석사 학위논문
Master's Thesis

양안 기반의 굴절 매질을 이용한 스테레오 합성

Stereo Fusion using a Refractive Medium on a Binocular Base

백 승 환 (白 承 煥 Baek, Seung-Hwan)
전산학과
Department of Computer Science

KAIST

2015

양안 기반의 굴절 매질을 이용한 스테레오 합성

Stereo Fusion using a Refractive Medium on a Binocular Base

Stereo Fusion using a Refractive Medium on a Binocular Base

Advisor : Professor Kim, Min Hyuk

by

Baek, Seung-Hwan
Department of Computer Science
KAIST

A thesis submitted to the faculty of KAIST in partial fulfillment of the requirements for the degree of Master of Science in the Department of Computer Science . The study was conducted in accordance with Code of Research Ethics[1].

2014. 12. 15.

Approved by

Professor Kim, Min Hyuk

[Advisor]

[1]Declaration of Ethical Conduct in Research: I, as a graduate student of KAIST, hereby declare that I have not committed any acts that may damage the credibility of my research. These include, but are not limited to: falsification, thesis written by someone else, distortion of research findings or plagiarism. I affirm that my thesis contains honest conclusions based on my own careful research under the guidance of my thesis advisor.

양안 기반의 굴절 매질을 이용한 스테레오 합성

백 승 환

위 논문은 한국과학기술원 석사학위논문으로
학위논문심사위원회에서 심사 통과하였음.

2014년 12월 15일

심사위원장 김 민 혁 (인)

심사위원 윤 성 의 (인)

심사위원 최 호 진 (인)

MCS

20133331

백 승 환. Baek, Seung-Hwan. Stereo Fusion using a Refractive Medium on a Binocular Base. 양안 기반의 굴절 매질을 이용한 스테레오 합성. Department of Computer Science . 2015. 39p. Advisor Prof. Kim, Min Hyuk. Text in English.

ABSTRACT

The performance of depth reconstruction in binocular stereo relies on how adequate the predefined baseline for a target scene is. Wide-baseline stereo is capable of discriminating depth better than the narrow one, but it often suffers from spatial artifacts. Narrow-baseline stereo can provide a more elaborate depth map with less artifacts, while its depth resolution tends to be biased or coarse due to the short disparity. In this thesis, we propose a novel optical design of heterogeneous stereo fusion on a binocular imaging system with a refractive medium, where the binocular stereo part operates as wide-baseline stereo; the refractive stereo module works as narrow-baseline stereo. We then introduce a stereo fusion workflow that combines the refractive and binocular stereo algorithms to estimate fine depth information through this fusion design. The quantitative and qualitative results validate the performance of our stereo fusion system in measuring depth, compared with homogeneous stereo approaches.

Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Motivation

Camera was invented to capture the real world as an image which is the two-dimensional projection of lights passing through the optical pipeline of the camera. The 2D imaging technology significantly influences daily life of people with emerging of mobile cameras. However, when we capture a scene with a traditional camera, we miss depth information which is the one dimension consisting of the three dimensional space in which we live. As a consequence, we face many challenges in understanding and analyzing the captured scene. In some cases such as collision detection and manufacturing, distance itself could be a key parameter, which is the purpose of the applications. Also object detection and recognition could be solved in a more effective and efficient way if the depth information is given. Since the benefits of depth data are significant, it leads many researchers to dive in various approaches for estimating depth.

One of the technologies to acquire depth is to employ binocular parallax, which is the difference in the image positions of an object on different view points. Our human visual system also makes use of this binocular cue for understanding a scene using our two eyes. As the principle of this cue is concise and simple, many researchers have been working on this binocular stereo. On the other hand, refractive stereo is another way to extract depth information. A refractive stereo system consists of a camera with a refractive medium in front of the camera in general. The system enables us to capture the scene with intentional distortion. We can recover depth information by estimating the displacement between the corresponding pixel positions of an object on the two images captured with and without the refractive medium. This procedure is similar to the binocular stereo, in which the corresponding pixel needs to be searched on two different images.

Refractive stereo enables us to estimate depth with a single camera, using an additional optical element (refractive medium). However, the displacement occurred by the refraction on the medium is shorter than that of binocular system in general. To this end, it restricts the depth resolution of the system which is an important factor for some applications.

1.2 Scope

There are many existing binocular 3D sensors and corresponding depth estimation methods. However, it is still challenging to obtain a depth map with high depth resolution and less artifacts. In this thesis, we

propose a fusion system of refractive stereo and binocular stereo in order to obatin high quality depth estimates in terms of both depth artifacts and depth resolution. Also we present a fusion workflow for the proposed system.

We take inspiration from refractive stereo to combine these two heterogeneous stereo systems, where a stereo fusion system is designed with a refractive medium placed on one of the binocular stereo cameras. Multiple images are captured on the camera equipped with the refractive medium by placing the medium in different poses. Also we capture an image from the other camera. Using the captured images, we estimate a depth map with high depth resolution and less spatial artifacts using our proposed workflow.

We pre-calibrate the parameters for the refractive medium, and the geometric and the radiometric parameters of our fusion system in order to estimate depth information. Parameters of the medium are the thickness, the refractive index and the pose of the medium with respect to the camera. Geometric calibration is the relation between the poses of the two cameras, which is necessary for matching two corresponding points on the stereo images. The refraction effect on the medium alters the spectral power distribution of the lights passing through the medium. Therefore, color correction of refracted images is necessary to match the refracted images with the image captured from the other camera without the medium.

This proposed stereo fusion workflow takes the advantages of both refractive and binocular stereo, producing depth estimates with less artifacts and a high depth resolution. Our system can be easily adopted to existing binocular systems by placing a refractive module in front of one camera. We demonstrate the effectiveness of our method in acquiring a high depth resolution and less spatial artifacts in this thesis.

1.3 Contributions

The contributions of this thesis are three-fold.

- **A stereo fusion system that combines refractive and binocular system.** We design a stereo system on a binocular base with a refractive medium. The medium is placed in front of a camera, and the refracted images are captured by rotating the medium.
- **Calibrations for the proposed fusion system.** We develop several calibration methods for our fusion system including radiometric, geometric and refractive calibration.
- **Depth estimation workflow that combines heterogenous stereo images.** After obtaining the necessary parameters for depth estimation in the calibration step, we estimate depth information with our proposed workflow. The resulting depth map has a high depth resolution with less artifacts.

Most of these contributions have been presented in the following publication:

- **Seung-Hwan Baek** and Min H. Kim, *Stereo Fusion using a Refractive Medium on a Binocular Base*, *Proc. Asian Conference on Computer Vision*, November., 2014. (oral presentation)

1.4 Thesis Outline

This thesis is organized as follows. Chapter 2 presents background knowledge about binocular stereo and refractive stereo. Chapter 3 explains previous work related to this thesis. Chapter 4 describes the hardware design and calibration methods (geometric, refractive, color) for our setup. Chapter 5 presents a depth estimation method for refractive stereo and a stereo fusion approach combining the two different designs: refractive and binocular. Experimental evaluations of this proposed method are presented in Chapter 6, including quantitative and qualitative results. Chapter 7 summarizes this thesis and discusses the limitations of the proposed system with possible modifications to resolve the problems. Chapter 8 concludes this thesis.

Chapter 2. Background

This chapter describes background knowledge about binocular stereo and refractive stereo to help the understanding of the thesis.

2.1 Binocular Stereo

Binocular stereo utilizes two cameras with a specific displacement between the cameras. Two images captured by the cameras contain a pair of corresponding pixels projected from a surface. Binocular disparity describes pixel-wise displacement of parallax between the corresponding points on a pair of stereo images.

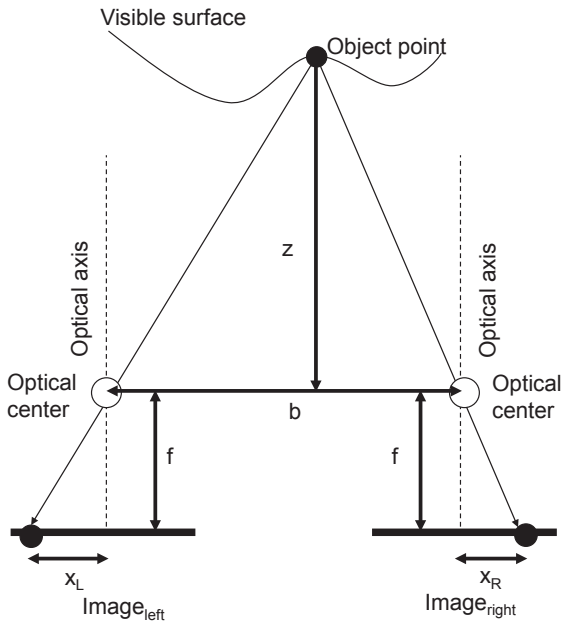


Figure 2.1: *For an object point, pixel positions on two image planes are different depending on the depth of the point, the focal length of the cameras and the baseline of the system.*

The disparity of a pixel is related to the depth of a surface corresponding to the pixel. Fig. 2.1 shows a schematic diagram of a binocular system. The rays from an object point on a visible surface are projected into two cameras passing through the optical centers of the cameras. The pixel-wise displacement between the corresponding pixels is called as *disparity* for the object point:

$$d = |X_L - X_R| \,, \tag{2.1}$$

where X_L and X_R are the projected points on the left image plane and the right image plane.

We can recover the depth z using simple trigonometry as:

$$z = \frac{fb}{d}, \tag{2.2}$$

where f is the focal length of the camera lens; b is the distance between the cameras, so called baseline and d is the disparity of the pair of corresponding pixels. In order to obtain a depth map from stereo images, we first estimate a disparity map, and then the depth map can be easily computed if we know the other two parameters (the focal length and the baseline) following Eq. (2.2).

As we mentioned, disparity is the displacement of the corresponding pixels projected from an object. Therefore, computing the disparity is accompanied with searching the corresponding points (correspondence problem). With an ideal alignment of a binocular stereo system, the corresponding points lies on a same row with a different column. This condition is called an *epipolar constraint*, and we can solve correspondence problem efficiently by restricting the search range of correspondence into one dimension when the epipolar constraint holds. The details of the constraint is explained in Sec. 2.3.

Disparity estimation of binocular stereo usually assumes that epipolar constraint holds for the stereo input images, and the method for estimating disparity consists of following four steps [26]:

1. Matching cost computation,
2. Cost aggregation,
3. Disparity computation, and
4. Disparity refinement.

The problem of estimating the disparity of a pixel on a left image is exactly equivalent to the correspondence search, which is to find the corresponding pixel on a right image. Note that it is also possible to select a pixel on the right image, and then find the corresponding point on the left image. As the corresponding point of a pixel lies on a same row due to the epipolar constraint, we can estimate the corresponding pixel using exhaustive search efficiently. For given disparity candidates, we directly compute the pixel positions of the corresponding point for a given original pixel. The remaining task is to select a best disparity of which the corresponding pixel candidate is most plausible.

To see the plausibility of the candidates, the color of each corresponding pixel is compared to the color of the original pixel. However, in this procedure, we implicitly assume that all surfaces on a scene follow *Lambertian reflectance*, which is a surface property that the color of the reflected ray on the surface is the same regardless of the viewing directions. It is the reason why stereo algorithms cannot recover depth of surfaces with non-Lambertian reflectance such as specular reflection. In this thesis, we assume

that the Lambertian reflectance assumption holds for our target scenes. Now, since the rays from a same object point have same color under the Lambertian surface assumption, we set the correspondence metric as a color difference between the corresponding pixel and the original pixel. The results of the pixelwise correspondence matching are called matching costs. As we mentioned, matching costs are computed for all target disparity candidates per a pixel.

The matching costs itself could contain many errors caused by noises and pixel errors. In order to build more robust matching costs, a cost aggregation technique is usually employed. In essence, cost aggregation is a filtering process for the matching costs.Underlying idea of cost aggregation is that the matching cost of a pixel should be similar to that of the neighbour pixel having similar color. Cost aggregation removes errors while keeping the consistency of the matching costs with the RGB image and imposing spatial consistency on the matching costs. There have been various cost aggregation methods, and the methods can be classified into two groups broadly: local cost aggregation and non-local cost aggregation. The difference between the two approaches is a window size of which inner pixels are considered as neighbour pixels enforcing the similarity of matching costs. Non-local cost aggregation enforces the similarity between the all pixels on an image while local aggregation only utilizes neighbour pixels on an window of which size is much smaller than that of an image.

We should select an optimal disparity for a pixel on the image using the aggregated costs, and this process is called as disparity computation. The methods of disparity computation can be also categorized into two groups: local approaches and global approaches. The local approaches select the optimal disparity, which minimizes or maximizes the aggregated costs for a pixel. The other approaches (global) estimates the best disparity considering the aggregated costs of the pixel as well as the aggregated costs of the neighbor pixels. In general, local methods require less computational cost than global methods while global methods produce a more elaborate depth map than that of local methods.

After obtaining a disparity map, we can refine the depth estimates using filtering methods such as a median filter and a box filter. This step is called as disparity refinement, which is not a necessary procedure of depth estimation. However, refinement could be useful to improve the quality of depth estimates with only a single image and the corresponding depth map.

2.2 Baseline vs. Disparity

Computational cost and depth accuracy are strongly correlated to the number of disparity candidates, which is proportional to the system baseline as shown in Eq. (2.2).

• **Wide-Baseline Stereo.** *Wide-baseline stereo* reserves more pixels for disparity than narrow-baseline stereo does. Therefore, the wide-baseline systems can discriminate depth with a higher resolution. On the other hand, the search range of correspondences increases, and in turn, it increases the chances of false

matching. The estimated disparity map is plausible in terms of depth but includes many small regions without depth as spatial artifacts (of holes) on the depth map. This missing information is caused by occlusion and false matching in unfeatured or pattern-repeated regions, where the corresponding point search fails.

• **Narrow-Baseline Stereo.** *Narrow-baseline stereo* has a relatively short search range of correspondence. We have relatively less chances for false matching so that we can enhance the accuracy and efficiency in cost computation. In addition, the level of spatial noise in the disparity map is low, as the occluded area is small. However, narrow-baseline stereo reserves a small number of pixels to discriminate depth. The depth-discriminative power decreases accordingly, whereas the spatial artifacts in the disparity map reduce. It trades off the discriminative power for the reduced spatial artifacts in the disparity map.

2.3 Binocular Calibration

Binocular stereo uses two cameras for obtaining a binocular cue. In order to meet the epipolar constraint of binocular stereo, we need to set the poses of the image sensors as parallel to each other. Also the levels of the image sensors need to be same. However, in real experimental setup, it is extremely difficult to manually align two cameras to be an ideal poses satisfying epipolar constraint. Therefore, instead of manually adjusting the cameras, the poses of the camera are roughly fixed in general. Then we usually transform the two image planes into two virtual image planes, which meet the epipolar constraint. This process is often called as rectification.

Rectification computes the extrinsic relation between the two camera viewpoints. Camera calibration, which is the process for estimating the intrinsic and extrinsic parameters of a camera, is necessary in advance of rectification. Intrinsic parameters of a camera includes the coordinates of the principal point, the focal length of the lens, its aspect ratio, lens distortion and skewness. Extrinsic parameters are rotation and translation matrices from the world coordinate system to the camera coordinate system. After capturing two cameras by capturing a chessboard target with several different poses, the properties of the two cameras are obtained [31]. Finally, we build up transformation matrices that enable us to have two virtual image planes holding epipolar constraints.

2.4 Depth from Refraction

Refractive stereo estimates depth using the refraction of light via a transparent medium. There has been several studies that tried to formulate the geometric relationship between refraction and depth [6, 12]. Here we formulate the foundations of general depth estimation from refractive stereo.

Suppose a 3D point p in a target scene is projected to p_d on an image plane through the optical center

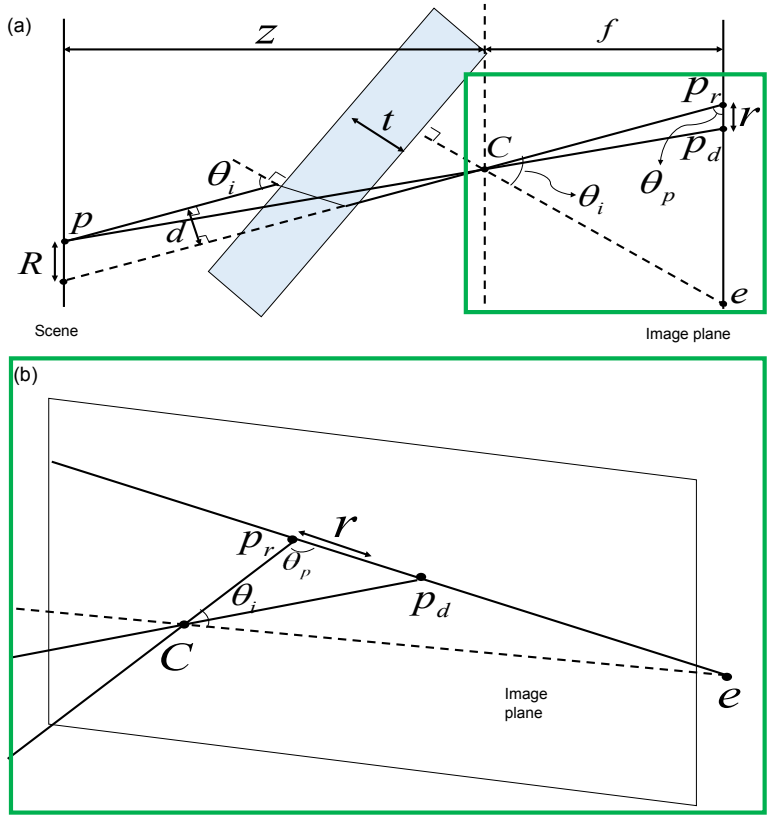


Figure 2.2: *(a) A cross-section view of the light path in refractive stereo. (b) A close-up view of the refractive light transport in 3D.*

of an objective lens C directly without any transparent medium (Fig. 2.2(a)). Inserting a transparent medium in the light path changes the transport of the incident beam from p and reach at p_r on the image plane with a lateral displacement d (between w/ and w/o the medium). The displacement between p_d and p_r on the image plane is called *refractive disparity*.

Now we formulate the depth z of p using simple trigonometry following [11, 12]:

$$z = f\frac{R}{r}, \tag{2.3}$$

where r is a refractive disparity, found by searching a pair of corresponding points; f is the focal length; R is the ratio of lateral displacement d to $\sin(\theta_p)$.

$$R = \frac{d}{\sin(\theta_p)}, \tag{2.4}$$

Here θ_p is the angle between $\overrightarrow{p_r C}$ and the image plane. In order to obtain the value of R, we first compute

$\cos(\theta_p)$ using the following equation:

$$\cos(\theta_p) = \frac{\overrightarrow{p_r e} \cdot \overrightarrow{p_r C}}{|\overrightarrow{p_r e}| \left|\overrightarrow{p_r C}\right|}. \tag{2.5}$$

Then, we simply assign $\sin(\theta_p)$ into Eq. (2.4) after computing $\sin(\theta_p)$ with a simple equation:

$$\sin^2(\theta_p) + \cos^2(\theta_p) = 1. \tag{2.6}$$

Lateral displacement d, the parallel-shifted length of the light passing through the medium, is determined as [14]:

$$d = \left(1 - \sqrt{\frac{1 - \sin^2(\theta_i)}{n^2 - \sin^2(\theta_i)}}\right) t \sin(\theta_i) \ , \tag{2.7}$$

where t is the thickness of the medium; n is the refractive index of the medium; θ_i is the incident angle of the light. $\sin(\theta_i)$ can be obtained in a similar manner with the case of $\sin(\theta_p)$ using the following equation:

$$\cos(\theta_i) = \frac{\overrightarrow{p_r C} \cdot \overrightarrow{eC}}{\left|\overrightarrow{p_r C}\right| \left|\overrightarrow{eC}\right|}. \tag{2.8}$$

The refracted point p_r lies on a line, the so-called *essential line*, passing through an *essential point* e (an intersecting point of the normal vector of the transparent medium to the image plane) and p_d (Fig. 2.2(b)). This property can be utilized to narrow down the search range of correspondences onto the essential line, allowing us to compute matching costs efficiently. It is worth noting that disparity in refractive stereo depends on not only the depth z of p but also the projection position p_d of light and the position of the essential point e, whereas the disparity in traditional stereo depends on only the depth z of the point p. Prior to estimating a depth, we calibrate these optical properties in refractive stereo in advance. See Sec. 4.2 for calibration.

Chapter 3. Related Work

In this chapter, we briefly overview recent depth-from-stereo algorithms that are the most relevant to our work, and categorize them into two groups by the number of cameras employed. One group estimates depth from multi-view stereo; the other group employs a single camera equipped with additional optics, such as an aperture mask or glass window.

3.1 Multi-View Stereo

Multi-view stereo utilizes many images with different viewpoints. In general, we call a system as a multi-view system when the system consists of more than two different viewpoints.

Okutomi and Kanade [25] proposed a multi-baseline stereo method, which is a variant of multi-view stereo. The proposed system consists of multiple cameras on a rail. They presented the matching cost design for the multi-baseline setup. Instead of computing the color difference of a pixel on the reference view and the corresponding point on the other view, color differences of every views are summed up. This method is straightforward, however, the multi-baseline stereo gives more accurate depth estimates over binocular stereo does.

Furukawa and Ponce [9] presented a hybrid patch-based multi-view stereo algorithm that is applicable to objects, scenes, and crowded scene data. Their method produced a set of small patches from matched features, which allows to fill in the gaps between neighboring feature points, yielding a fine mesh model.

Gallup et al. [10] estimated the depth of the scene by adjusting the baseline and resolution of images from multiple cameras so that the depth estimation becomes computationally efficient. This system can exploit the advantages of multi-baseline stereo while requiring a mechanical support of the moving cameras.

Nakabo et al. [23] presented a variable-baseline stereo system on a linear slider. They controlled the baseline of the stereo system depending on the target scene for estimating the accurate depth map.

Zilly et al. [32] introduced a multi-baseline stereo system with various baselines. Four cameras are configured in multiple baselines on a rail. The two inner cameras establish a narrow-baseline stereo pair while two outer cameras form a wide-baseline stereo pair. They then merge depth maps from two different baselines. We take inspiration from this work [32] to extend the multiple baseline idea, i.e., we extend the structure of traditional binocular stereo by adopting a refractive medium to one of the cameras. Refer to [27] for in-depth investigation on other multi-view methods.

3.2 Single-View Stereo

Nishimoto and Shirai [24] first introduced a refractive camera system by placing a refractive medium in front of a camera. Rather than computing depth from depth from refraction described in Sec. 2.4, their method estimates depth using a pair of a direct image and a refracted one, assuming that the refracted image is equivalent to one of the binocular stereo images.

Lee and Kweon [17] presented a single camera system that captures a stereo pair with a bi-prism. The bi-prism is installed in front of the objective lens to separate the input image into a stereo pair with refractive shift. The captured image includes a stereo image pair with a baseline. Depth estimation is analog to the traditional methods.

Gao and Ahuja [11, 12] proposed a seminal refractive stereo method that captures multiple refractive images with a glass medium titled at different angles. This method requires optical calibration of the every pose of the medium. It was extended by placing a glass medium on a rotary stage in [12]. The rotation axis of the titled medium is mechanically aligned to the optical axis of the camera resulting in that the position of an essential point lies on a circle with a specific radius. Although the mechanical alignment is cumbersome, this method achieves more accurate depth than the previous one does.

Shimizu and Okutomi [28, 29] introduced a mixed approach that combines the refraction and the reflection phenomena. This method superposes a pair of reflection and refraction images via the surface of a transparent medium. This overlapped image is utilized as a pair of stereo images.

Chen et al. [5, 6] proposed a calibration method for refractive stereo. This method finds the pairs of matching points on refractive images with the SIFT algorithm [20] to estimate the pose of a transparent medium. They then search corresponding features using the SIFT flow [19]. By estimating the rough scene depth, they recover the refractive index of a transparent medium.

In addition to the refraction-based approaches, Levin et al. [18] introduced a coded aperture-based approach, where they insert a coded aperture blade inside a camera lens instead of a conventional aperture. It allows to estimate depth by evaluating the blur kernels of the coded aperture.

Bando et al. [1] presented a color-filtered aperture in a commodity camera, where the sub-apertures of red, green and blue colors are windowed at different positions. This optical design enables the camera to form three color channels with geometric shift at different positions to yield depth. They extract depth from the shifted channels, analogous to traditional depth from defocus.

In this thesis, we adopt an optical hardware structure of refractive stereo [12] and combine it on a binocular stereo base.

Chapter 4. System Implementation

We propose a novel stereo fusion system by taking advantages of the refractive and binocular stereo systems. This chapter describes the technical details about the hardware design and the calibration methods for the proposed system.

4.1 Hardware Design

Our stereo fusion system consists of two cameras and a transparent medium on a mechanical support structure. The focal length of the both camera lenses is the same as 8 mm. The cameras are placed on a rail in parallel with a baseline of 10 cm to configure binocular stereo. We place a transparent medium on a rotary stage for refractive stereo in front of one of the binocular stereo cameras. See Fig. 4.1 for our system diagram and Fig. 4.2 for our actual prototype. Note that refractive stereo presents a smaller disparity than traditional binocular stereo because it creates the disparity from the change of the light direction by refraction. Therefore, we regard refractive stereo as being equivalent to narrow-baseline stereo in terms of disparity in this work. i.e., refractive stereo is equivalent to short-baseline stereo in terms of disparity. Binocular stereo structure is equivalent to wide-baseline stereo in our system. Refer to Sec. 2.2 for more details.

Our transparent medium is a block of clear glass. The measured refractive index of the medium is 1.41 ($n = \sin(20°)/\sin(14.04°)$); the thickness of the medium is 28 mm. We built a customized cylinder to hold the medium, cut in 45° from the axis of the cylinder. We spin the titled medium about the optical axis from 0° to 360° in 10° intervals while capturing images. The binocular stereo baseline and the tilted angle of the medium are fixed rigidly while capturing.

For the input images of a scene, we use multiple refracted images from the camera of the refractive module by rotating the refractive medium. And we also obtain another image from the other camera without the glass. Note that we do not capture the scene on the camera of refractive module without the glass.

4.2 Calibration

Our stereo fusion system requires several stages of calibration prior in order to estimate depth information. This section summarizes our calibration processes.

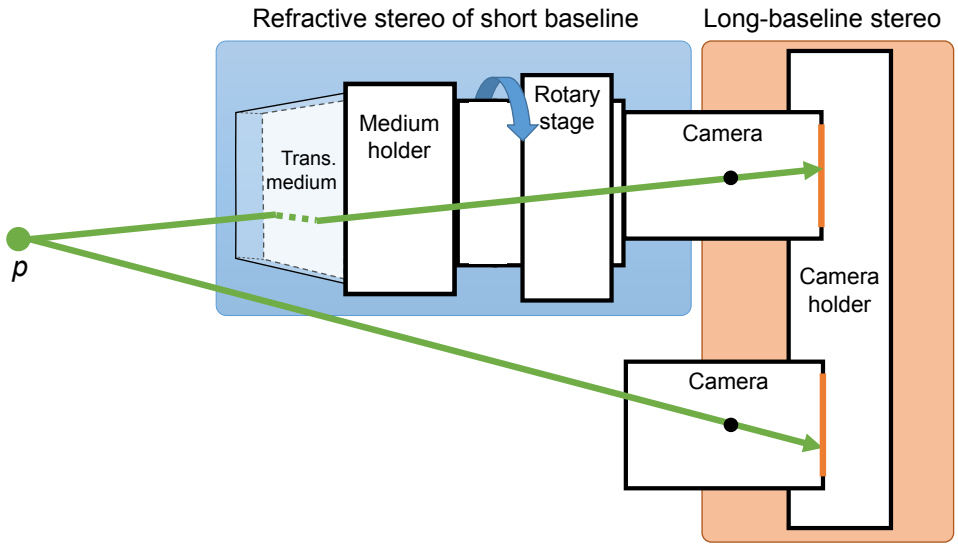


Figure 4.1: *The schematic diagram of our stereo fusion system. A point p is captured by both the refractive stereo and the binocular stereo module.*

Figure 4.2: *Our system prototype.*

4.2.1 Geometric Calibration

We first calibrate the extrinsic/intrinsic parameters of the cameras, including the focal length of the objective lens, the center point of the image plane and the lens distortion in order to convert the image coordinates into the global coordinates.

For the geometric calibration, we captured 14 different poses of a chessboard as shown in Fig. 4.3. This allows us to derive an affine relationship between the two cameras and rectify the coordinates of

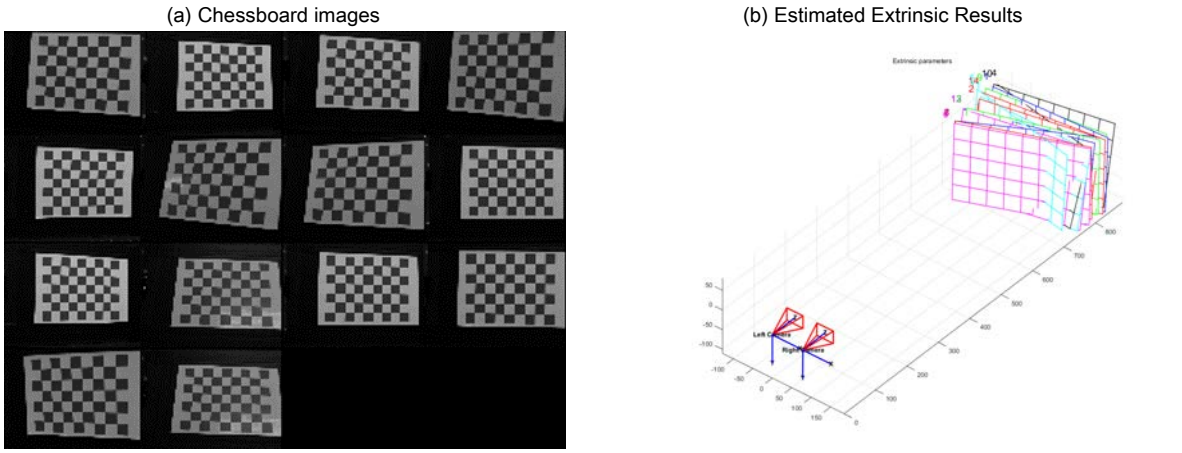


Figure 4.3: *(a) shows the multiple images captured for different poses of a chessboard. In our experiment, we utilizes 14 calibration images. (b) presents the estimated poses of the stereo system with respect to the chessboards.*

these cameras with respect to the constraint epipolar line [31].

However, since we do not have the direct image on the refractive module for a scene, first we need to recover a synthetic direct image (see Sec. 5.1.5). Then, the epipolar constraint is satisfied by applying the transformation to the synthetic direct image.

4.2.2 Refractive Calibration

Refractive stereo requires additional calibrations of the optical properties such as glass thickness, the refractive index and the essential point. Here we describe the calibration detail of the essential points.

Analogous to the epipolar line in binocular stereo, refractive stereo forms an essential point e, where the essential lines forge to the essential point e outside the image plane, i.e., a refracted point p_r passes through an unrefracted pixel p_d and reaches the essential point e on the essential line (see Fig. 2.2(b)).

Gao and Ahuja [11, 12] estimated the essential point by solving an optimization problem with a calibration target at a known distance. They precomputed the positions of the essential points for all angles by manually adjusting the normal axis of the glass, so that the accuracy of estimating the essential points does not depend on a target scene.

Instead of placing an target at a known distance and solving the optimization problem, Chen et al. [6] directly estimated the essential point on target scene images with a fact that all essential lines meet at the essential point. They estimated the position of the essential point by computing intersection points of lines passing through each matching point on the superposed images with and without the medium. This method is considerably simpler than solving the optimization problem [11, 12]; however, the goal of refractive stereo is to estimate the corresponding point of a pixel. In that sense, this method makes the calibration process become a chicken-and-egg problem. On the other hands, searching corresponding features with SIFT [6] is not consistent often such that the calibration accuracy is bound to the SIFT performance.

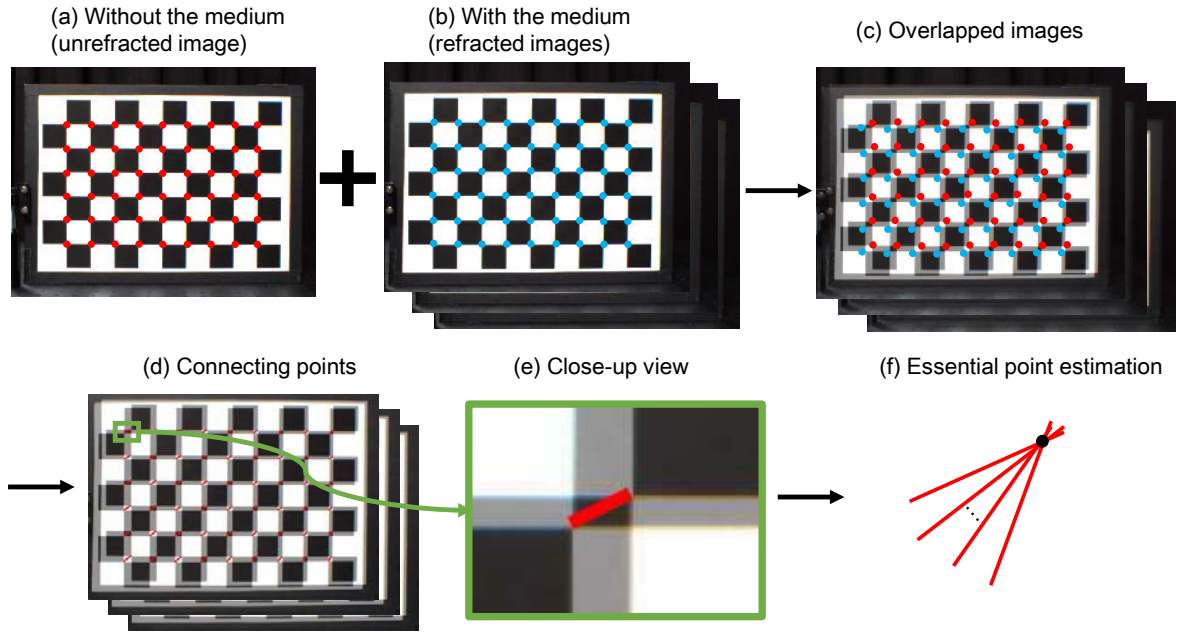


Figure 4.4: *(a) and (b) We first captured a chessboard with and without the medium for the target poses of the medium. (c) After extracting corner points on the chessboard, we superposed the two images for each pose of the medium. (d) and (e) Then, a line connecting corresponding points for each corner is drawn on the overlapped image. (f) The extended red lines that link the correspondences of features forge to an essential point outside the image. Therefore, an essential point of a pose is estimated by arithmetic mean of the line-by-line intersection points.*

Our calibration method takes advantages of the both methods [6, 12] to estimate the essential points with 36 poses of the refractive medium on the rotary stage in advance. We take an image of a checkerboard without the medium once to compare it with other refracted images in different poses of the medium. Once we take a refracted image in a pose, we extract corner points from the both direct and refracted images as shown in Fig. 4.4(a) and (b). Note that the same feature points appear at different positions due to refraction. Superposing these two images, we draw lines by linking the corresponding points with all feature corners with the fact observed by Chen et al. [6] (Fig. 4.4(c)). We then compute the arithmetic mean of the coordinates of the intersection points to determine an essential point per rotation angle

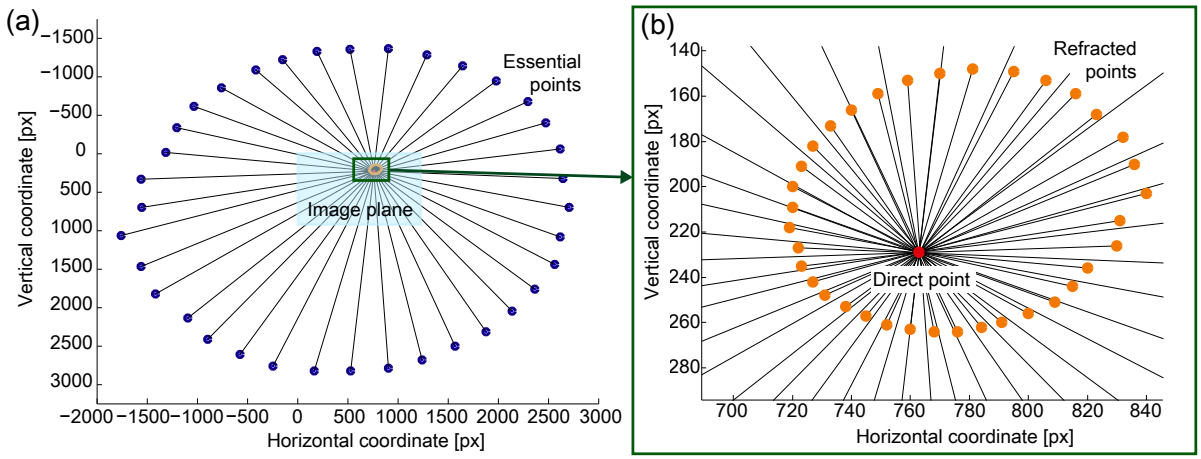


Figure 4.5: *(a) presents the calibrated results of the 36 essential points (blue dots) in our system. (b) shows an example of the locations of 36 refracted points (orange dots) from a direct point (without the medium, red point) in the coordinates of (763,229) at a distance of 30 cm. The location of the direct point has been refracted to 36 different positions per rotation due to refraction.*

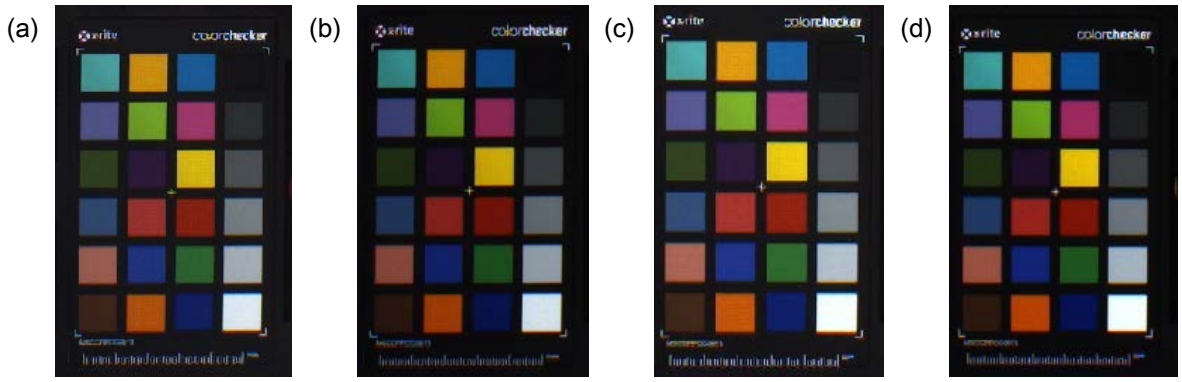


Figure 4.6: *(a) and (b) show the RGB color patches and the linearized RGB color patches on the camera of refractive module with the medium. Also (c) and (d) are the RGB and linearized RGB color patches on the other camera of a binocular module.*

(Fig. 4.4(f)). We repeat this process with the 36 rotation poses of the medium predetermined in 10-degree intervals. Fig. 4.5 shows the estimated essential points.

4.2.3 Color Calibration

Matching costs are calculated by comparing the intrinsic properties of color at the feature points. Since we introduce a transparent medium on a camera in binocular stereo, it is critical to achieve consistent camera responses with and without the medium. Note that we need to match color characteristics between the camera with a transparent medium on the refractive module and the the other camera without the medium.

To do so, we employ a GretagMacbeth ColorChecker target of 24 color patches. We first capture an image from the refractive module with the medium, and an image from the other camera without the

medium. Then, we linearize the two RGB images with known gamma values by inverse gamma correction. For the 24 color patches, now we have two set of linear RGB colors, A and B (cameras with and without the medium with inverse gamma correction), which are measured from the both cameras (Fig 4.6). Note that the dimension of the A and B is both 24×3. Now, we determine a 3×3 affine transformation M of A to B as a camera calibration function using least-squares [16].

We apply this color transform M for the linear RGB image which is generated from the images taken by the camera with the medium. It produces a reconstructed image as if the image is taken from the refractive module camera having consistent color responses with the other camera without the medium.

Chapter 5. Depth Reconstruction in Stereo Fusion

Our stereo fusion workflow is composed of two steps. We first estimate an intermediate depth map from a set of refractive stereo images (from the camera with the medium) and reconstruct a synthetic direct image. Then, this virtual image and a direct image (from the other camera without the medium in a baseline) are used to estimate the final depth map referring to the intermediate depth map from refractive stereo. Fig. 5.1 overviews the workflow of our stereo fusion method.

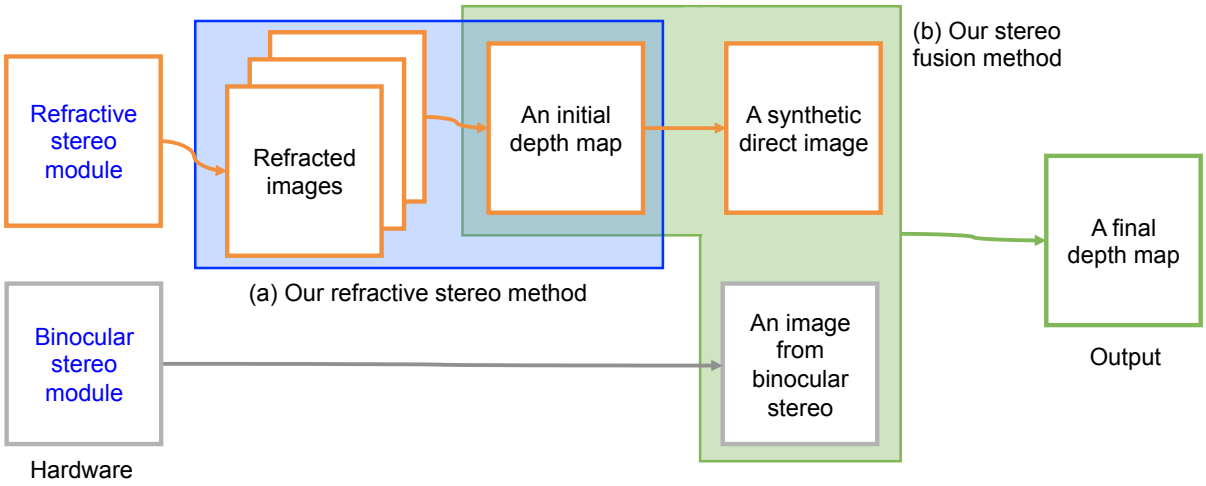


Figure 5.1: *Schematic diagram of our stereo fusion method. (a) Our refractive stereo method estimates an intermediate depth map from refractive stereo. (b) Our stereo fusion method reconstructs a final depth map from a pair of an image from binocular stereo and a synthetic direct image using the intermediate depth map.*

5.1 Depth from Refraction

Depth reconstruction from binocular stereo has been well-studied including matching cost computation, cost aggregation, disparity computation, and disparity refinement [26], whereas depth reconstruction from refraction has been relatively less discussed. In this section, we describe our approach for refractive stereo for reconstructing an intial depth map.

5.1.1 Matching Cost in Refractive Stereo

General binocular stereo algorithms define the *matching cost volumes* of every pixels per disparity [26], where a disparity (proportional to the inverse of the depth [25]) implies a certain depth directly in binocular stereo. This relationship can be applied for all the pixels in the stereo image uniformly. It is

important to note that the disparity in refractive stereo however changes, different from binocular stereo, by not only the depth but also the coordinates on the image plane and the pose of the medium. It means that the refracted points of a single direct point could have different refractive disparities depending on the coordinates on the image plane and the pose of the medium. We therefore define the matching cost volumes for our refractive stereo based on the depth, rather than the disparity. This allows us to apply a cost volume approach for refractive stereo.

5.1.2 Disparity- vs. Depth-based Matching Cost

Disparity is a bitmap metric which can be converted into depth if the focal length and the baseline of a stereo system are known. Since the unit of disparity is pixel, the position of a corresponding point (r', c') with a specific disparity *disp* for a pixel (r, c) can be computed by adding the disparity to the column of the original pixel following:

$$(r', c') = (r, c + disp), \tag{5.1}$$

where the original pixel (r, c) lies on the right camera, and the corresponding pixel (r', c') is on the left camera.

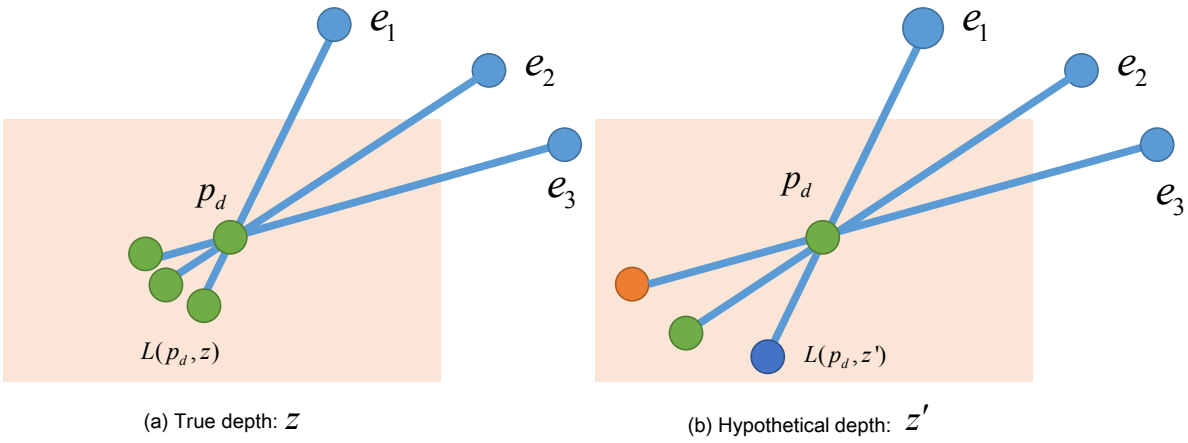
However, in the case of refractive stereo, the refractive disparity varies along the position of a pixel on the image plane. If we uniformly sample the refractive disparity candidates over the every pixels on the image plane, then the corresponding depths of the refractive disparities are completely different for each pixel resulting in non-uniform depth resolutions per a pixel. Therefore, we need to use not disparity but depth for the metric of matching costs.

There have been some works utilizing depth as a metric metric of matching costs. Okutomi and Kanade [25] presented the inverse-depth-based volume having an important benefit which is proportionality to disparity in a multi-baseline stereo system. In contrast to the multi-baseline stereo, refractive stereo cannot satisfy the proportionality to disparity as pixel position also changes the relation between the refractive disparity and depth. Therefore, we define a *refractive matching cost volume* as a depth-based one.

Suppose we have a geometric position set P of the refracted points $p_r(p_d, z, e)$ of a direct point p_d at a depth z (see Fig. 2.2) with an essential point e ($e \in E$): All $R(p_d, z)$ depends on the coordinates of the unrefracted point p_d and the depth z of the point:

$$P(p_d, z) = \{p_r(p_d, z, e) | e \in E\}. \tag{5.2}$$

This set P can be derived analytically by refractive calibration (Sec. 2.4) so that we precompute this set P for computational efficiency, inspired by [12].



(a) True depth: z (b) Hypothetical depth: z'

Figure 5.2: *A pixel p_d on a direct image is refracted into different pixel positions depending on the pose of the medium. We now assume that three poses of the medium of which essential points are e_1, e_2 and e_3. (a) When the depth of p_d is z, the color of the refracted pixels $L(p_d, z)$ would have high similarity when Lambertian assumption holds. (b) However, if we test the depth of p_d is z' which is different from the true depth z, $L(p_d, z')$ has low similarity since the positions of refracted pixels have wrong estimates.*

We denote L as the set of colors observed at the refracted positions P, where l is a color vector in a linear RGB color space ($l \in L$). Assuming that the surface of the direct point p_d is Lambertian, the colors of the refracted points $L(p_d, z)$ would be the same (see Fig. 5.2). We use the similarity of $L(p_d, z)$ for the matching cost C of p_d with a hypothetical depth z [15]:

$$C(p_d, z) = \frac{1}{|L(p_d, z)|} \sum_{l \in L(p_d, z)} K(l - \bar{l}). \tag{5.3}$$

K is an Epanechnikov kernel [8] following:

$$K(l) = \begin{cases} 1 - \|l/h\|^2, & \|l/h\| \leq 1 \\ 0, & \text{otherwise} \end{cases}, \tag{5.4}$$

where h is a normalization constant ($h = 0.01$). $\bar{l}$ is a mean color vector of the all element in a set L. We compute l with five iterations in $L(p_d, z)$ using the mean shift method [7] as:

$$\bar{l} = \frac{\sum_{l \in L(p_d, z)} K(l - \bar{l}) l}{\sum_{l \in L(p_d, z)} K(l - \bar{l})}. \tag{5.5}$$

z in our refractive stereo is a discrete depth, of which range is set between 60 cm and 120 cm in 3 cm intervals. Note that we build a refractive cost volume per depth for all the pixels in the refractive image.

5.1.3 Cost Aggregation for Depth Estimation

In order to improve the spatial resolution of the intermediate depth map in refractive stereo, we aggregate the refractive matching cost using a window kernel G.

Advanced cost aggregation techniques, such as guided image [13] and bilateral weights [22], require a prior knowledge of the scene, i.e., a unrefracted direct image. However, we do not capture the direct image in our experiments because this requires detaching the medium for every scene. Therefore, we first aggregate the refractive matching costs using a Gaussian kernel G:

$$G(p_d, q_d) = \frac{1}{2\pi\sigma^2} \exp \frac{-||p_d - q_d||^2}{2\sigma^2}, \tag{5.6}$$

where σ is 9.6.

We filter the refractive matching at a pixel p_d in a depth z, where this kernel convolves $C(p_d, z)$ with the matching costs of neighboring pixels with a weighting factor $G(p_d, q_d)$ [30]:

$$C^A(p_d, z) = \sum_{q_d \in w} G(p_d, q_d) C(q_d, z), \tag{5.7}$$

where q_d is a pixel inside a squared window w, of which size is 7×7.

Finally, we compute the optimal depth $Z(p_d)$ of the point p_d that maximizes the aggregated matching costs:

$$Z(p_d) = \arg\max_z C^A(p_d, z). \tag{5.8}$$

5.1.4 Synthetic Direct Image Reconstruction

Even though the levels of the two cameras are the same on the rail as traditional binocular stereo, our stereo pair includes more than horizontal parallax due to the refraction effect. Prior to fusing the binocular stereo and the refractive depth input, we first reconstruct a synthetic image I_d (a direct image without the medium) by computing the mean radiance of the set $L(p_d, Z(p_d))$ using the mean shift method (Eq. (5.5)). Note that set L consists of colors gathered from the refracted images.

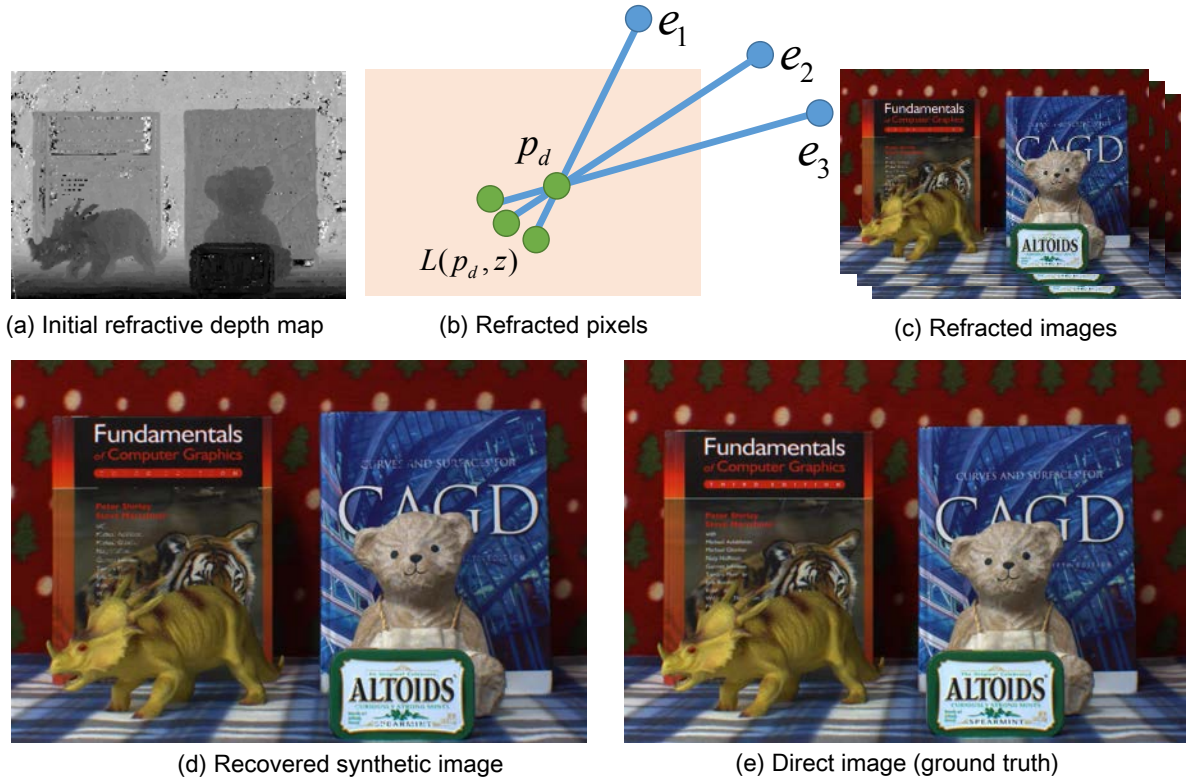


Figure 5.3: *(a) shows an initial depth map. (b) is the the computed positions of refracted pixels using the estimated refractive depth. We recover a synthetic direct image (d) by compute arithmetic mean of the refracted colors (b) on the refracted images (c).*

Fig. 5.3 presents the initial depth map Z and the reconstructed synthetic direct image I_d. If the refractive depth estimates $Z(p_d)$ contains some errors, the resulting synthetic image I_d have also contains errors. However, the visual effect of the artifacts on the reconstructed image does not have significant impact as the wrong depth estimates might be caused by the featureless regions.

5.1.5 Depth and Direct Image Refinement

Reconstructing the direct image allows us to apply a depth refinement algorithm with a weighted median filter [21] by treating the synthetic direct image as guidance in order to fill in the holes of the estimated depth map. The weighted median filter replaces the depth $Z(p_d)$ using the median from the histogram $h(p_d, \cdot)$:

$$h(p_d, z) = \sum_{q_d \in w} W(p_d, q_d) f(q_d, z), \tag{5.9}$$

where $f(q_d, z)$ is defined as follows:

$$f(q_d, z) = \begin{cases} 1, \text{ if } Z(q_d) - z = 0 \\ 0, \text{ otherwise} \end{cases}. \tag{5.10}$$

W is a weight function with a guided image filter [13], defined as:

$$W(p_d, q_d) = \frac{1}{|w|^2} \sum_{k:(p_d,q_d)\in w_k} (1 + (l_d(p_d) - \mu_k)(\Sigma_k + \epsilon U)^{-1}(l_d(q_d) - \mu_k)), \tag{5.11}$$

where $l_d(p_d)$ is a linear RGB color of p_d on the direct image I_d; U is an identity matrix; k is the center pixel of window w_k including p_d and q_d; $|w|$ is the number of pixels in w_k; μ_k and Σ_k are the mean vector and covariance matrix of I_d in w_k. In our experiments, we set the size of w_k as 9×9, and ϵ as 0.001.

This median filter allows us to refine the hole artifacts in the depth map while preserving sound depth. After refining the depth map, the direct image is reconstructed again with the updated depth map. Fig. 5.4 shows the result of the refinement, which are the updated depth map and the direct image.

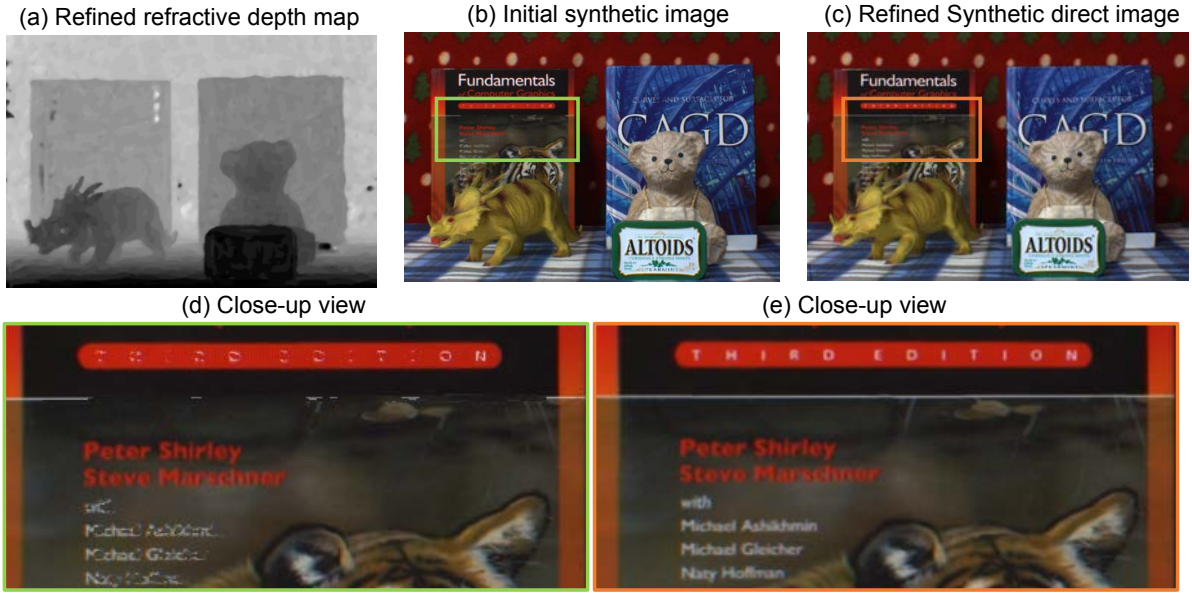


Figure 5.4: *(a) is the refined depth map with weighted median filtering. A synthetic direct image (c) is computed again using the refined depth map (a), used for binocular stereo later. (b) is the initial synthetic image. The refined synthetic image has more details (e) than the initial synthetic image (d).*

After obtaining the final synthetic direct image, we apply the color calibration matrix M (see Sec. 4.2) to the synthetic image so that the synthetic image is used for improving the quality of a depth map by comparing with the image taken from the other camera.

5.1.6 Optimal Number of Refractive Images

We conduct an additional quantitative experiment that measures the point-wise errors of depth estimates in order to find out the optimal number of input refractive images, while maintaining the sufficient spatial resolution of a refractive depth map. We compute the root-mean-square error (RMSE) of depths on a planar surface, which is the red square (Fig. 5.5(a)). Fig. 5.5(b) shows that the RMSE decreases very

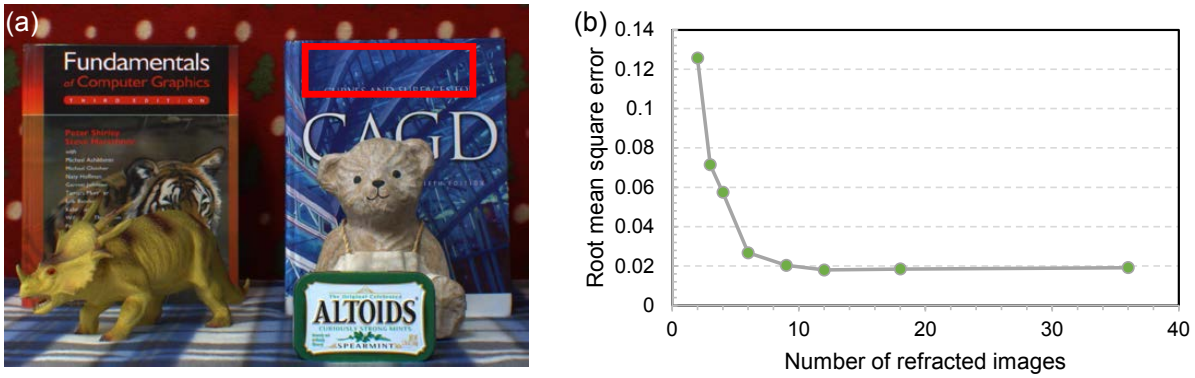


Figure 5.5: *(a) The red square indicates the area used for finding an optimal number of input refractive images. The book cover is a planar surface orthogonal to the camera optical axis with a constant depth. (b) The depth error drops down fast significantly up to six refractive inputs with different angles. No significant improvement is observed with more than six inputs.*

fast while increasing the number of input up to six refractive images. Hence, we determine the optimal number of input refractive images as six. Note that we use six refractive images with 60° intervals for capturing results in this thesis.

5.1.7 Parallax Occlusion

In the case of binocular stereo, only a pair of left and right images is used; therefore, the depth output of the binocular method suffers from typical occlusion artifacts. In contrast, we reconstruct a refractive depth map from a set of refracted images (with the rotation of the medium pose in 360°) so that the refractive depth does not suffer from parallax occlusion.

5.2 Depth in Stereo Fusion

As described in Sec 2.2, our binocular stereo with a wider baseline allows us to discriminate depth with a higher resolution than refractive stereo (equivalent to narrow-baseline stereo). We take inspiration from a coarse-to-fine stereo method [2, 4] to develop our stereo fusion method. Our refractive stereo yields an intermediate depth map with a high spatial resolution, which is on a par with narrow-baseline stereo. However, it is not surprising that the z-depth resolution of this depth map is discrete and coarse on the other hand. We utilize the fine depth map from refractive stereo in order to increase the z-depth resolution as high as possible with a high spatial resolution by limiting the search range of matching cost computation in binocular stereo using the refractive depth map. To this end, we can significantly reduce the chances of false matching while estimating depth from binocular stereo between the direct and synthetic images. This enables us to achieve a fine depth map from binocular stereo, taking advantages of a high spatial resolution in refractive stereo.

5.2.1 Matching Cost in Stereo Fusion

Now we have a direct image I_b from the camera without the medium in the binocular module and the synthetic image I_d reconstructed from the refractive stereo module (Sec. 5.1.5) with its depth map. Depth candidates with uniform intervals are not related linearly to the disparities with pixel-based intervals. We hence define a cost volume for stereo fusion on the disparity instead in order to fully utilize the image resolution. To fuse the depth from binocular and refractive stereo, we build a fusion matching cost volume $F(p_d, d)$ per disparity for all pixels as next. The fusion matching cost F is defined as a norm of the intensity difference:

$$F(p_d, d) = \|l_d(p_d) - l_b(p'_d)\|, \tag{5.12}$$

where p'_d is a shifted pixel by a disparity d from p_d; $l_b(p'_d)$ is a color vector of p'_d on image I_b.

5.2.2 Cost Aggregation in Stereo Fusion

To improve the robustness of cost matching, we employ the bilateral image filter W as weight again, as shown in Eq. (5.9). The size of the kernel w is 9×9, and the value of ϵ is 0.001.

Since the guided filter consists of multiple box filters, the efficient implementation would be feasible through an optimization. However, applying the guided image filter on our refractive stereo fusion algorithm caused a significant computational load while estimating a depth map. Therefore, we ended up with choosing the bilateral filter alternatively. We could achieve a significant improvement in computational cost, by applying it and narrow down the search range in our stereo fusion.

The aggregated cost of the fusion matching costs is defined as:

$$F^A(p_d, d) = \sum_{q_d \in w} W(p_d, q_d) F(q_d, d). \tag{5.13}$$

Here W is the bilateral image filter [30] defined as

$$W(p_d, q_d) = exp\left\{ -\frac{d(p_d, q_d)}{\sigma_s^2} - \frac{c(p_d, q_d)}{\sigma_c^2} \right\}, \tag{5.14}$$

where $d(p_d, q_d)$ is the Euclidean distance between p_d and q_d, $c(p_d, q_d)$ is the sum of differences of colors of RGB channels, σ_s and σ_c are the standard deviations for spatial distance and color difference. In our experiment, we select the window size, σ_s and σ_c as 9, 7 and 0.07.

Suppose the depth of point p_d is estimated as $Z(p_d)$ from refractive stereo. As we compute the refractive matching cost and aggregate the cost *per discrete depth interval* Δz in refractive stereo, let the actual depth of p_d be in between $(Z(p_d) - \Delta z)$ and $(Z(p_d) + \Delta z)$ as Z_{prev} and Z_{post}. The corresponding disparities of Z_{prev} and Z_{post} can be computed as d_{prev} and d_{post} using Eq.(2.2). Note that d_{post} is

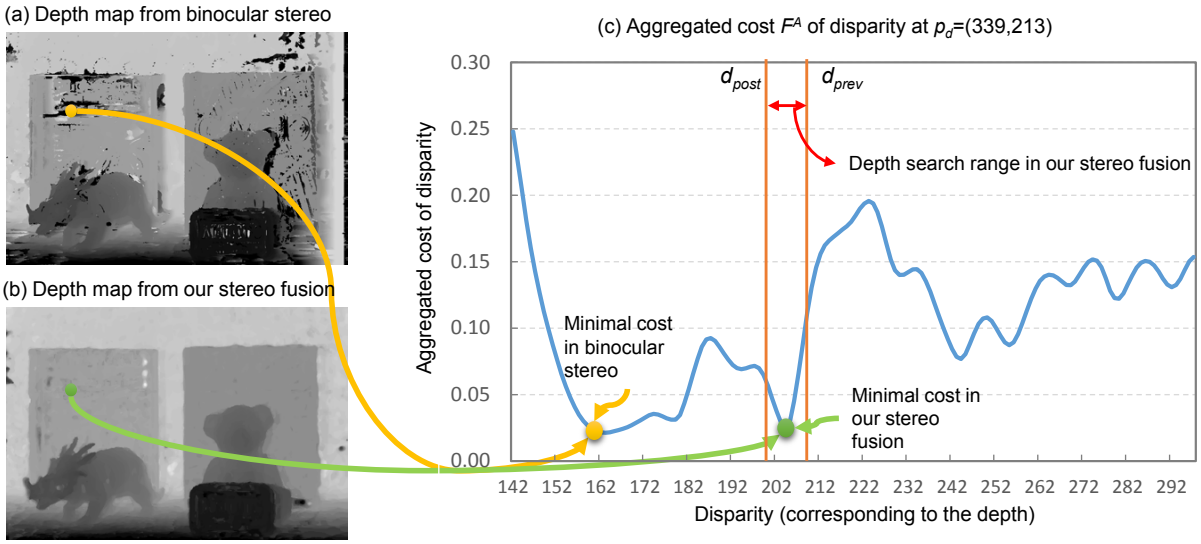


Figure 5.6: *The binocular depth map (a) includes artifacts due to false matching caused by occlusions, featureless regions and repeated patterns. Using the intermediate refractive depth map (b), we can limit the search range of a corresponding point p_d between d_{post} and d_{prev} for instance. This significantly reduces false matching frequency in estimating depth.*

smaller than d_{prev}. We therefore estimate the optimal disparity $D(p_d)$ by searching the aggregated cost volume $F^A(p_d, d)$ within the range $[d_{post}, d_{prev}]$ as below:

$$D(p_d) = \arg\min_d F^A(p_d, d). \tag{5.15}$$

Note that we compute Eq.(5.13) within the range of $[d_{post}, d_{prev}]$ exclusively for computational efficiency.

Fig. 5.6 shows an example of our method. The true disparity of an orange pixel on a scene (a) is around 200. However, since the disparity of the minimal aggregated cost for the orange pixel is around 160, the orange pixel on a depth map from binocular stereo (a) has a wrong depth estimate. In order to solve this artifact problem, we take a coarse-to-fine approach. We can guide the search range of disparity as we already estimates the refractive depth map with less spatial artifacts. As we mentioned before, the search range is set to $[d_{post}, d_{prev}]$. In the end, the true disparity is correctly estimated in our final depth map.

Chapter 6. Results

We conducted several experiments to evaluate the performance of our stereo fusion method. We computed depth maps, of which resolution is 1280×960 with 140 depth steps, on a machine equipped with an Intel i7-3770 CPU and 16GB RAM with CPU parallelization (GPU-based acceleration would be feasible as future work.) The computation times for estimating the depth map from six refractive inputs are ∼77 secs. for the first-half stage of refractive stereo and ∼33 secs. for the second-half stage of stereo fusion. The total computation time on runtime is ∼110 secs. We precomputed the refracted essential points per pixel in the image plane beforehand for computational efficiency.

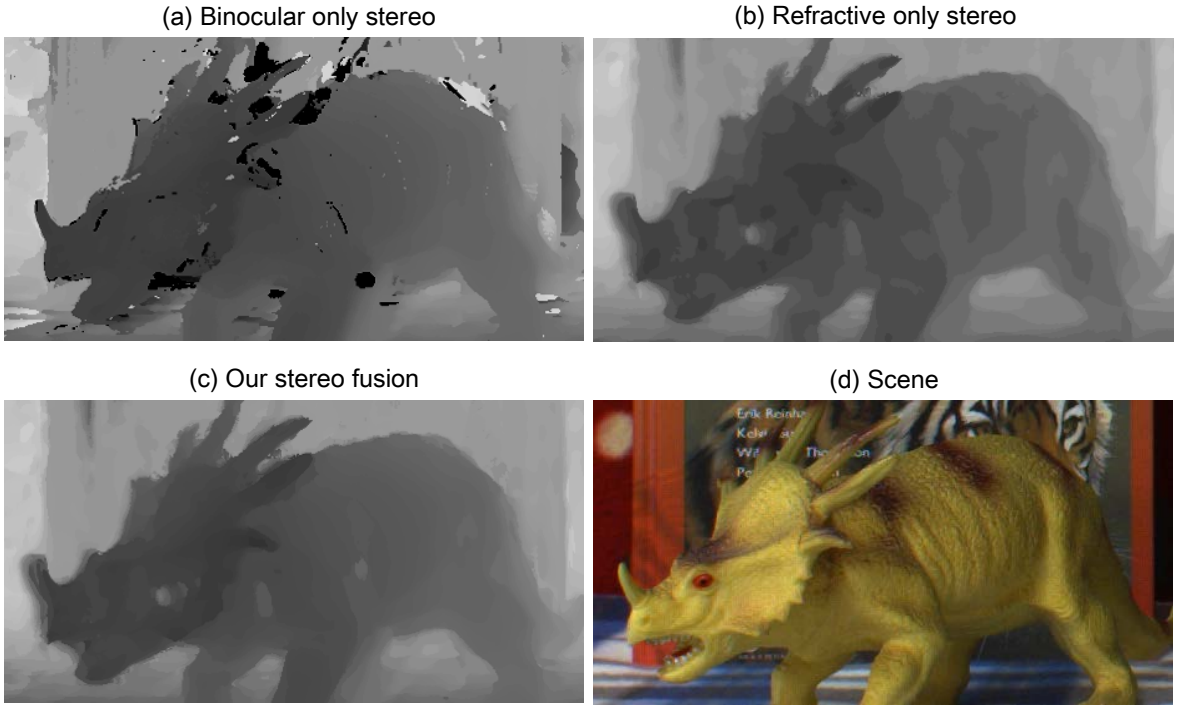


Figure 6.1: *The top two rows compares the three different depth maps of binocular only stereo (a), refractive only stereo (b) from the intermediate stage of our fusion method and our stereo fusion (c) for a scene (d). The depth map of binocular stereo depth is fine but suffers from false matching. Refractive stereo presents depth without artifacts, but its depth is coarse and discrete. Our stereo fusion method estimates depth as fine as binocular stereo without suffering any false match.*

The two rows in Fig. 6.1 compares three different depth maps by binocular only stereo (a), refractive only stereo (b) and our proposed stereo fusion method (c). Although the depth estimation of binocular only stereo (a) appears sound, (a) suffers from typical false matching artifacts around the edges of the front object due to occlusion. Refractive only stereo (b), obtained from the intermediate stage of our fusion method, presents depth without artifacts, but the depth resolution is significantly discretized

and coarse. Our stereo fusion (c) overcomes the shortcomings of the homogeneous stereo methods. It estimates depth as fine as binocular stereo without severe artifacts.

In addition, we quantitatively evaluated the accuracy of our stereo fusion method compared with others in Fig. 6.2(d). We measured three points in the scene using a laser distance meter (Bosch GLM 80) and compared the measurements by the three methods. The accuracy of our method is as high as the binocular only method (aver. distance error: ∼2 mm), outperforming the refractive only method (aver. error: ∼6 mm).

Figure 6.2: *For a target scene, the ground truths and estimated depths of three points (i), (ii), and (iii) are given (see Table 6.1).*

Target point	Binocular only stereo [mm]	Refractive only stereo [mm]	Our stereo fusion [mm]	Ground truth [mm]
(i)	856 (+2)	863 (+9)	856 (+2)	854
(ii)	784 (+2)	784 (+2)	784 (+2)	782
(iii)	873 (+3)	863 (-7)	873 (+3)	870

Table 6.1: *For three points (i,ii,ii) appearing on Fig. 6.2, the estimated depth values are shown in this table. Our refractive stereo cannot distinguish the differences of between (i) and (iii), which is* 16*mm. However, our stereo fusion discriminate the depths as same level of binocular stereo.*

We qualitatively compared our stereo fusion method with a global stereo method [3] (b), a local stereo method [13] (c) and a refractive stereo method [6] (d) in Figs. 6.3, 6.4 and 6.5.

We compared our proposed method with a renowned graphcut-based algorithm [3] with an image of the same resolution. Global stereo methods in general allow for an accurate depth map, while requiring high computational cost. It is not surprising that this global method was about eight times slower than our method (see Figs. 6.3(b), 6.4(b) and 6.5(b)) even if it produces an elaborate depth map.

We also compared our method with a local binocular method [13], which computes the matching cost as the norm of intensity difference and aggregates the cost using the weight of the guided filter [13]. Its computing time was ∼212 secs., which is almost two times slower than our method, with the same scene (see Figs. 6.3(c), 6.4(c) and 6.5(c)) This local method struggles with typical false matching artifacts. Also in terms of speed, our method computes the aggregated costs only for the a few candidates resulting in high computation speed while the local stereo method needs to consider all aggregated costs for many disparity candidates.

A refractive method using SIFT flow [6] is compared to ours (Figs. 6.3(d), 6.4(d) and 6.5(d) and Figs. 6.3(e), 6.4(e) and 6.5(e)). The same number of six refractive images were employed for both methods. While the refractive method suffers from wavy artifacts caused by SIFT flow and its depth resolution is very coarse, typical to refractive stereo, our method estimates depth accurately with less spatial artifacts in all test scenes.

6.1 Multi-baseline Stereo

Multi-baseline stereo methods such as trinocular stereo employ multiple views with various baselines. In this sense, multi-baseline approach is the most similar method to our approach, where the refractive stereo is equivalent to the short baseline stereo; the binocular stereo is equivalent to the long baseline stereo.

We built a trinocular stereo setup, where the distance between the right and the middle camera is set to 2 cm and the distance between the middle and the left one is set to 11 cm to yield multiple baselines. For fair comparison, we compute the depth maps from the trinocular stereo in two ways. Fig. 6.6(b) and Fig. 6.6(c) presents results of trinocular stereo. Fig. 6.6(b) is the result of a multi-baseline method [25], where two pairs of matching costs are calculated from the short and the long baseline pairs, and these costs are combined as total matching cost to yield a depth map. Also the aggregated costs are compute with the guided filter.

Fig. 6.6(c) is another implementation of trinocular stereo. Similar to our coarse-to-fine approach, we compute matching cost volumes from the short-baseline stereo pair and aggregate the volumes through the guided filter to yield an intermediate depth map. We then use this depth map to narrow down the search range of correspondence same as ours.

As shown in Fig. 6.6, our stereo fusion achieves an more elaborate depth map than the both trinocular

stereo methods. We could speculate that our improvement is feasible as our refractive method utilizes the oval shape of corresponding points. This oval-shaped patterns can provide unique signatures in computing the correspondences in our system.

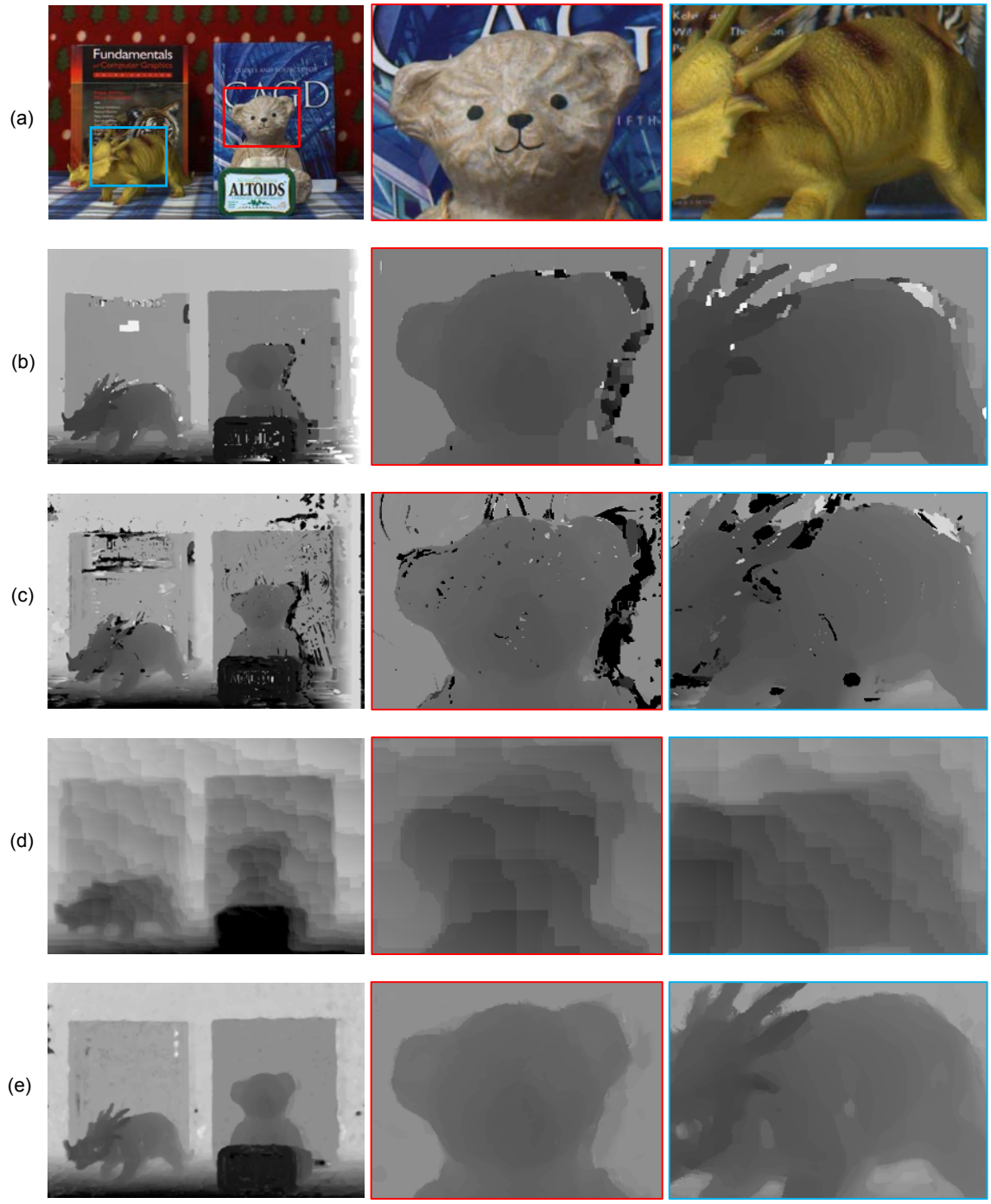


Figure 6.3: *The depth maps for a scene (a) are computed by four different methods. (b) and (c) show global [3] and local binocular stereo [13] methods. (d) presents a refractive stereo method [6]. Our method (e) estimates depth accurately without suffering from severe artifacts.*

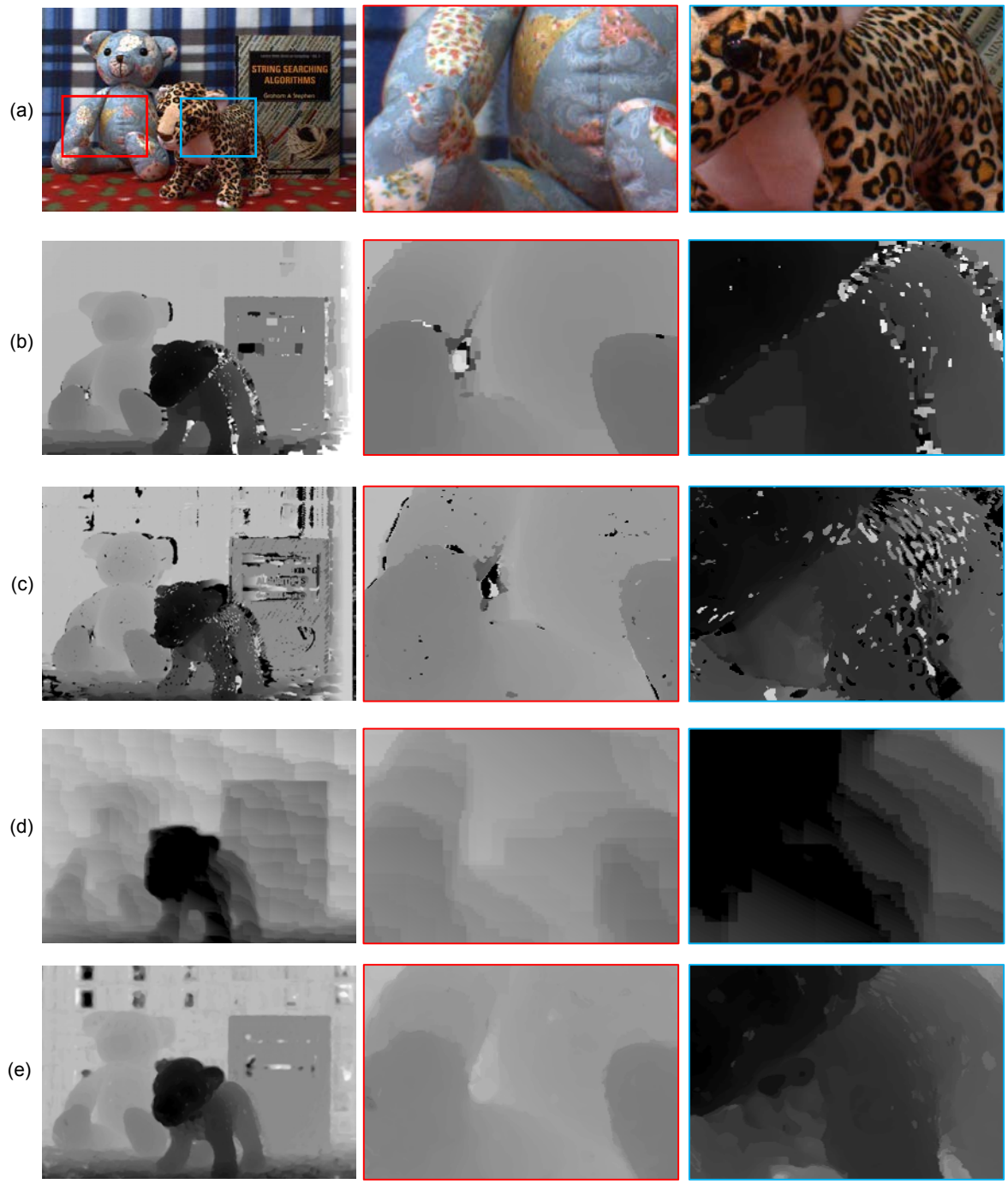


Figure 6.4: *The depth maps for a scene (a) are computed by four different methods. (b) and (c) show global [3] and local binocular stereo [13] methods. (d) presents a refractive stereo method [6]. Our method (e) estimates depth accurately without suffering from severe artifacts.*

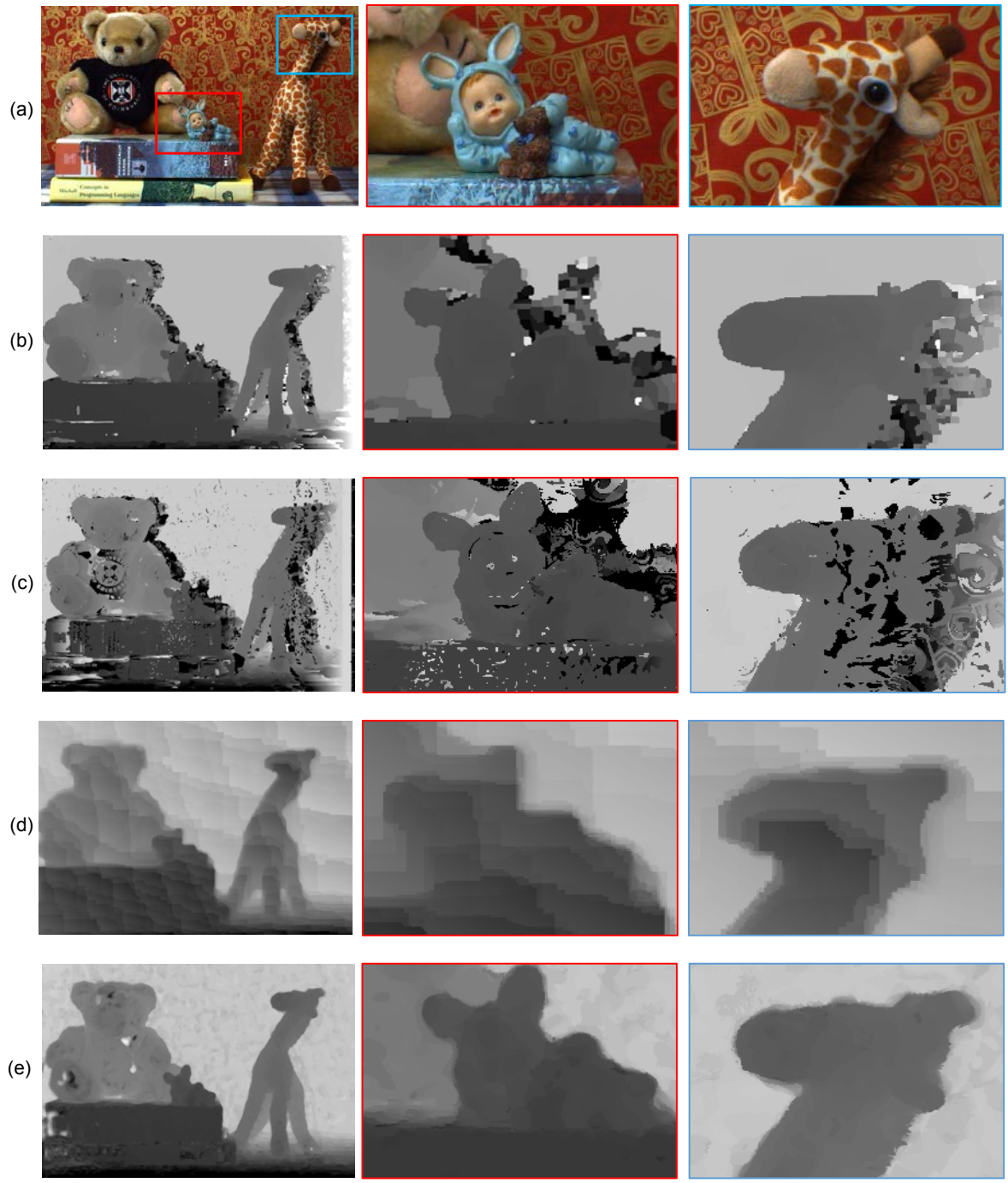


Figure 6.5: *The depth maps for a scene (a) are computed by four different methods. (b) and (c) show global [3] and local binocular stereo [13] methods. (d) presents a refractive stereo method [6]. Our method (e) estimates depth accurately without suffering from severe artifacts.*

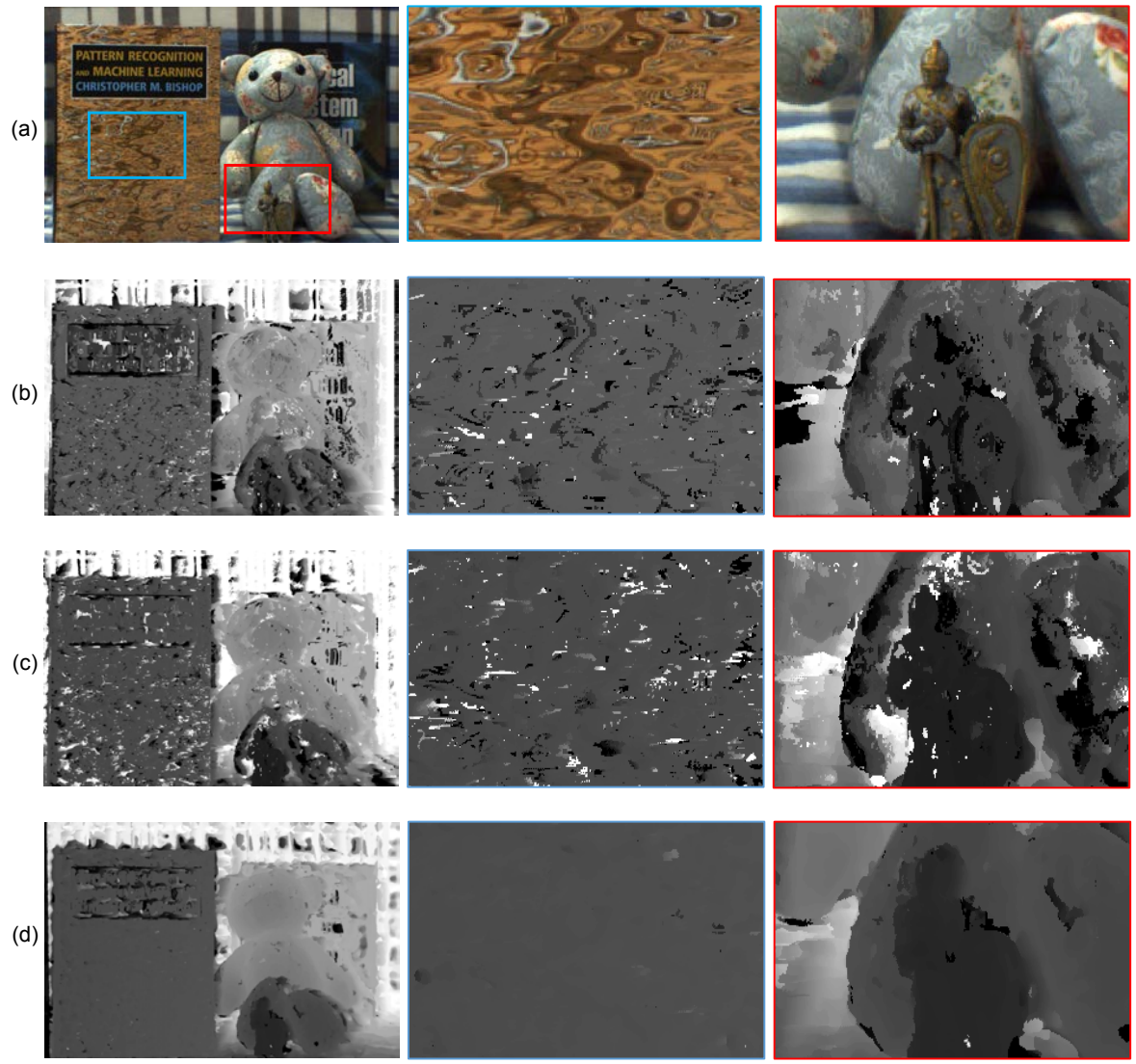


Figure 6.6: *Multi-baseline stereo methods are compared with our method. (a) is the scene image captured by the camera of the refractive module. (b) is a depth map using a trinocular stereo method [25]. (c) is also a trinocular stereo method, implemented with the coarse-to-fine approach same as ours. (d) is the result of our stereo fusion method with the same number of input images.*

Chapter 7. Discussions and Future Work

In this thesis, we proposed a stereo fusion system and workflow for estimating a depth map with less artifacts and high depth resolution. Our refractive system enables us to obtain a depth prior which can be used for increasing the final depth resolution with less artifacts. Synthetic image generation using the estimated refractive depth map is used as a binocular stereo image to improve the depth resolution of the refractive depth map as high as possible. Since coarse-to-fine approach reduces the computational time of our process, we obtain a high-quality depth map faster than the global and even the local method.

We conducted the experiment to find out the optimal number of refracted images which produces a refractive depth map with less artifacts. In the experimental results, we observed that six refracted images would be enough for our case. However, the number of refracted images can be set according to the purpose of an usage.

Our hardware design requires at least one rotation of the medium in order to obtain a depth map using more than two refracted images. It restricts the applications of our system to the static scene only. Also now the medium needs to be manually rotated. Nevertheless, we expect that we can solve the problem by employing the auto-rotary stage for the refractive module and miniaturizing the refractive module. The refractive module of our system currently fixed to the optical breadboard (see Fig. 4.2.). We can reduce the size of the refractive part and place the medium in front of the camera with a body tube connecting the refractive module with the camera. The refractive module of our system currently fixed to the optical breadboard (see Fig. 4.2.). This direction will enable us to easily integrate the auto-rotating refractive module with the existing binocular stereo systems.

Our pipeline currently consists of two steps: refractive depth estimation and stereo fusion. Since our stereo fusion takes a coarse-to-fine approach, errors on the refractive depth map can be transferred to the final depth map. Our assumption on this problem is that the refractive depth map usually have less spatial artifacts due to the narrow-baseline and the oval-shape of corresponding points. In the future, we expect that it is possible to overcome this problem by modifying the fusion algorithm. We will fuse the two different depth maps using multi-scale approach, and obtain a final depth map by solving the multi-scale problem using optimization.

Chapter 8. Conclusions

We have presented a novel optical design, a mixture of binocular and refractive stereo and a stereo-fusion workflow. Our stereo fusion system extracts depth information with high depth resolution and less artifacts with competitive speed to other local and global binocular methods. We validate that our proposed method takes the advantages of both traditional binocular and refractive stereo. Also quantitative and qualitative evaluation shows that our fusion method outperforms the traditional homogeneous methods in terms of artifacts and depth resolution. In addition, our stereo fusion can be easily integrated into any existing binocular stereo, yielding a significant improvement in the number of artifacts.

References

[1] Yosuke Bando, Bing-Yu Chen, and Tomoyuki Nishita. Extracting depth and matte using a color-filtered aperture. *ACM Trans. Graph. (TOG)*, 27(5):134:1–9, 2008.

[2] Stephen T Barnard. Stochastic stereo matching over scale. *Int. J. Comput. Vision (IJCV)*, 3(1):17–32, 1989.

[3] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 23(11):1222–1239, 2001.

[4] Jer-Sen Chen and Gérard Medioni. Parallel multiscale stereo matching using adaptive smoothing. In *Proc. European Conference on Computer Vision (ECCV)*, pages 99–103. 1990.

[5] Zhihu Chen, KK Wong, Yasuyuki Matsushita, Xiaolong Zhu, and Miaomiao Liu. Self-calibrating depth from refraction. In *Proc. Int. Conf. Comput. Vision (ICCV)*, pages 635–642, 2011.

[6] Zhihu Chen, Kwan-Yee K Wong, Yasuyuki Matsushita, and Xiaolong Zhu. Depth from refraction using a transparent medium with unknown pose and refractive index. *Int. J. Comput. Vision (ICJV)*, pages 1–15, 2013.

[7] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 24(5):603–619, 2002.

[8] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[9] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(8):1362–1376, 2010.

[10] David Gallup, J-M Frahm, Philippos Mordohai, and Marc Pollefeys. Variable baseline/resolution stereo. In *Proc. Comput. Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[11] Chunyu Gao and Narendra Ahuja. Single camera stereo using planar parallel plate. In *Proc. Int. Conf. Pattern Recognition (ICPR)*, volume 4, pages 108–111, 2004.

[12] Chunyu Gao and Narendra Ahuja. A refractive camera for acquiring stereo and super-resolution images. In *Proc. Comput. Vision and Pattern Recognition (CVPR)*, pages 2316–2323, 2006.

[13] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *Proc. European Conference on Computer Vision (ECCV)*, pages 1–14. Springer, 2010.

[14] Eugene Hecht. *Optics.* Addison-Wesley, Reading, Mass, 1987.

[15] Changil Kim, Henning Zimmer, Yael Pritch, Alexander Sorkine-Hornung, and Markus Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph. (TOG)*, 32(4):73:1–12, 2013.

[16] Min H. Kim and Jan Kautz. Characterization for high dynamic range imaging. *Computer Graphics Forum (Proc. EUROGRAPHICS 2008)*, 27(2):691–697, 2008.

[17] DooHyun Lee and InSo Kweon. A novel stereo camera system by a biprism. *IEEE Trans. Robotics and Automation*, 16(5):528–541, 2000.

[18] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graphics (TOG)*, 26(3):70:1–9, 2007.

[19] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):978–994, 2011.

[20] David G Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision (IJCV)*, 60(2):91–110, 2004.

[21] Ziyang Ma, Kaiming He, Yichen Wei, Jian Sun, and Enhua Wu. Constant time weighted median filtering for stereo matching and beyond. In *Proc. Int. Conf. Computer Vision (ICCV)*, pages 1–8, 2013.

[22] Stefano Mattoccia, Simone Giardino, and Andrea Gambini. Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering. In *Proc. Asian Conference on Computer Vision (ACCV)*, pages 371–380. Springer, 2010.

[23] Yoshihiro Nakabo, Toshiharu Mukai, Yusuke Hattori, Yoshinori Takeuchi, and Noboru Ohnishi. Variable baseline stereo tracking vision system using high-speed linear slider. In *Proc. Int. Conf. on Robotics and Automation (ICRA)*, pages 1567–1572, 2005.

[24] Y Nishimoto and Y Shirai. A feature-based stereo model using small disparities. In *Proc. Comput. Vision and Pattern Recognition (CVPR)*, pages 192–196, 1987.

[25] Masatoshi Okutomi and Takeo Kanade. A multiple-baseline stereo. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 15(4):353–363, 1993.

[26] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision (IJCV)*, 47(1-3):7–42, 2002.

[27] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. Comput. Vision and Pattern Recognition (CVPR)*, pages 519–528, 2006.

[28] Masao Shimizu and Masatoshi Okutomi. Reflection stereo-novel monocular stereo using a transparent plate. In *Proc. Canadian Conf. Computer and Robot Vision (CRV)*, pages 14–14. IEEE, 2006.

[29] Masao Shimizu and Masatoshi Okutomi. Monocular range estimation through a double-sided half-mirror plate. In *Proc. Canadian Conf. Computer and Robot Vision (CRV)*, pages 347–354. IEEE, 2007.

[30] Kuk-Jin Yoon and In So Kweon. Adaptive support-weight approach for correspondence search. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 28(4):650–656, 2006.

[31] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 22(11):1330–1334, 2000.

[32] Frederik Zilly, Christian Riechert, Marcus Müller, Peter Eisert, Thomas Sikora, and Peter Kauff. Real-time generation of multi-view video plus depth content using mixed narrow and wide baseline. *J. Visual Communication and Image Representation*, 25(4):632–648, 2013.

Summary

Stereo Fusion using a Refractive Medium on a Binocular Base

양안 기반 스테레오 시스템의 깊이 정보는 카메라 사이의 거리인 베이스라인에 따라서 영향을 받게 된다. 긴 베이스라인은 짧은 베이스라인보다 깊이를 구별하는 능력이 뛰어나지만, 깊이 정보의 공간적 결함이 많다. 반면 짧은 베이스라인은 깊이 정보의 결함이 긴 베이스라인일때 보다 더 적지만 디스페리티가 작기 때문에 깊이 해상도가 낮다. 본 논문에서는 서로 다른 스테레오 시스템을 융합하는 새로운 광학적 디자인을 제안한다. 굴절 매질을 기존의 양안 기반 시스템에 장착하여 양안 시스템으로부터 긴 베이스라인의 장점과 굴절 매질 스테레오를 통해 짧은 베이스라인의 장점 두가지를 모두 가지는 깊이 정보를 획득할 수 있다. 또한 본 디자인에 맞는 깊이 해상도와 공간적 결함의 관점에서 높은 질의 깊이 맵을 획득하기위한 융합 알고리즘 또한 소개한다. 본 시스템의 성능은 질적 그리고 양적인 실험 결과를 통해 기존의 여러가지 스테레오 방법들과 비교하여 입증된다.

감 사 의 글

이 석사 학위 논문을 완성하기까지 많은 분들의 도움을 받았습니다. 연구란 무엇인지 가르쳐주시고 2년내내 옆에서 가장 많이 도와주신 김민혁 교수님 정말 감사합니다.

철없는 자식 항상 응원해주시고 힘이 되주시는 어머니, 아버지 감사합니다. 집에 갈때마다 큰 활력소가 되어주는 벌써 대학생이 된 동생 승희에게도 고맙다는 말을 전하고 싶습니다.

같은 연구실 멤버들에게도 정말 감사합니다. 똑똑하고 연구실 멤버들 잘 챙겨주는 인창이형, 결단력 있는 연구실 큰형 길주형, 우직하면서 수학 잘하는 영범이형, 사람들 잘 쟁켜주는 똑똑한 해봄이형, 연구실 막내 광학의 달인 석준이, 분위기 메이커 동갑 주호 모두 함께할 수 있어서 너무 감사했습니다.

고등학교 때부터 지금까지 계속 서로에게 힘이 되어주고 있는 선호, 승현이, 윤래, 원준이, 준희, 희용이에게도 말로 못할 고마움을 전하고 싶습니다.

마지막으로 언급하지 못했지만 살아가면서 인연이 닿은 많은 분들이 있어 지금의 제가 있을 수 있을 수 있었습니다. 그 분들께 정말 감사드립니다.

이 력 서

이 름 : 백 승 환

생 년 월 일 : 1991년 2월 13일

E-mail 주 소 : shwbaek@vclab.kaist.ac.kr

학 력

2009. 3. – 2013. 2. 서강대학교 컴퓨터공학부 (B.S.)

경 력

2013. 3. – 2014. 12. 한국과학기술원 전산학과 일반조교

연 구 업 적

1. **Seung-Hwan Baek** and Min H. Kim, *Stereo Fusion using a Refractive Medium on a Binocular Base*, Proc. Asian Conference on Computer Vision (ACCV), November., 2014, **Oral presentation, Best Application Paper Award and Best Demo Award**