aims to search dense correspondences between stereo images for as many pixels as possible. The dense matching is usually carried out after sparse reconstruction, assuming the camera is calibrated and the poses of cameras are known. Dense matching can be classified into two categories: local and global methods, according to the principle they are based on. Local methods compare correspondences one point at a time, not considering neighboring points/measures, while global methods typically seek a disparity assignment that minimizes a global cost function which includes a data term and a smoothness term.

$$\mathrm{E(D)} = E_{data}(D) + E_{smoonthness}(D) \tag{3.12}$$

The data term is the sum of the matching costs of all pixels for a disparity assignment D. The smoothness term is a numeric representation for a disparity assignment D which is typically calculated by adding penalties for each pixel based on the smoothness of its neighboring disparities.

The improvement of accurate stereo matching techniques, as well as the efficiency and computational cost, is one of the most important and investigated topics in computer vision and photogrammetry. Especially over the last few years, a large number of new matching algorithms have been developed. However, two methods – Patch Based Multi-View Stereo (PMVS) [39] and Semi-global Matching based on mutual information [5, 50] - must be mentioned separately in particular due to the high accuracy, robustness and computational efficiency. These two methods have been widely accepted and applied in stereo vision systems. Their successful results have encouraged the algorithms implementation by many researchers. Semi-global matching realizes a pixel-wise matching and relies on the application of a consistency constraint during the cost aggregation. Combining many 1D constraints realized along several paths, symmetrically from all directions through the image, the method performs the approximation of a global 2D smoothness constraint which allows for detection of

occlusions, fine structure and depth discontinuities. It has been implemented in many public domain and commercial software packages, such as OpenCV, SURE and its library interface libTSgm [52] which is developed by the Institute for Photogrammetry of Stuttgart University. The experiments have shown competitive results in point cloud generation from multi-view images even compared with laser scanned data. Figure 18 is the point cloud generated using the semi-global matching method on aerial images.

Figure 18. 3D point cloud derived from aerial images using SURE.

The other top-performing method that has to be mentioned is Patch Based Multi-View Stereo (PMVS) developed by Furukawa. The scene representation in PMVS consists of a set of small rectangular patches. Each patch is parameterized by its center c and normal direction n, approximating a local tangent plane of the true surface. Each patch is associated with a photometric discrepancy score which measures the difference in its appearance between two images. This is done by the patch projection back into images and only patches with low discrepancy score are chosen to represent the surface. This is followed by patch expansion and filtering for a fixed number of iterations. The expansion step creates a new patch next to existing ones and the parameters of the new patch are adjusted to minimize the photometric discrepancy which is similar with the patch initialization step. The filtering step eliminates

incorrect matches either in front or behind the observed surface using visibility constraints. The PMVS have been demonstrated to perform efficiently and produce accurate results on many different datasets. Some example results are shown in Figure 20. A Graphical User Interface (GUI) application for 3D reconstruction using SfM called VisualSFM has integrated PMVS as its dense reconstruction algorithm. The program is easy to use and runs in the following three steps: image matching, sparse reconstruction and dense reconstruction using the PMVS algorithm. Sample results on close range images are shown in Figure 20.

Figure 19. The overall approach of CMVS. From left to right: a sample input image; detected features; reconstructed patches after the initial matching; final patches after expansion and filtering; polygonal surface extracted from reconstructed patches.

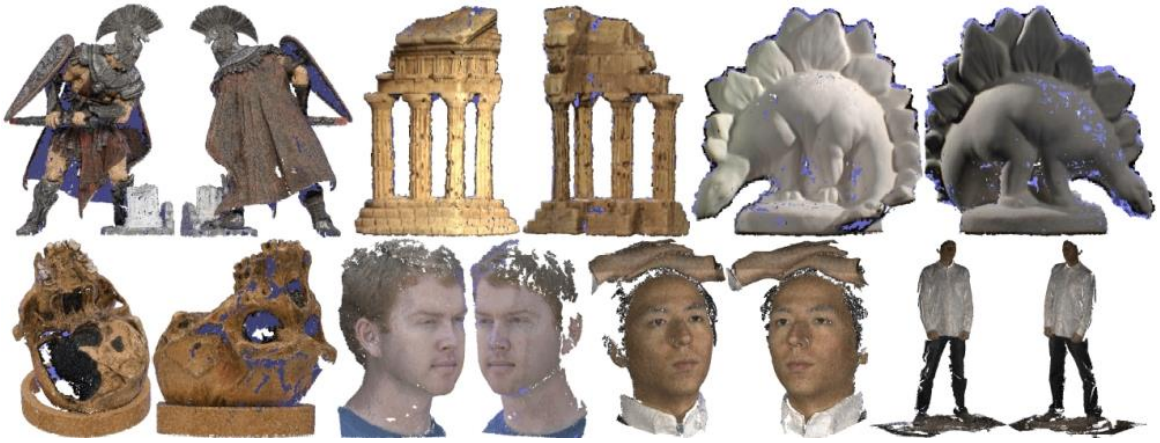


Figure 20. Sample results of PMVS2.

In this work, we have experimented with our underwater image dataset using these two state-of-the-art methods which have achieved great success and have been widely accepted in the computer vision community. Unfortunately, neither of them was able to produce stable or

comprehensive reconstruction result on our experimental underwater images. The reason in general is that the different methods used on the surface are not robust to changes caused by the underwater medium and hence the results are unstable [35]. In the last experimental section of this thesis, the results using these two methods will be compared with our proposed quasi-dense matching method.

Chapter IV - QUASI-DENSE MATCHING

Most difficulties for accurate dense stereo matching are caused by occlusions, object boundaries, and fine structures which can appear blurred. The specific properties of the medium make the dense matching even more difficult, thus the previous works [29-31] usually result in sparse and low resolution models because only few distinct features can be robustly matched with the idea that these features are more robust to artifacts from the medium effects [28]. The sparse approach is sufficient for computing or tracking camera positions, but is not adequate for full representation of the scene or objects as it merely reconstructs sparsely distributed 3D points. When an application demands the detailed structure, dense matching is especially important. Hence, a modified quasi-dense matching method is required and is proposed in this work to produce a satisfactory reconstruction.

4.1 Review of Quasi-Dense Matching

Quasi-dense matching was first proposed by Lhuillier and Quan [56, 57] and developed further to be applicable to wide baseline images [58] and multiple views [60]. This method is considered to be the "golden mean" between sparse and dense approaches [61] and it was motivated by the deficiencies of the existing sparse approaches. The method is based on the propagation of point correspondences into their neighborhoods and is able to deliver a set of high density 3D points from calibrated images. This propagation strategy could also be justified as the seed matches are the points of interest which are the local maxima of the texture richness so the matches could be extended to its neighbors which still have rich texture though are not local maxima.

The method first sorts the sparse reliable point correspondences from sparse matching by their correlation scores. These sorted point correspondences are called the seed points. All initial

seed matching are starting points of concurrent propagations. At each step of propagation, the best corresponding pixels are chosen from the seed points and removed from the current set of seed matches. Then, in the immediate spatial neighborhood of this chosen point, potential matches are searched and the best ones are added to the current list of seed points and to the set of accepted matches according to a combined consideration of local constraints such as correlation, gradient disparity and confidence (Figure 21).

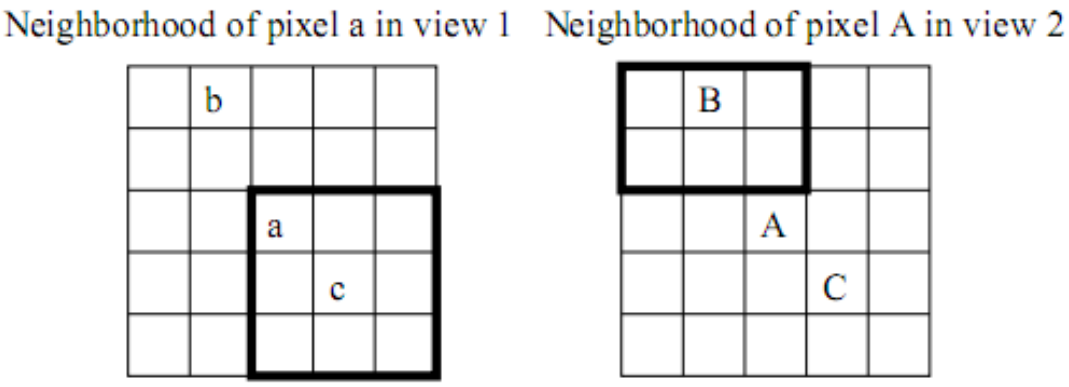


Figure 21. Definition of neighborhood N(a, A) of pixel match (a, A). It is a set of matches included in the two 5×5-neighhborhood $N_5(a)$ and $N_5(A)$. Possible matches for b (resp. C) are in the 3×3 black frame centered at B (resp. c). The complete definition of N(a, A) is $\{(\mathrm{b},\mathrm{B}), \mathrm{b} \in N_5(a), B \in N_5(A), (B-A)-(b-a) \in \{-1,0,1\}^2\}$.

Only new matches that have not been matched yet are chosen to be the final matches to guarantee matching uniqueness. The whole process ends when no more seed points can be used to propagate for finding new matches. The propagation algorithm can be described as follows. Let $\mathrm{s(x)} = \max\{|\mathrm{I(x}+\Delta) - \mathrm{I(x)}|, \Delta \in \{(1,0),(-1,0),(0,1),(0,-1)\}\}$ be the estimate of luminance roughness for the pixel at x, which is used to stop propagation into insufficiently textured areas with $\mathrm{s(a)} < \mathrm{t}$ where $\mathrm{t} = 0.01$ and $\mathrm{I(a)}$ is scaled to be from 0 to 1. The input of the algorithm is the set "Seed" of the current seed matches and the output is an injective displacement mapping "Map".

```
Input:  Seed
Output: Map
while Seed ≠ ∅ do
    pull the best match (a,b) from Seed
    Local ← ∅
    (Store in Local new candidate matches)
    for each (c,d) in N(a,b) do
      if (c,*) and (*,d) not in Map and
        s(c)>t, s(d)>t and ZNCC(c,d)>0.5
      then store match (c,d) in Local
      end-if
    end-for
    (Store in Seed and Map good candidate matches)
    while Local ≠ ∅ do
      pull the best match (c,d) from Local
      if (c,*) and (*,d) not in Map
      then store match (c,d) in Map and Seed
      end-if
    end-while
end-while
```

Figure 22. Pseudo code of quasi-dense matching algorithm.

The quasi-dense matching method is a very efficient algorithm both in time and space. The most significant property of this method is the robustness which is necessary in underwater imagery matching due to the low image quality. At each step, only the most reliable point from the seeds are chosen to propagate for new matches and this drastically limits the possibility of bad matches. In other words, the risk of bad propagation is considerably diminished by the best-first strategy. Since the propagation approach produces denser but not completely dense pixel correspondences, it is called quasi-dense matching. This quasi-dense matching algorithm can be applied not only to surface reconstruction, it is also useful for fundamental matrix estimation [62]. The latter application is implemented by locally fitting planar patches encoded by homographies in order to obtain more accurate correspondences with sub-pixel accuracy. In this case, the epipolar geometry is not added as a constraint in correspondence propagation [62]. Figure 23 illustrates the good performance of the quasi-dense matching algorithm on the "flower garden" dataset to generate a depth map. It should be noticed that the pair of "flower

garden" images are very difficult to match by classical correlation or dynamic programming as the disparity is large and the ordering constraint along the eipolar lines is violated. Even a small number of initial seed matches with a large portion of false matches could provoke and avalanche of correct matches in the whole image areas.

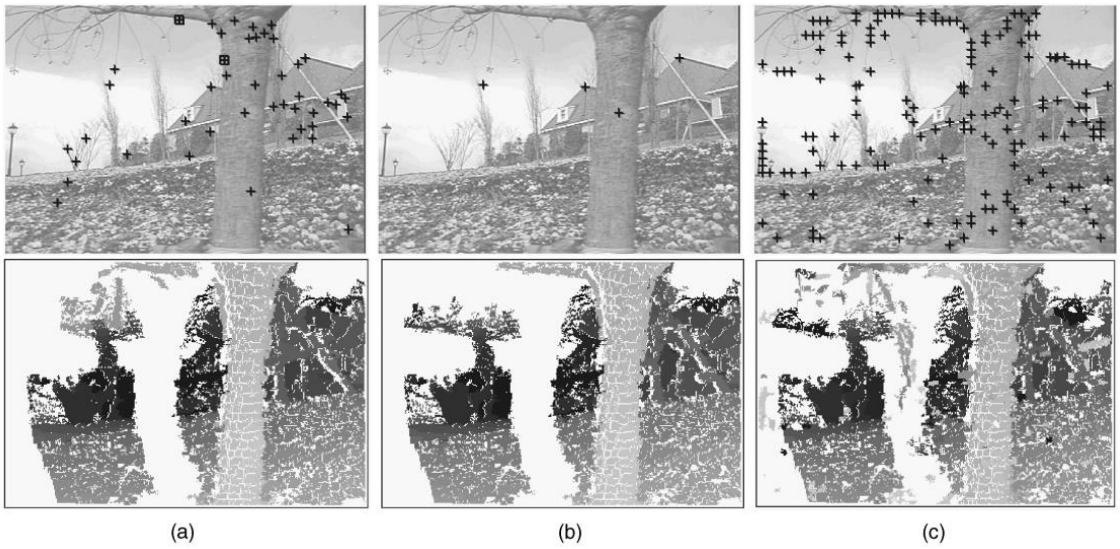


Figure 23. The disparity maps produced by propagation with different seed points and without the epipolar constraint. (a) Automatic seed points with the match outliers marked with a square instead of a cross. (b) Four seed points manually selected. (c) Four seed points manually selected plus 158 match outliers with a strong correlation score.

4.2 Quasi-Dense Matching based on Affine Transformation

Usually underwater images have large pair-wise baselines because of the difficult imaging conditions. In addition, due to the low image quality, not too many feature correspondences can be established in the sparse matching step. Hence present research is motivated by the properties and the success achieved on difficult image datasets of this traditional quasi-dense matching algorithm. However, since the brightness and color constancy does not hold for underwater imagery, the traditional intensity-based matching methods that compare the gray-scales of pixels from two small image windows (patches) cannot be used for underwater image matching as it usually can for the matching of in-air images. Straightforward application of this method leads to a high probability of identifying a wrong match as a correct one (false positive). Even with the epipolar constraint added at the propagation step, there are still a certain number

of incorrect matches in the set of final accepted matches. The reason for incorrect matches for underwater images is that the correlation score obtained from the intensity-based matching cannot be accepted as confidently as that for air imaging. However, the correlation measurement can still be an important indicator for similarity estimation between patches as long as the intensity-based matching method can be combined with other measurements to identify the correct image correspondences. Adaptive Least Square Matching (ALSM) incorporates adaptive geometric properties between the patches that are matched. This method utilizes an iterative optimization of the local geometric warping parameters between image patches such that the sum of absolute gray-scale differences [59] between corresponding pixels is minimized. In this paper, both the warping parameters and gray-scale similarity measurement contribute to the final decision about the goodness of the match. The "match-and-expand" procedure is used to implement the dense pixel-wise matching [56]. Our inspiration comes from [53], where the local geometric transformation is incorporated in a procedure for matching images with a wide baseline. Starting with a sparse set of feature matches (seed matches), the search for pixel-wise relationships is then iteratively expanded into the neighborhoods of seeds. In the process of searching for new matches, the ALSM technique is used instead of the traditionally used normalized cross-correlation. With ALSM, the correlation measurement is more meaningful. Simultaneously the optimal warping parameters are obtained and then compared with the parameters of the neighboring seed. For viewpoint changes, the most convenient transformation is the affine one [45]. Therefore, we choose affine transformation as the model used in adaptive lease square matching.

4.2.1 Adaptive Least Square Matching

The ALSM (Gruen, 1985) technique has been widely used in photogrammetry and computer vision since it was proposed in 1985. The process of the ideal pinhole camera imaging is in principle the projection from a Euclidean 3-dimensional coordinate system to a 2-dimensional

plane. Due to the change of viewpoints and properties of the medium, the patches from different images that correspond to the same object in a 3D scene differ both in shape and color (gray-scale levels of the contributing pixels). Traditionally, ALSM assumes that there is a geometric affine transformation and a radiometric linear gray-scale transformation between corresponding image pixels within the patches.

We denote the small corresponding image patches as discrete two-dimensional functions $f(x,y)$ and $g(x,y)$ respectively where $f(x,y)$ is the function that describes the patch in the reference image and $g(x,y)$ describes the patch in the target image. If the image location is represented by homogeneous vectors, the pixel mapping can be written in matrix form.

$$\begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} a_0 & a_1 & a_2 \\ b_0 & b_1 & b_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} \tag{4.1}$$

where a_0, a_1, a_2 and b_0, b_1, b_2 are the affine geometric parameters. (x_2, y_2) is the image coordinate in the patch of target image $g(x,y)$ and (x_1, y_1) is the image coordinate in the patch of the reference image $f(x,y)$.

The relationship of the gray-scales between corresponding pixels is assumed to be linear,

$$g(x_2, y_2) = h_0 + h_1 f(x_1, y_1) \tag{4.2}$$

where h_0 and h_1 are constant offset and gain. To summarize, each pixel in the patch of the target image is related to the pixel in the patch of the reference image through an affine transformation. Also, the corresponding pixels are linearly related by their gray-scales.

Linearizing (4.1) and (4.2) gives:

$$v = c_1 dh_0 + c_2 dh_1 + c_3 da_0 + c_4 da_1 + c_5 da_2 + c_6 db_1 + c_7 db_2 + c_8 db_3 - \Delta \tag{4.3}$$

where $c_1, \ldots, c_8$ are the partial derivatives of $g(x_2, y_2)$ with respect to the geometric affine and radiometric linear gray-scale parameters; $dh_0, dh_1, \ldots, db_3$ are the corrections of the parameters and Δ is the constant brightness difference.

Each pixel pair in the corresponding patches gives a linear equation in the form of equation

(4.3) and from all the pixel pairs between the image patches we obtain a set of linear equations which can be written in matrix form:

$$V = AX - L \tag{4.4}$$

Equation (4.4) is the observation equation where A is the design matrix and X is the correction vector of observables (geometric and radiometric parameters):

$$X = (dh_0 \ dh_1 \ da_1 \ da_2 \ da_3 \ db_0 \ db_1 \ db_2)^T \tag{4.5}$$

The brightness difference between the corresponding pixels in the matching patches is minimized by the iterative solution of the observation equation (4.4) and the updating of the correction vector X until the correction of each element in X becomes smaller than a predefined threshold value.

With a rough estimate of the locations of the corresponding image patches, the optimal affine and brightness transformation parameters can be obtained using ALSM. The center of the image patch is considered to be the final refined correspondence location. ALSM is applied to every eligible putative match around the seed match. The best ones (i.e., with the highest similarity score) are chosen as the final matches and added to the seed list for further propagation. However, if the initial guess of the patch location is far from the true location, the ALSM will not converge to a final solution within a certain number of iterations. In this case the putative match is rejected.

4.2.2 Quasi-Dense Matching using ALSM

After the first stage of obtaining an initial sparse set of seed matches, these seed matches are used to produce a quasi-dense correspondences between the two images.

In order to adjust the traditional quasi-dense matching for underwater imagery, our proposed approach incorporates the ALSM in the propagation process to minimize the number of mismatches. The modification is described below.

The first extension to the match propagation algorithm is applied at the first stage of the initial

matching. The initial set of feature matches is matched using ALSM in order to obtain the affine parameters for the transformation between the initial seed patches (image regions around x and x′). The affine parameters are adhered to the seed matches and are used later in the propagation step.

The original quasi-dense algorithm does not include the normalization of the local image patches and the search for new matches is performed in the non-altered neighborhood of x′. This prevents the original quasi-dense algorithm from the application on wide-baseline images because in such images viewpoints differ substantially and hence the deformation between two corresponding patches can be significant. In underwater imagery, viewpoints are almost always far from each other because it is difficult to control viewpoints due to currents and platform instability. Here the search for the putative matches is proposed not in the original neighborhood, but in the image patch warped by the affine transformation with the parameters found for the seed match (x, x').

Each putative match (u, u') from the local normalized image patch is first checked for the epipolar constraint and the 2D disparity limit as it is normally done in the original propagation algorithm. The location of the match that satisfies the above constraints is used as an initial estimate for a putative match to ALSM. The refined location for this putative match (u_r, u_r') is returned from ALSM together with its geometric affine and radiometric gray-scale linear parameters. Actually, u is equivalent to u_r since the center of the image patch in the reference image does not change during the ALSM computation. u_r', the correspondence of u_r in the target image, is obtained by iterating equation (4.4). Now the uniqueness of (u_r, u_r') should be checked to guarantee that the match has not been covered on the previous propagation steps. If the match (u_r, u_r') already exists in the final dense match list, it is rejected. The affine parameters of (u_r, u_r') are compared with those of the seed match (x, x'). The match (u_r, u_r') is accepted and added to the list of local candidates only if the differences

between the parameters are sufficiently small. The underlying assumption for this comparison of affine parameters is that the local observed surface is smooth, and hence the local geometric warping parameters for the two neighboring matches vary continuously. Likewise, the proposed algorithm proceeds until no seed match can be taken out from the seed list for propagation. Figure 24 illustrates the workflow of the proposed algorithm. Compared with the seeds in the original proposed algorithm, the seed matches in our approach have the geometric affine transformation parameters as an extra property which is used for normalization of local image patches and the surface smoothness check.

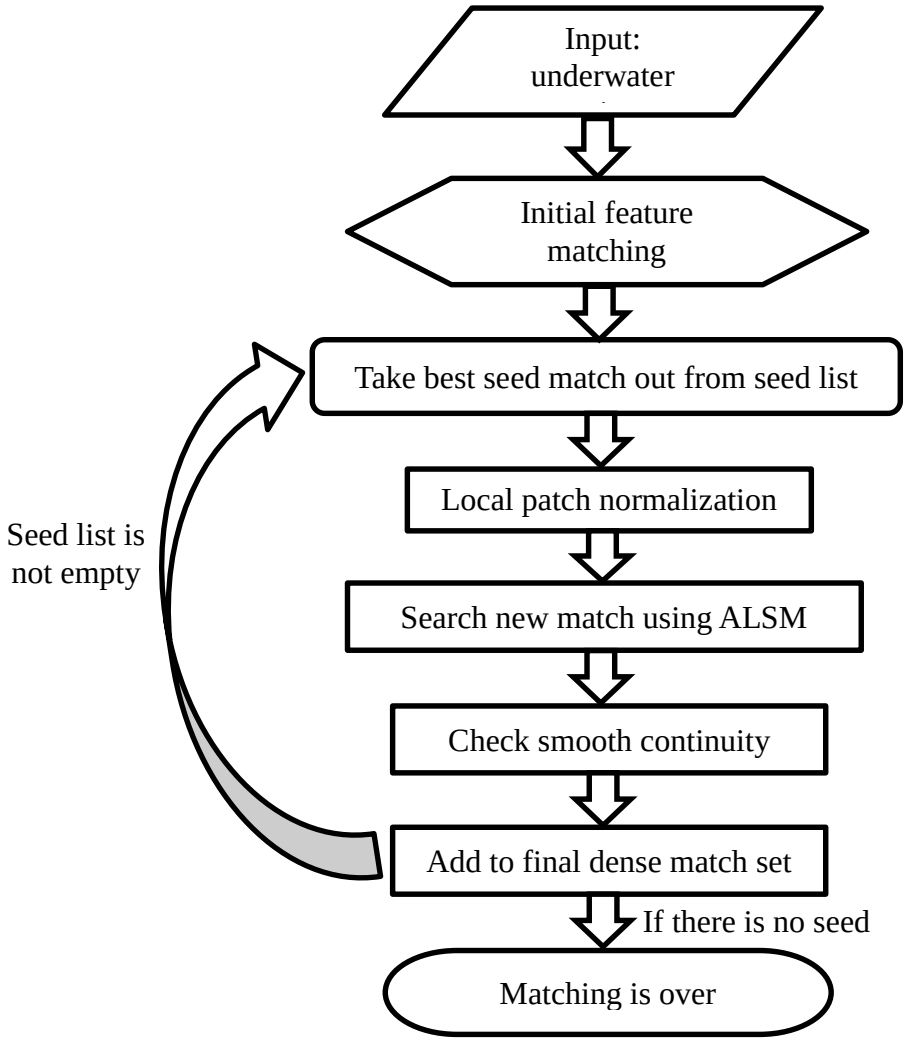


Figure 24. The workflow of the proposed algorithm.

Chapter V - 3D RECONSTRUCTION OF CAMERAS AND STRUCTURE

5.1 Image Preprocessing

The main reason that reconstructing the 3D surface from underwater imagery has always been challenging work is that the image quality is degraded due to light propagation through water. As mentioned before, various reasons can be found for the degradation, e.g non-uniform illumination of the imaged objects, light scattering and attenuation effects, or suspended particles present underwater result in image contamination. A major difficulty in the processing of underwater imagery comes from light attenuation which limits the visibility distance to approximately twenty meters in clear water and about five meters or less in turbid water. Thus use of artificial illumination is necessary for acquisition of underwater images. Unfortunately, artificial lights tend to illuminate non-uniformly producing a bright spot in the center of the image and poorly illuminated surroundings. However, the degraded underwater image can be preprocessed before the reconstruction process in order to counteract this effect to a certain degree [32]. Different methods have been proposed in order to enhance the underwater image quality. Usually methods consists of several successive steps which respectively correct non-uniform illumination, suppress noise, enhance contrast and adjust colors [33]. For 3D reconstruction purpose, the major drawback is that the degraded image quality limits the inadequacy of the feature points that can be extracted, which makes it difficult to determine image orientation and sequence connection. Even though the presented research is not intended to perform deep investigation of image processing techniques, simple image enhancement can dramatically increase the number of features and this could be crucial for 3D reconstruction work based on underwater images. Finding more feature correspondences results in the

enhancement of both the accuracy and the robustness of the recovery of the camera geometry. Figure 25 shows an original underwater image and the image after normalization enhancement.

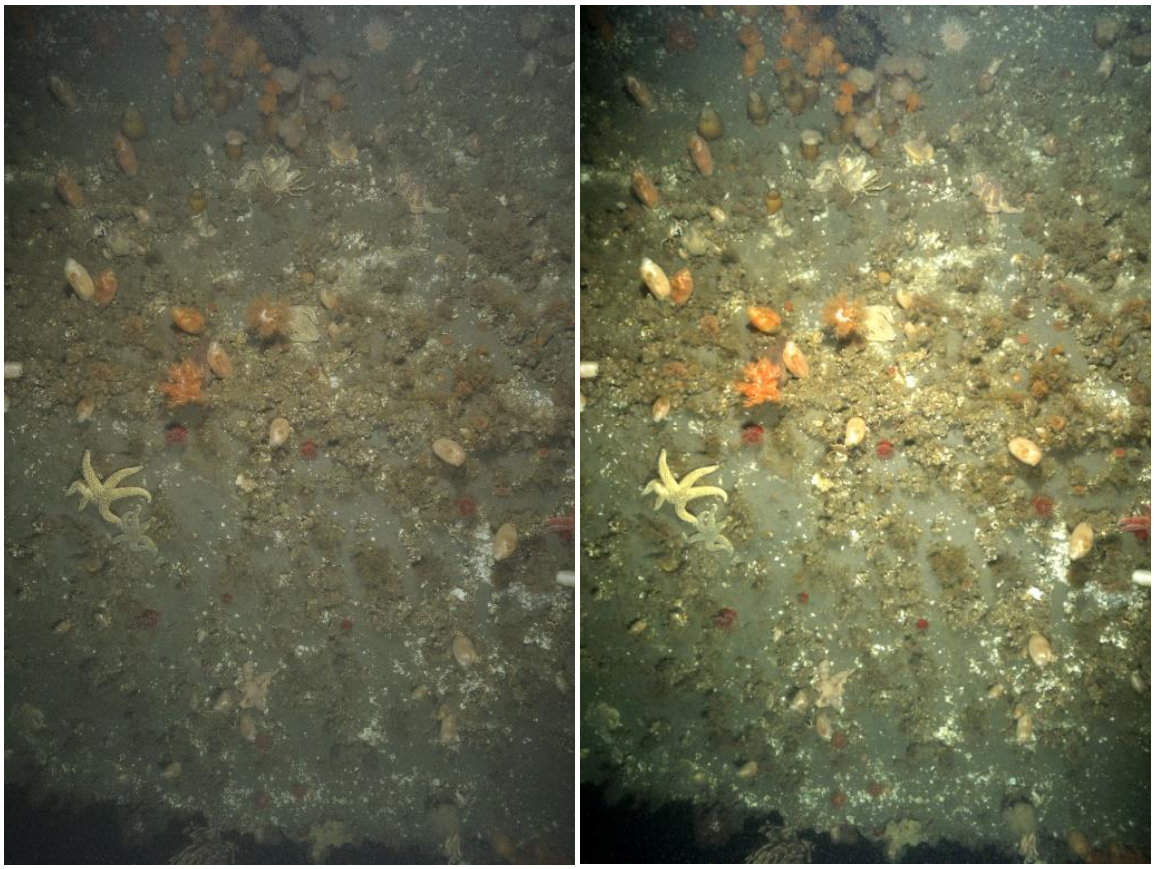


Figure 25. Image enhancement by normalization. Left: original image. Right: image after normalization.

We performed a test on a subset of the image dataset including 5 sequential images to obtain the statistics on the number of features extracted and matched before and after simple image enhancement. The result is shown in Figure 26. It can be seen that the number of features extracted and successfully matches increases dramatically after image normalization. While the image enhancement is not that important for images taken in air, it is quite meaningful for underwater images since few feature points can be successfully matched on the original images. It is highly recommended for researchers and engineers working on 3D reconstruction from underwater images to preprocess images in order to acquire better performance.

	Before Image Normalization			After Image Normalization		
	Number of features (left image)	Feature numbers(right image)	Number of Matches	Feature numbers (left image)	Feature numbers (right image)	Number of matches
Pair 1	2306	3134	136	7196	8891	225
Pair 2	3134	3198	195	8891	9461	331
Pair 3	3198	3880	171	9461	10454	311
Pair 4	3880	3682	72	10454	10323	107

Figure 26. The number of features and sparse matches on 4 sequential image pairs before and after image normalization.

5.2 Adding View from Image Sequence

Previous sections have introduced the geometry of two views. As is shown above, the fundamental matrix which represents the two-view geometry algebraically can be estimated by providing a set of sparse matches from two views up to a scale factor. The reconstruction of each pair of two views has their own scale. Since this work aims at reconstructing 3D models from an image sequence, the scale for all of the image projection matrices should be consistent with each other. The consistent scale for all pairs of images can be achieved by only one common point that can be seen in all of the three images. Let the common point be denoted as $\mathrm{X} = (\mathrm{X}, \mathrm{Y}, \mathrm{Z}, 1)^T$. Since the first image pair is chosen to be the reference, X can be obtained by triangulating the correspondence from the first image pair. The scale of the second image pair is adjusted to make X be exactly projected to p_3 in the third image. We will denote the scale ratio as s and s is only reflected on the length of the baseline as it is shown in Figure 27. According to the projection equation,

$$\mathrm{K}[\mathrm{R}'|\mathrm{st}']\mathrm{X} = p_3 \tag{5.1}$$

The undetermined ratio s is the only unknown and can be solved for directly. The ratio s is multiplied with the old translation vector t' of the second image pair to form the new

translation vector. After this scale adjustment, the third image is added and hence all the images are reconstructed in the same world coordinate system. If there are multiple common points that can be seen by all the three images, s has multiple solutions and can be averaged in order to increase the solution accuracy. The solution can be used as an initial value in sparse bundle adjustment.

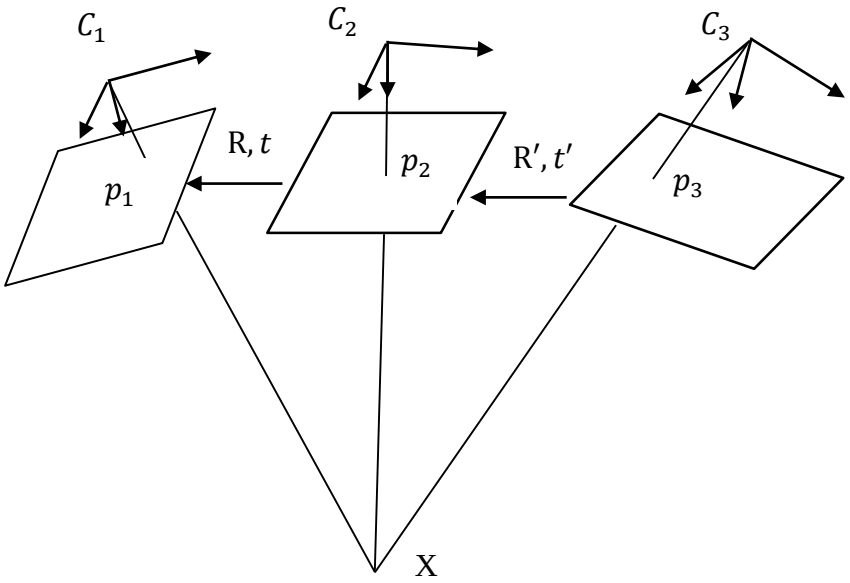


Figure 27. Add new image in the previous constructed image sequence.

5.3 Sparse Bundle Adjustment

In the previous sections of this thesis, the two view geometry and matching have been introduced. In section 5.2, it has been stated that by using a common world point that can be seen in three images, the new image from the sequence can be added into the same world coordinate system of the first two images. By adding one next image each time from the sequence, the whole image sequence can be oriented and reconstructed in the same coordinate frame. However, due to image noise and measurement error, the system error is accumulated and is not distributed normally in each image. Bundle adjustment aims at obtaining a

reconstruction which is optimal under certain assumptions regarding the noise pertaining to the observed image features. This is achieved by iteratively refining projection matrices and 3D points in order to minimize reprojection error. Bundle Adjustment (BA) is almost invariably used as the last step of every feature-based 3D reconstruction algorithm. Let us denote the set of 3D points by X_j and the set of project matrices by P^i. x_j^i is the coordinate of the j-th point as seen by the i-th camera. Due to the image noise, system and measurement error, the projection equation $x_j^i = P^i X_j$ will not be satisfied exactly. The sparse bundle adjustment strives to estimate projection matrices $\hat{P}^i$ and 3D points $\hat{X}_j$ which exactly project image points $\hat{x}_j^i$ as $\hat{x}_j^i = \hat{P}^i \hat{X}_j$, and simultaneously minimize the image distance between the reprojected point and the measured image points x_j^i for every view in which the 3D point appears i.e.

$$\min_{\hat{P}^i, \hat{X}_j} \sum_{ij} d(\hat{P}^i \hat{X}_j, x_j^i)^2 \tag{5.2}$$

Where $\mathrm{d}(\mathrm{x}, \mathrm{y})$ is the geometric distance between the homogeneous points x and y [13]. The popular sparse bundle adjustment package (SBA) [51] is used in our work. The reprojection error between the observed and predicted image points is to be minimized during BA. The minimization is achieved using a nonlinear least-squares algorithm. Levenberg-Marquartdt has been proven to be effective in nonlinear least-squares problems and this method is used in this sparse bundle adjustment package. In this work, bundle adjustment is performed each time a new image from the sequence is added. The reason for choosing the incremental bundle adjustment strategy is to guarantee that the bundle adjustment is able to converge in a small number of iterations. If the bundle adjustment is only carried out after all the images have been added, the error is accumulated for the images added later causing the initial value to be quite far away from the true value. In this case, the bundle adjustment is not guaranteed to converge.

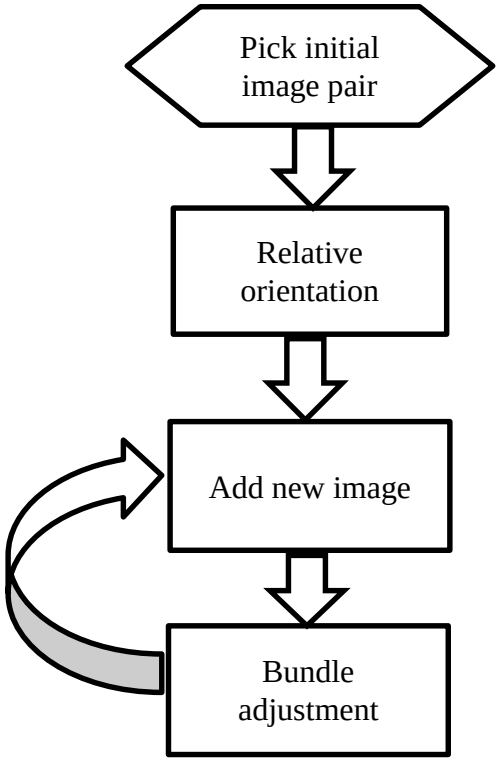


Figure 28. Incremental sparse bundle adjustment workflow.

Chapter VI - EXPERIMENT

The approach described above was tested on our underwater image dataset. The dimensions of the images are 712 pixels in width and 1072 pixels in height which are actually 4-times subsamples of the original images. Two different image pairs from the dataset were selected for the experiment. The first image pair covers a smooth surface of a sunk airplane and does not have dark (poorly textured) areas. The surface covered by the second image pair has depth discontinuities and large dark areas. The SIFT detector was used for the initial feature matching step and produced 304 feature matches and 155 feature matches respectively (Figure 29).

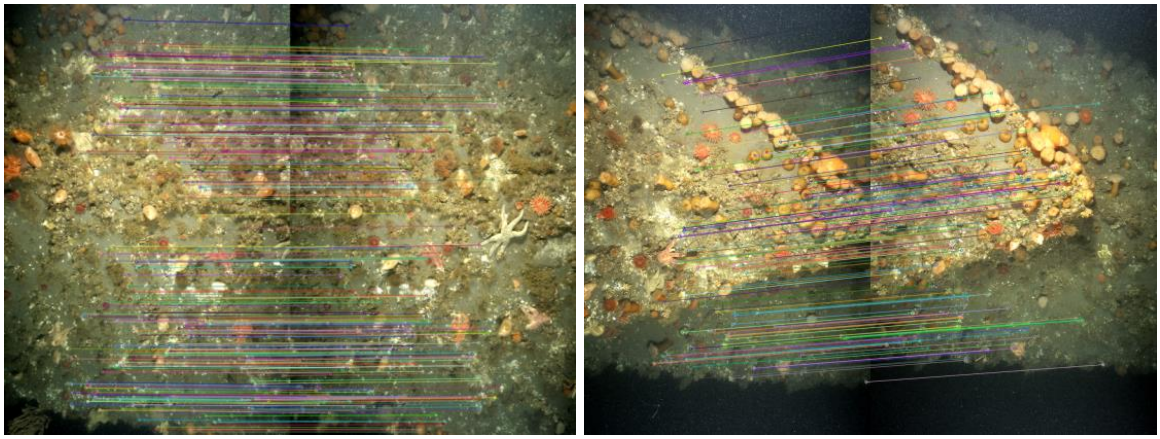


Figure 29. SIFT-detected matches for two underwater stereo image pairs.

The epipolar geometry of these image pairs has been recovered as described above. The quasi-dense matches obtained by the proposed approach of this thesis were triangulated to build point clouds and then surfaces were reconstructed from point clouds to test the performance of the approach. The generated point clouds from these two image pairs were also compared with the point clouds generated by CMVS andSGM. The results are shown in Figure 30.

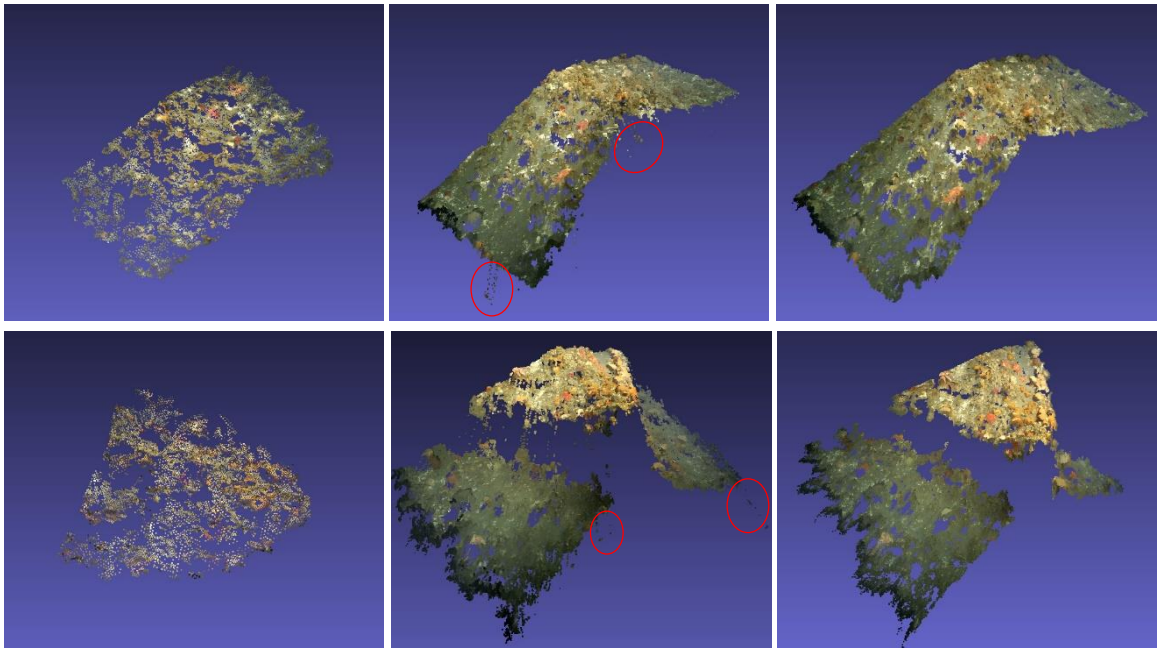


Figure 30. Comparison of the point clouds generated by three different approach. Top: the first experimental image pair; Bottom: the second experimental image pair. From left to right: point cloud generated by CMVS; point cloud generated by Semi-global Matching based on Mutual Information; point cloud generated by the proposed method of this thesis.

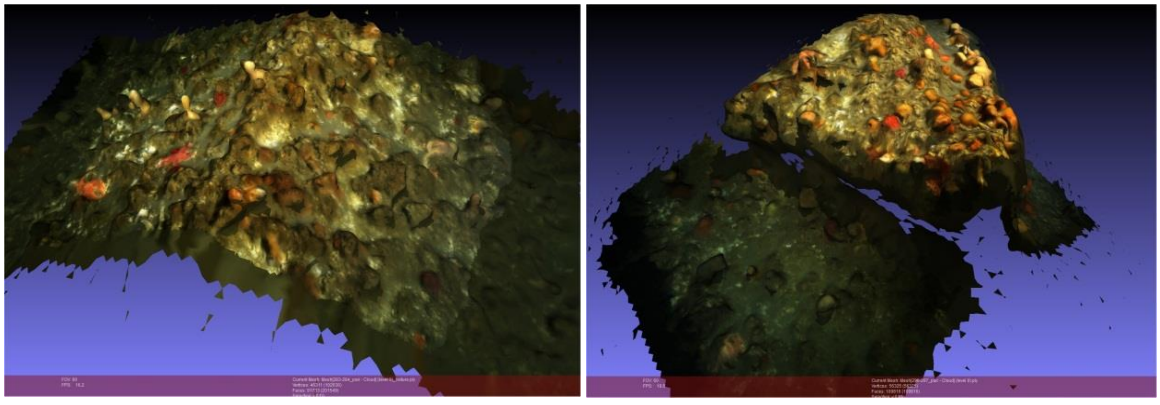
From the comparison illustrated by Figure 30, it can be seen that the performance of CMVS on dense matching of underwater images is the worst for underwater images. The density of points generated by CMVS is the sparsest and parts of the images are not matched at all. The second approach, semi-global matching, is efficient and gives a very satisfactory result with the largest point cloud density. However, due to the special property of underwater images – absence of color and grayscale, it is unavoidable that some outliers exist in the final result because this method uses multi-directional dynamic programming which is only based on the measurement of grayscales. These outliers generated by semi-global matching are red circled in Figure 30. The stereo matching method proposed in this thesis proved to be most suitable for underwater images since it combines both the measurement of image patch geometry and correlation to determine the best pixel correspondence. Although a pixel-to-pixel match is not satisfied most of the time, the density of the generated point cloud is still sufficient to reconstruct the 3D models with detailed surface structure. 391,016 quasi-dense matches were obtained from the

first pair of images and 284,338 from the second pair. The ratio of successfully matched pixels to all pixels is 0.51 and 0.37 respectively. For the second image pair, the pixel matching ratio is lower because of the existence of low contrast areas and significant depth discontinuities. (Note that the ratio of 1 cannot be achieved in principle because the images overlap only partially.) The number of points that have been generated by these three different methods are given in Figure 31. Even compared with the fully pixel-to-pixel dense matching algorithm, semi-global matching, our proposed method still gives a reasonably dense result.

	Number of points for pair 1	Number of points for pair 2
CMVS	13093	7356
SGM	483965	347676
QD_ALSM	391016	284338

Figure 31. The number of points that are matched and triangulated by three different methods.

The point clouds from this proposed method are reconstructed into surface meshes and the image texture is draped on these two meshes based on the camera's orientation parameters. The textured meshes are shown in Figure 32.



.

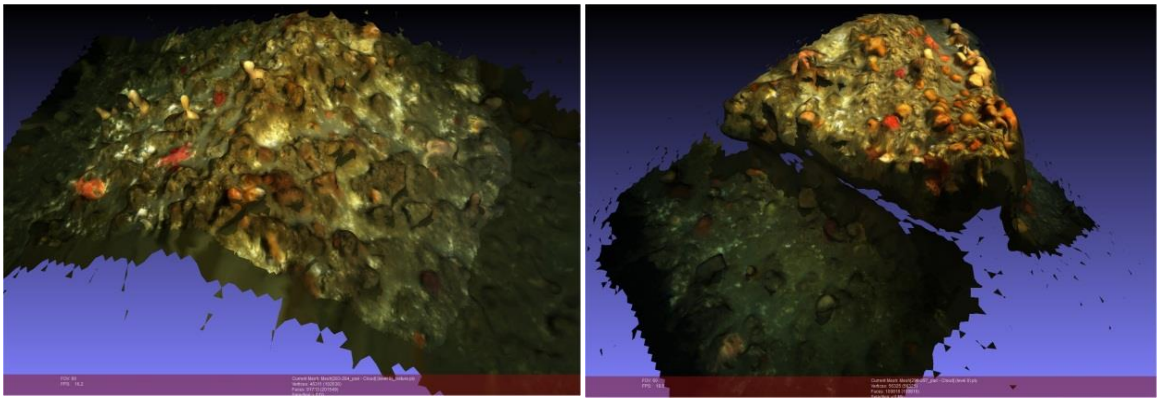


Figure 32. Reconstructed surfaces from the point clouds with image texture draped over the 3D surface. (The surface is constructed using the Poisson surface reconstruction algorithm [43] provided by MeshLab [37].)

From the constructed point clouds and textured surfaces of these two underwater image pairs, it follows that our proposed algorithm is robust enough to provide a sufficient number of image correspondences for the representative reconstruction of the 3D scene. Figure 33 and Figure 34 are the complete reconstruction results from the whole image sequence of our dataset with two different perspectives.

Figure 33. Full point cloud from all the images of top view sequence after bundle adjustment (Perspective 1).

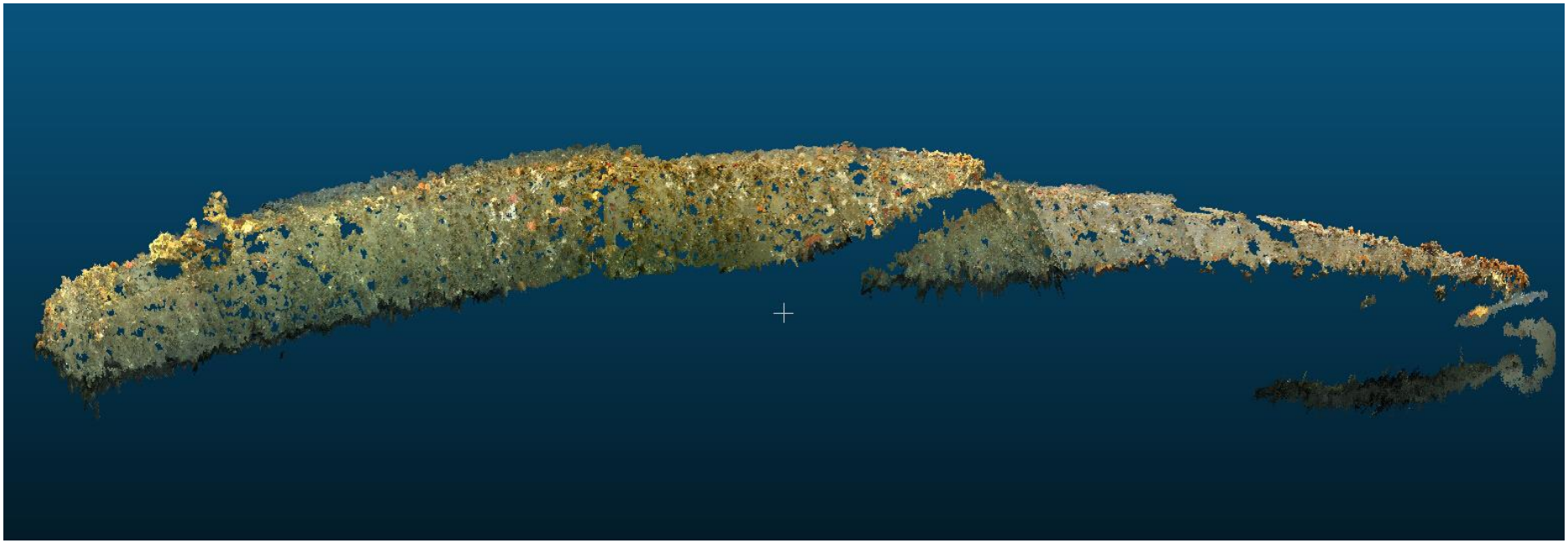


Figure 34. Full point cloud from all the images of top view sequence after bundle adjustment (Perspective 2).

Chapter VII - CONCLUSION AND FUTURE WORK

In this thesis, a quasi-dense matching algorithm, which is sufficiently robust for the application to underwater imagery, has been reported. The proposed algorithm works differently from existing methods in several ways to be better adapted to the underwater environment. The ALSM technique is incorporated in the process of the match propagation. In addition, the local geometric warping parameters for each putative match are also compared with those of the neighboring seed match to choose the best dense matches. The seed match is augmented by an array of affine transformation parameters which is returned from the ALSM and the match expansion procedure is carried out in the normalized (warped) neighborhood of the seed.

The experiments on our underwater image dataset demonstrated that this algorithm is robust to the change of environment factors (in comparison with air) and is able to provide a sufficient number of successful correspondences between the overlapping images. Even compared with state-of-art dense matching approaches from computer vision, such as CMVS and semi-global matching, this method produces a competitive result. Even with the camera calibrated in a tank rather than on site, the reconstruction result looks satisfactory. The algorithm described in this thesis can be useful in a wide range of applications such as underwater image modeling and rendering. The contribution of our work is the adaptions of existing 3D reconstruction methods to accommodate for specifics of underwater environment.

Since the initial propagation is based on the initial seeds, the distribution and quantity of the seed matches is important to the performance of this propagation based algorithm. In the experiment period, it has been found that many underwater stereo images are not capable of obtaining well-distributed feature matches with most of the matches concentrated in the well-illuminated areas. In addition, the number of initial feature matches is not sufficient either,

which reduces the robustness and accuracy of the estimation of epipolar geometry. The reasons for these phenomenon are low texture, low overlap, and non-uniform illumination. While it is possible to improve the illumination condition to acquire image of better quality and to take more images in order to increase overlap, from the algorithm perspective, better and improved robust feature detectors are required for underwater image matching. The quantity and distribution of sparse feature matches will not only improve the performance of this proposed algorithm, but also increase the image network robustness and connectivity, which is helpful for the underwater 3D reconstruction.

As we can see, the major drawback of this proposed method is the high computational cost. For each putative match in the propagation step around a seed, their geometric warping parameters are computed iteratively which are CPU-intensive and time consuming. A promising improvement can be made to increase the computation speed, which is to divide image into different blocks and implement the quasi-dense matching algorithm in a multi-threaded environment, since each block has its own seed matches and they can propagate simultaneously without any conflicts with each other.

LIST OF REFERENCES

[1] Hu, S., Qiao, J., Zhang, A., and Huang, Q. 2008. 3d reconstruction from image sequence taken with a handheld camera, in *International Society for Photogrammetry and Remote Sensing, Congress Beijing*, 559-562.

[2] Koch, R., Evers-Senne, J. F., Frahm, J. M., and Koeser, K. 2005. 3d reconstruction and rendering from image sequences, in *Proceedings of WIAMIS* (April, 2005).

[3] Snavely, N., Seitz, S. M., and Szeliski, R. 2008. Modeling the world from internet photo collections. *International Journal of Computer Vision. 80*, 2, 189-210.

[4] Brown, M., and Lowe, D. G. 2005. Unsupervised 3D object recognition and reconstruction in unordered datasets, in *Proceedings of the Fifth International Conference on 3-D Digitial Imaging and Modeling* (2005) IEEE, 56-63.

[5] Hirschmuller, H. 2008. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence. 30*, 2, 328-341.

[6] Prabhakar, C. J., and Kumar, P. U. 2012. 3D Surface Reconstruction of Underwater Objects. *arXiv preprint arXiv: 1211.2082.*

[7] Huang, C., and Zhou, W. 2013. An Underwater Image Matching Approach on Improved SURF. *Journal of Computational Information Systems. 9*, 8, 3019-3025.

[8] Sedlazeck, A., Koser, K., and Koch, R. 2009. 3d reconstruction based on underwater video from rov kiel 6000 considring underwater imaging conditions, in *Proceedings of OCEANS 2009-EUROPE*, 1-10.

[9] Pollefeys, M., Koch, R., Vergauwen, M., and Van Gool, L. 1998. Metric 3D surface reconstruction from uncalibrated image sequences. In *3D Structure from Multiple Images of Large-Scale Environments*. Springer, 139-154.

[10] Scharstein, D., and Szeliski, R. 2002. A taxonomy and evaluation of dense two-frames stereo correspondence algorithms. *International Journal of Computer Vision. 47*, 1-3, 7-42.

[11] Lee, Y., Toh, K. A., and Lee, S. 2008. Stereo image rectification based on polar transformation. *Optical Engineering. 47*, 8, 087205-087205.

[12] Megyesi, Z. 2009. Dense Matching Methods for 3D Scene Reconstruction from Wide Baseline Images. Ph.D. dissertation. Eotvos Lorand University, France.

[13] Hartley, R., and Zisserman, A. 2003. Multiple view geometry in computer vision. Cambridge university press.

[14] Luong, Q. T., and Faugeras, O. D. 1996. The fundamental matrix: Theory, algorithms, and stability analysis. *International journal of computer vision. 17*, 1, 43-75.

[15] Zhang, Z., Deriche, R., Faugeras, O., andLuong, Q. T. 1995. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence. 78*, 1, 87-119.

[16] Huang, J. F., Lai, S. H., and Cheng, C. M. 2007. Robust fundamental matrix estimation with accurate outlier detection. *Journal of information science and engineering. 23*, 4, 1213-1225.

[17] Fischler, M. A., and 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.

[18] Morel, J. M., and Yu, G. 2009. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences. 2*, 2, 438-469.

[19] Mikolajczyk, K., and Schmid, C. 2004. Scale & affine invariant interest point detectors. *International journal of computer vision. 60*, 1, 63-86.

[20] Harris, C., and Stephens, M. 1988. A combined corner and edge detector, in *Proceedings of the Alvey vision conference* (August 1988), 15, 50.

[21] Bay, H., Tuytelaars, T., and Van Gool, L. 2006. SURF: Speeded up robust features, in *Proceedings of Computer vision-ECCV 2006* (Graz, Austria, May 2006) Springer press, 404-417.

[22] Aclantarilla, P. F., Bartoli, A., and Davison, A. J. 2012. KAZE features, in *Proceedings of Computer Vision-ECCV 2012* (Firenze, Italy, October 2012) Springer press, 214-227.

[23] Rosten, E., and Drummond, T. 2005. Fusing points and lines for high performance tracking, in *Proceedings of the Tenth IEE International Conference on Computer Vision, 2005, ICCV 2005* (Beijing, China, October 2005) IEEE press, 1508-1515.

[24] Rosten, E., and Drummond, T. 2006. Machine learning for high-speed corner detection, in *Proceedings of Computer Vision-ECCV 2006* (Graz, Austria, May 2006) Springer press, 430-443.

[25] Leutenegger, S., Chli, M., and Siegwart, R. Y. 2011. BRISK: Binary robust invariant scalable keypoints, in *Proceedings of 2011 IEEE International Conference on Computer Vision (ICCV)* (Barcelona, Spain, November 2011) IEEE press, 2548-2555.

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

[27] Calonder, M., Lepetit, V., Strecha, C., and Fua, P. 2010. Brief: Binary robust independent elementary features, in *Proceedings of Computer Vision-ECCV 2010* (Crete, Greece, September 2010) Springer press, 778-792.

[28] Beall, C., Lawrence, B. J., lla, V., and Dellaert, F. 2010. 3D reconstruction of underwater structures, in *Proceedings of 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Taipei, Taiwan, October 2010) IEEE press, 4418-4423.

[29] Negahdaripour, S., and Madjidi, H. 2003. Stereovision imaging on submersible platforms for 3-d mapping of benthic habitats and sea-floor structures. *IEEE Journal of Oceanic Engineering. 28*, 4, 625-650.

[30] Williams, S. B., Pizarro, O., Johnson-Roberson, M., Mahon, I., Webster, J., Beaman, R., and Bridge, T. 2008. Auv-assisted surveying of relic reef sites, in *Proceedings of OCEANS 2008* (Quebec, Canada, September 2008) IEEE press, 1-7.

[31] Nicosevici, T., and Garcia, R. 2008. Online robust 3D mapping using structure from motion cues, in *Proceedings of OCEANS 2008-MTS/IEEE Kobe Techno-Ocean* (Kobe, Japan, April 2008) IEEE press, 1-7.

[32] Prabhakar, C. J., and Kumar, P. U. 2012. 3D surface reconstruction of underwater objects. arXiv preprint arXiv:1211.2802.

[33] Bazeille, S., Quidu, I., Jaulin, L., and Malkasse, J. P. 2006. Automatic underwater image pre-processing, in CMM'06, xx.

[34] Brandou, V., Allais, A. G., Perrier, M., Malis, E., Rives, P., Sarrazin, J., and Sarradin, P. M. 2007. 3D reconstruction of natural underwater scenes using the stereovision system IRIS, in *Proceedings of OCEANS 2007-Europe* (Aberdeen, Scotland, June 2007), 1-7.

[35] Meline, A., Triboulet, J., and Jouvencel, B. 2012. Comparative study of two 3D

reconstruction methods for underwater archaeology, in *Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Algarve, Portugal, October, 2012), 740-745.

[36] Allais, A-G., Brandou, V., Dentrecolas, S., Gilliotte, J-P., Perrier, M. IRIS – A Vision System to Reconstruct Natural Deep-sea Scenes in 3D, in Proceedings of the Seventeenth International Offshore and Polar Engineering Conference (Lisbon, Portugal, July 2007), 1-6.

[37] Visual Computing Lab ISTI – CNR. Meshlab. http://meshlab.sourceforget.net/

[38] Timothy, H. Y. Y. 2014. Underwater Camera Calibration and 3D Reconstruction. Master thesis. Department of Computing Science, University of Alberta, Edmonton, Canada.

[39] Furukawa, Y., and Ponce, J. 2010. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on Pattern Analysis and Machine Intelligence. 32*, 8, 1362-1376.

[40] Wilby, A., Wu, G., and Naughton, P. Underwater Stereo Vision and 3D Reconstruction.

[41] Cavan, N. 2011. Reconstruction of 3d points from uncalibrated underwater video.

[42] Chum, O., Werner, T., and Matas, J. 2004. Epipolar geometry estimation via ransac benefits from the oriented epipolar constraint. IEEE.

[43] Kazhdan, M., Bolitho, M., and Hoppe, H. 2006. Poisson surface reconstruction, in *Proceedings of the fourth Eurographics symposium on Geometry processing (vol. 7)* (Sardinia, Italy, June 2006).

[44] Moravec, H. P. 1980. Obstacle avoidance and navigation in the real world by a seeing robot rover. Ph.D. dissertation. Department of Computer Science, Stanford University, Stanford, USA.

[45] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., and Van Gool, L. 2005. A comparison of affine region detectors. *International Journal of Computer Vision. 65*, 1-2, 43-72.

[46] Hogue, A., German, A., Zacher, J., and Jenkin, M. 2006. Underwater 3d mapping: Experiences and lessons learned, in *Proceedings of the 3rd Canadian Conference on Computer and Robot Vision, 2006* (Quebec City, Canada, May 2006), 24-24.

[47] Bouguet, J. Y. 2013. Camera Calibration Toolbox for Matlab. http://www.vision.caltech.edu/bouguetj/calib_doc

[48] Loop, C., and Zhang, Z. 1999. Computing Rectifying Homographies for Stereo Vision, in *Proceedings of IEEE Computer Vision Society Conference on Computer Vision and Patter Recognition (Vol. 1)* (Fort Collins, USA, June 1999).

[49] Bryant, M., Wettergreen, D., Abdallah, S., and Zelinsky, A. 2000. Robust camera calibration for an autonomous underwater vehicle, in *Proceedings of Australian Conference on Robotics and Autom* (Melbourne, Australia, September 2000), 111-116.

[50] Hirschmuller, H. 2005. Accurate and efficient stereo processing by semi-global matching and mutual information, in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (Vol. 2)*(San Diego, USA, June 2005), 807-814.

[51] Lourakis, M. I., and Argyros, A. A. 2009. SBA: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software (TOMS). 36*, 1, 2.

[52] Rothermel, M., Wenzel, K., Fritsch, D. and Haala, N. 2012. SURE: Photogrammetric surface reconstruction from imagery, in *Proceedings LC3D Workshop* (Berlin, Germany, December 2012).

[53] Kannala, J., and Brandt, S. S. 2007. Quasi-dense wide baseline matching using match propagation, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007 (CVPR'07)* (Minneapolis, USA, June 2007) IEEE, 1-8.

[54] Wu, C. 2011. VisualSFM: A Visual Structure from Motion System. http://ccwu.me/vsfm.

[55] Wu, C., Agarwal, S., Curless, B., and Seitz, S. M. 2011. Multicore bundle adjustment, in *Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Colorado Springs, USA, June 2011) IEEE, 3057-3064.

[56] Lhuillier, M., and Quan, L. 2002. Match propagation for image-based modeling and rendering. *IEEE Transactions on Pattern Analysis and Machine Intelligence. 24*, 8, 1140-1146.

[57] Lhuillier, M. 1998. Efficient dense matching for textured scenes using region growing, in *Proceedings of the Ninth British Machine Vision Conference* (Southampton, UK, September 1998), 14-17.

[58] Megyesi, Z., and Chetverikov, D. 2004. Affine propagation for surface reconstruction in wide baseline stereo, in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR) 2004 (Vol. 4)* (Cambridge, UK, August 2004) IEEE, 76-79.

[59] Srikham, M., Pluempitiwiriyawej, C., and Chanwimaluang, T. 2010. Comparison of dense matching algorithms in noisy image, in *Proceedings of the Second International Conference*

on Digital Image Processing (Singapore, February 2010) International Society for Optics and Photonics, 75461V-75461V.

[60] Koskenkorva, P., Kannala, J., and Brandt, S. S. 2010. Quasi-dense wide baseline matching for three views, in *Proceedings of the 20th International Conference on Pattern Recognition* (Istanbul, Turkey, August 2010) IEEE, 806-809.

[61] Khropov, A., Konushin, A. 2006. Guided Quasi-Dense Tracking for 3D Reconstruction, in *Proceedings of International Conference Graphicon* (Novosibirsk Akademgorodok, Russia, July 2006), 1-5.

[62] Longuet-Higgins, H. C. 1987. A computer algorithm for reconstructing a scene from two projections. *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, MA Fischler and O. Firschein, eds.* 61-62.