Stereo Fusion using a Refractive Medium on a Binocular Base

Seung-Hwan Baek* Min H. Kim*

KAIST

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 paper, 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.

1 Introduction

Classical stereo matching algorithms employ a pair of binocular stereo images. Such stereo algorithms estimate depth by evaluating the distance of corresponding features, so-called disparity, via computing matching costs and aggregating the costs [1]. However, owing to the nature of triangulation in estimating depth, the depth accuracy strongly depends on its baseline between the stereo pair. For instance, a wide baseline elongates the range of the correspondence search so that the matching problem cannot be solved with high precision in typical locally-optimizing approaches [2]. On the contrary, a narrow baseline shortens the resolution of disparity; therefore, the accuracy of estimated depth could be degraded [3,4].

Recently, Gao and Ahuja [5,6] introduced a single depth camera based on refraction. Chen et al. [7] further extended this refractive mechanism. Such refractive stereo systems estimate depth from the change of light direction; therefore, the disparity in refractive stereo in general is smaller than that in binocular stereo, i.e., its performance is similar to that of binocular stereo with a narrow baseline. We take inspiration from refractive stereo to combine these two heterogeneous stereo systems, where a stereo fusion system is designed with a

* {shwbaek;minhkim(corresponding author)}@vclab.kaist.ac.kr

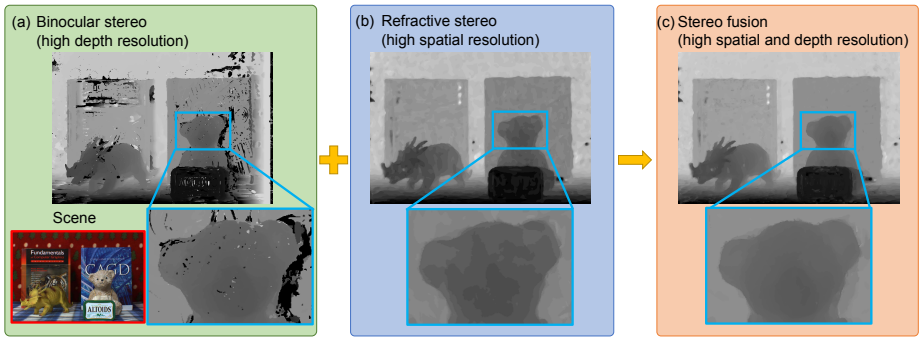


Fig. 1: *(a) A binocular stereo detects depth accurately; whereas, it suffers from spatial artifacts caused by occlusions and featureless regions. (b) A refractive stereo improves the spatial resolution with less artifacts, but its depth resolution is coarse with fewer steps yet. (c) Our stereo fusion significantly improves the spatial and depth resolutions by combining these two heterogeneous stereo methods.*

refractive medium placed on one of the binocular stereo cameras. In this paper, we introduce a novel optical design that combines binocular and refractive stereo and its depth process workflow that allows us to fuse heterogeneous stereo inputs seamlessly to achieve fine depth estimates. Fig. 1 shows a brief overview of our method.

2 Related Work

In this section, we briefly overview recent depth-from-stereo algorithms that are the most relevant to our work.

Multi-Baseline Stereo. Okutomi and Kanade [3] proposed a multi-baseline stereo method. The proposed system consists of multiple cameras on a rail. Gallup et al. [8] 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. Nakabo et al. [9] 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. [4] 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. We take inspiration from this work [4] 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. The camera viewpoints in the multi-baseline systems are secured mechanically at fixed locations in general. This design restricts the spatial resolution along the camera array while reconstructing the depth map. In contrast, our fusion system controls the baseline dynamically by rotating the medium. Our system requires a smaller space to operate consequently while acquiring input than the multi-baseline stereo systems.

Refractive Stereo. Nishimoto and Shirai [10] first introduced a refractive camera system by placing a refractive medium in front of a camera. 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 [11] presented a single camera system that captures a stereo pair with a bi-prism. Gao and Ahuja [5,6] proposed a seminal refractive stereo method that captures multiple refractive images with a glass medium titled at different angles. The rotation axis of the titled medium is mechanically aligned to the optical axis of the camera. Shimizu and Okutomi [12,13] 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. Chen et al. [7,14] proposed a calibration method for refractive stereo. This method finds the pairs of matching points on refractive images with the SIFT algorithm [15] to estimate the pose of a transparent medium. In this paper, we adopt an optical hardware structure of refractive stereo [6] and combine it on a binocular stereo base.

3 Light Transport in Stereo

3.1 Baseline vs. Disparity

Binocular disparity describes pixel-wise displacement of parallax between corresponding points on a pair of stereo images taken from different positions. Therefore, it is natural that computing the disparity is accompanied with searching the corresponding points on an epipolar line. As disparity d depends on its depth, we can recover a depth z as:

$$z = fb/d, \tag{1}$$

where f is the focal length of the camera lens; b is the distance between the cameras, the so-called baseline. As shown here, the disparity is proportional to the baseline, i.e., when the baseline is wide, the range of disparity increases.

Wide-baseline stereo reserves more pixels for disparity than narrow-baseline stereo does. Therefore, 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. Although the estimated disparity maps are plausible in terms of depth, but it contains many empty regions, where the depth values were not correctly estimated for occlusion and false matching. It is frequently observed in homogeneous regions and heavily-textured regions, where the corresponding point search fails.

Narrow-baseline stereo has a relatively short search range of correspondences. Hence, the false matching rarely happens so that we can achieve the accuracy and efficiency in the 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. It trades off the discriminative

power for the reduced spatial artifacts in the disparity map. 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.

3.2 Depth from Refraction

Our stereo fusion system includes a refractive stereo module on a binocular stereo architecture. 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 [6,7]. Here we briefly 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 of an objective lens C directly without any transparent medium (Fig. 2a). 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: $z = fR/r$, where r is a refractive disparity, found by searching a pair of corresponding points; f is the focal length; $R = d/\sin(\theta_p)$ is the ratio of lateral displacement d to $\sin(\theta_p)$. Here θ_p is the angle between $\overrightarrow{p_rC}$ and the image plane. $\cos(\theta_p)$ can be computed by the normalized dot product of $\overrightarrow{p_re}$ and $\overrightarrow{p_rC}$ (Fig. 2b). Lateral displacement d, the parallel-shifted length of the light passing through the medium, is determined as [16]:

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

where t is the thickness of the medium; n is the refractive index of the medium; θ_i (the incident angle of the light) can be obtained by computing $\cos(\theta_i)$ as the normalized dot product of $\overrightarrow{Cp_r}$ and $\overrightarrow{Ce}$. 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. 2b). This property can be utilized to narrow down the search range of correspondences onto the essential line, allowing us to compute matching costs

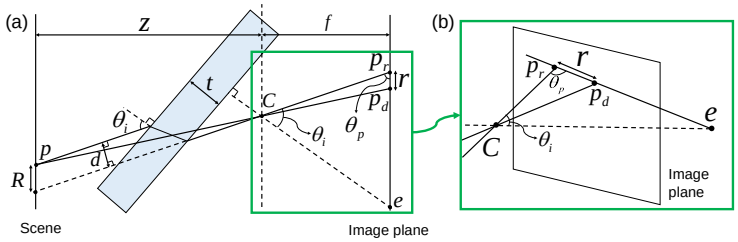


Fig. 2: *(a) Cross-section view of the light path in refractive stereo. (b) A close-up view of the refractive light transport in 3D.*

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.

4 System Implementation

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. 3 for our hardware design. Note that this refractive stereo module is equivalent to narrow-baseline stereo while the binocular stereo structure is equivalent to wide-baseline stereo in our system.

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.

4.2 Calibration

Our stereo fusion system requires several stages of calibration prior to the depth estimation. This section summarizes our calibration processes.

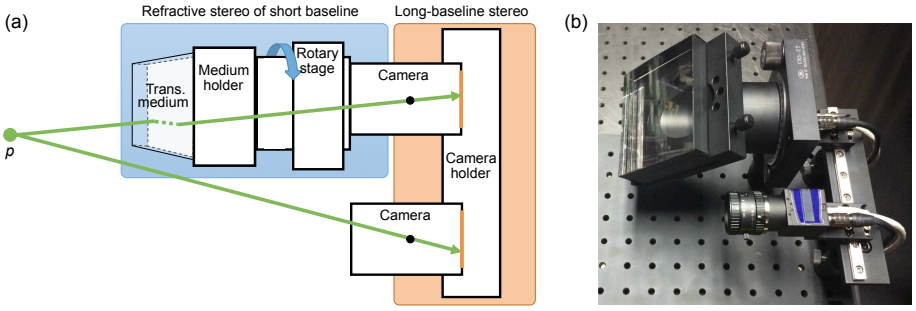


Fig. 3: *(a) The schematic diagram of our stereo fusion system. A point p is captured by both the refractive stereo and the binocular stereo module. (b) Our system prototype.*

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 to the global coordinates. This allows us to derive an affine relationship between the two cameras and rectify the coordinates of these cameras with respect to the constraint epipolar line [17].

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 a unrefracted pixel p_d and reaches the essential point e on the essential line (see Fig. 2b).

Gao and Ahuja [5,6] estimate 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. Chen et al. [7] directly estimate the essential point on target scene images with a fact that the 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 [5,6]; however, searching corresponding features with SIFT [7] is not consistent often such that the calibration accuracy is bound to the SIFT performance.

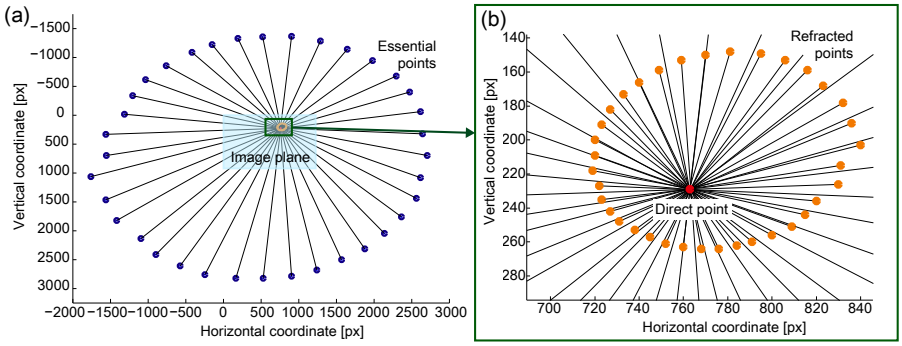


Fig. 4: *(a) presents the calibrated results of the 36 essential points (blue dots) in our system. (b) shows an example of the locations of 36 refracted points (orange dots) from a direct point (w/o the medium, red point) in the coordinates of (763,229) at a distance of 30 cm. The location of the direct point has been refracted to 36 different positions per rotation due to refraction.*

Our calibration method takes advantages of the both methods [6,7] 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 following the proposed method in [6]. 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. [7]. We then compute the arithmetic mean of the coordinates of the intersection points to determine an essential point per rotation angle. We repeat this process with the 36 rotation poses of the medium predetermined in 10-degree intervals. See Fig. 4.

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. To do so, we employ a GretagMacbeth ColorChecker target of 24 color patches. Two sets of linear RGB colors, A and B (cameras with and without the medium with inverse gamma correction), are measured from the both cameras. We determine a 3×3 affine transformation M of A to B as a camera calibration function using least-squares [18]. We apply this color transform M for the linear colors through the medium, acquiring consistent colors through the medium.

5 Depth Reconstruction in Stereo Fusion

Our stereo fusion workflow consists of two main steps. We first estimate an intermediate depth map from a set of refractive stereo images (from the camera with the medium) and reconstruct a virtual 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 overviews the workflow of our stereo fusion method.

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 [1], 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.

Matching Cost in Refractive Stereo. General binocular stereo algorithms define the *matching cost volumes* of every pixels per disparity [1], where a disparity implies a certain depth directly in binocular stereo. This relationship can be

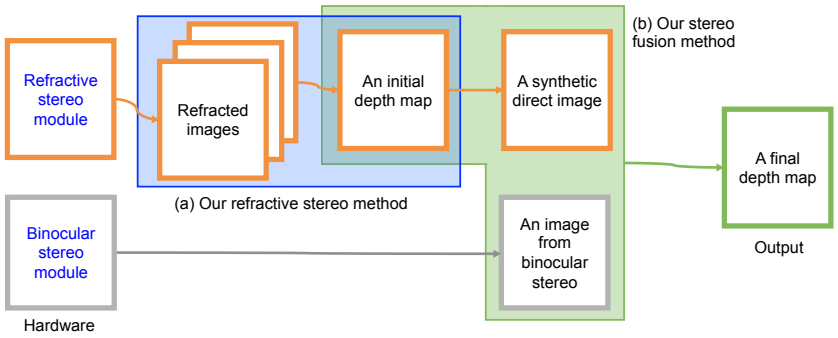


Fig. 5: *Schematic diagram of our stereo fusion method. (a) Our refractive stereo method estimates an intermediate depth map from refractive stereo. (b) Our stereo fusion method reconstructs a final depth map from a pair of an image from binocular stereo and a synthetic direct image using the intermediate depth map.*

applied for all the pixels in the stereo image uniformly. In contrast to binocular stereo, the length of disparity in refractive stereo changes 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.

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) with an essential point e ($e \in E$): $P(p_d, z) = \{p_r(p_d, z, e) | e \in E\}$. This set P can be derived analytically by refractive calibration (Sec. 3.2) so that we precompute this set P for computational efficiency.

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. We use the similarity of $L(p_d, z)$ for the matching cost C of p_d with a hypothetical depth z [19]:

$$C(p_d, z) = \frac{1}{|L(p_d, z)|} \sum_{l \in L(p_d, z)} K(l - \bar{l}), \tag{3}$$

where K is an Epanechnikov kernel [20]: $K(l) = 1 - \|l/h\|^2$ when $\|l/h\|$ is less than or equal to 1; otherwise, $K(l) = 0$; h is a normalization constant ($h = 0.01$); $\bar{l}$ is computed iteratively in $L(p_d, z)$ (for five iterations) using the mean shift method [21] as:

$$\bar{l} = \sum_{l \in L(p_d, z)} K(l - \bar{l}) l \Big/ \sum_{l \in L(p_d, z)} K(l - \bar{l}) \, . \tag{4}$$

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.

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. An aggregated cost function C^A is defined as $C(p_d, z)$ convolved by a weighting factor G [22]:

$$C^A(p_d, z) = \sum_{q_d \in w} G(p_d, q_d) C(q_d, z), \tag{5}$$

where q_d is a pixel inside a squared window w, of which size is 7×7. We aggregate the refractive matching costs using a Gaussian kernel $G(p_d, q_d)$, defined as $(1/2\pi\sigma^2)\exp(-(||p_d - q_d||^2)/2\sigma^2)$, where σ is 9.6. Finally, we compute the optimal depth $Z(p_d)$ of the point p_d that maximizes the aggregated matching costs: $Z(p_d) = \arg\max_z C^A(p_d, z)$.

Depth and Direct Image Refinement. 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. (4)).

Reconstructing the direct image allows us to apply a depth refinement algorithm with a weighted median filter [23] by treating the 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 $h(p_d, \cdot)$:

$$h(p_d, z) = \sum_{q_d \in w} W(p_d, q_d) f(q_d, z), \tag{6}$$

where $f(q_d, z)$ is defined as 1 when $Z(q_d) - z$ is 0; otherwise, $f()$ is 0. W is a weight function with a guided image filter [24], defined as $W(p_d, q_d) = \frac{1}{|w|^2} \sum_{k:(p_d,q_d)\in w_k} (1 + (l_d(p_d) - \mu_k)(\Sigma_k + \epsilon U)^{-1}(l_d(q_d) - \mu_k))$, 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; 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. 6 shows the result of the refinement, which are the updated depth map and the direct image.

Optimal Number of Refractive Images. The number of input refractive images is critical to the quality of depth estimation in refractive stereo. We were motivated to find out an optimal number of input refractive images so that we evaluated point-wise errors in estimating depth on a planar surface in a known distance. Fig. 7b shows that the root-mean-square error decreases very fast while increasing the number of input up to six refractive images. Hence, we choose the optimal number of input refractive images as six. Note that we use six refractive images with 60° intervals for capturing results in this paper.

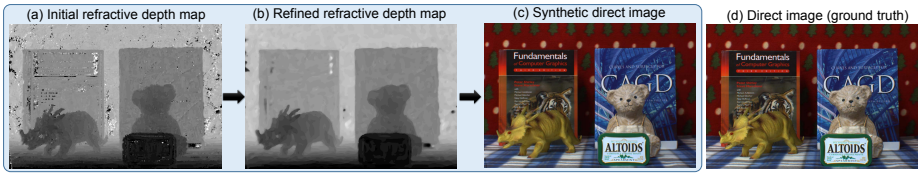


Fig. 6: *(a) shows an initial depth map and (b) is the refined map with weighted median filtering. A refracted image is computed as a synthetic direct image (c), used for binocular stereo later. (c) is compared to a ground truth photograph (d).*

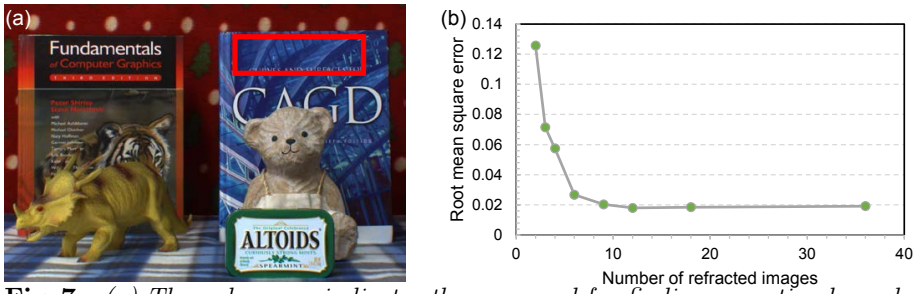


Fig. 7: *(a) The red square indicates the area used for finding an optimal number of input refractive images. The book cover is a planar surface orthogonal to the camera optical axis with a constant depth. (b) The depth error drops down fast significantly up to six refractive inputs with different angles. No significant improvement is observed with more than six inputs.*

5.2 Depth in Stereo Fusion

As described in Sec 3.1, 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 [25,26] 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.

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) 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: $F(p_d, d) = \|l_d(p_d) - l_b(p'_d)\|$, 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.

Cost Aggregation in Stereo Fusion. The aggregated cost of the fusion matching costs is defined as:

$$F^A(p_d, d) = \sum_{q_d \in w} W(p_d, q_d) F(q_d, d). \tag{7}$$

Here W is the bilateral image filter [22] defined as $W(p_d, q_d) = exp(-(d(p_d, q_d)/\sigma_s^2 + c(p_d, q_d)/\sigma_c^2)$, 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.(1). Note that d_{post} is 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:

$$D(p_d) = \arg\min_d F^A(p_d, d). \tag{8}$$

We compute Eq.(7) within the range of $[d_{post}, d_{prev}]$ exclusively for computational efficiency.

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.

The first row in Fig. 9 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)

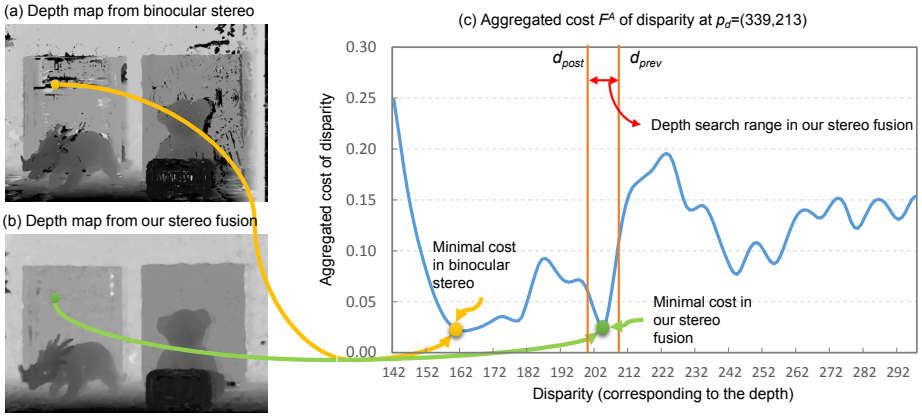
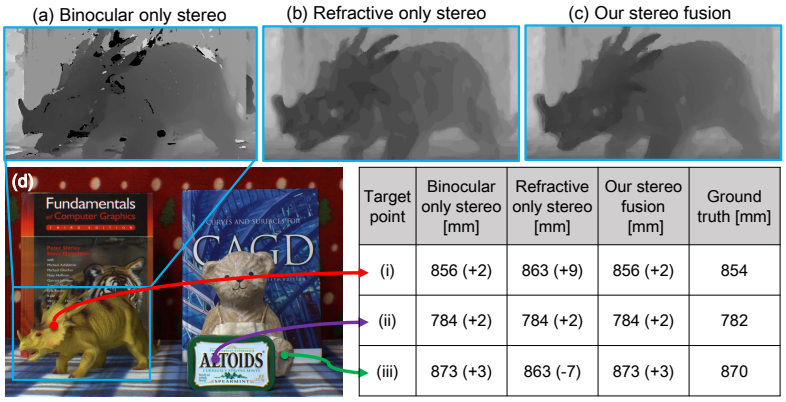


Fig. 8: *The binocular depth map (a) includes artifacts due to false matching caused by occlusions, featureless regions and repeated patterns. Using the intermediate refractive depth map (b), we can limit the search range of a corresponding point p_d between d_{post} and d_{prev} for instance. This significantly reduces false matching frequency in estimating depth.*

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 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. 9d. 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).

We compare our proposed method with a renowned graphcut-based algorithm [27] 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 Fig. 10a). We also compare our method with a local binocular method [24], which computes the matching cost as the norm of intensity difference and aggregates the cost using the weight of the guided filter [24]. Its computing time was ∼212 secs. with the same scene (see Fig. 10b). This local method struggles with typical false matching artifacts. A refractive method using SIFT flow [7] is compared to ours (Figs. 10c and 10d). The same number of six refractive images were employed for both methods. While the refractive method suffers from wavy artifacts of 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.



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

Fig. 9: *The top row compares the three different depth maps of binocular only stereo (a), refractive only stereo (b) and our stereo fusion (c). (d) quantitatively compares the accuracy of three methods. Three points were measured by a Bosch laser distance meter. The accuracy of our stereo fusion is as high as binocular stereo; our depth map does not suffer from artifacts as refractive stereo.*

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. Fig. 11a and Fig. 11b presents results of trinocular stereo. Fig. 11a is the result of a multi-baseline method [3], 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. Fig. 11b 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 the search range of correspondence same as ours. As shown in Fig. 11, 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.

7 Conclusion

We have presented a novel stereo fusion method with an optical design and stereo fusion algorithm. We validate that our proposed method takes the advantages of both traditional binocular and refractive stereo. In addition, our fusion design can be easily integrated into any existing binocular stereo, yielding a significant improvement in depth accuracy.

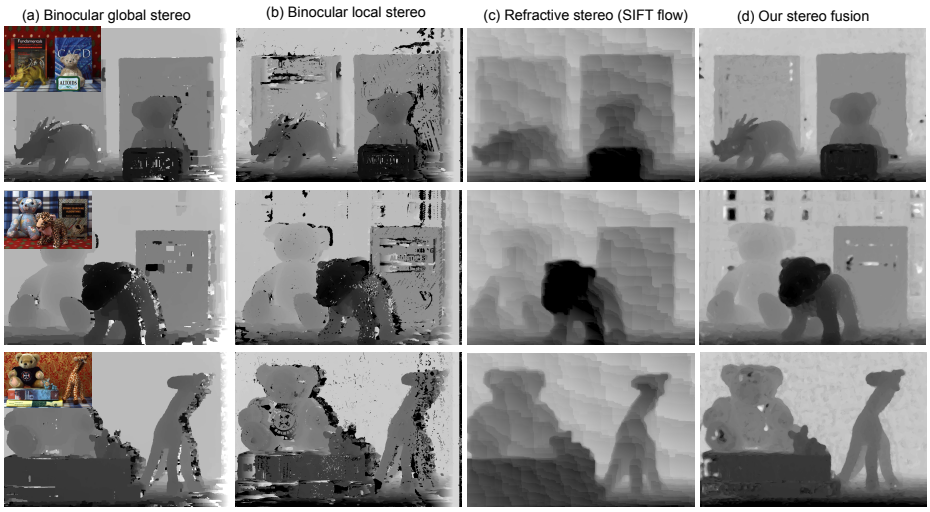


Fig. 10: *The depth maps of three different scenes in each row were computed by four different methods. The first two columns (a) and (b) show global [27] and local binocular stereo [24] methods. The third column (c) presents a refractive stereo method [7]. Our method (d) estimates depth accurately without suffering from severe artifacts.*

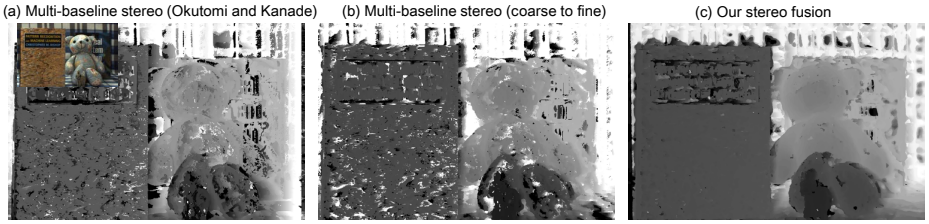


Fig. 11: *Multi-baseline stereo methods are compared with our method. (a) is a depth map using a trinocular stereo method [3]. (b) is also a trinocular stereo method, implemented with the coarse-to-fine approach same as ours. (c) is the result of our stereo fusion method with the same number of input images.*

However, there are some issues remained as future work. Our current method requires the rotation of the refractive module at least more than one time for a depth estimate. Therefore, our method is only allowed for static scenes. Suppose a depth map from refractive stereo contains errors, these errors remain from the stereo fusion stage, where we refine the depth resolution of the refractive depth map using binocular stereo.

Acknowledgement. Min H. Kim gratefully acknowledges Korea NRF grants (2013R1A1A1010165 and 2013M3A6A6073718) and additional support by Microsoft Research Asia and an ICT R&D program of MSIP/IITP (10041313).

References

1. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vision (IJCV) **47** (2002) 7–42
2. Rhemann, C., Hosni, A., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2011) 3017–3024
3. Okutomi, M., Kanade, T.: A multiple-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **15** (1993) 353–363
4. Zilly, F., Riechert, C., Mller, M., Eisert, P., Sikora, T., Kauff, P.: Real-time generation of multi-view video plus depth content using mixed narrow and wide baseline. J. Visual Communication and Image Representation **25** (2013) 632–648
5. Gao, C., Ahuja, N.: Single camera stereo using planar parallel plate. In: Proc. Int. Conf. Pattern Recognition (ICPR). Volume 4. (2004) 108–111
6. Gao, C., Ahuja, N.: A refractive camera for acquiring stereo and super-resolution images. In: Proc. Comput. Vision and Pattern Recognition (CVPR). (2006) 2316–2323
7. Chen, Z., Wong, K.Y.K., Matsushita, Y., Zhu, X.: Depth from refraction using a transparent medium with unknown pose and refractive index. Int. J. Comput. Vision (ICJV) (2013) 1–15
8. Gallup, D., Frahm, J.M., Mordohai, P., Pollefeys, M.: Variable baseline/resolution stereo. In: Proc. Comput. Vision and Pattern Recognition (CVPR). (2008) 1–8
9. Nakabo, Y., Mukai, T., Hattori, Y., Takeuchi, Y., Ohnishi, N.: Variable baseline stereo tracking vision system using high-speed linear slider. In: Proc. Int. Conf. on Robotics and Automation (ICRA). (2005) 1567–1572
10. Nishimoto, Y., Shirai, Y.: A feature-based stereo model using small disparities. In: Proc. Comput. Vision and Pattern Recognition (CVPR). (1987) 192–196
11. Lee, D., Kweon, I.: A novel stereo camera system by a biprism. IEEE Trans. Robotics and Automation **16** (2000) 528–541
12. Shimizu, M., Okutomi, M.: Reflection stereo-novel monocular stereo using a transparent plate. In: Proc. Canadian Conf. Computer and Robot Vision (CRV), IEEE (2006) 14–14
13. Shimizu, M., Okutomi, M.: Monocular range estimation through a double-sided half-mirror plate. In: Proc. Canadian Conf. Computer and Robot Vision (CRV), IEEE (2007) 347–354
14. Chen, Z., Wong, K., Matsushita, Y., Zhu, X., Liu, M.: Self-calibrating depth from refraction. In: Proc. Int. Conf. Comput. Vision (ICCV). (2011) 635–642
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision (IJCV) **60** (2004) 91–110
16. Hecht, E.: Optics. Addison-Wesley, Reading, Mass (1987)
17. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **22** (2000) 1330–1334
18. Kim, M.H., Kautz, J.: Characterization for high dynamic range imaging. Computer Graphics Forum (Proc. EUROGRAPHICS 2008) **27** (2008) 691–697
19. Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., Gross, M.: Scene reconstruction from high spatio-angular resolution light fields. ACM Trans. Graph. (TOG) **32** (2013) 73:1–12
20. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley (1973)
21. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **24** (2002) 603–619

22. Yoon, K.J., Kweon, I.S.: Adaptive support-weight approach for correspondence search. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **28** (2006) 650–656
23. Ma, Z., He, K., Wei, Y., Sun, J., Wu, E.: Constant time weighted median filtering for stereo matching and beyond. In: Proc. Int. Conf. Computer Vision (ICCV). (2013) 1–8
24. He, K., Sun, J., Tang, X.: Guided image filtering. In: Proc. European Conference on Computer Vision (ECCV), Springer (2010) 1–14
25. Barnard, S.T.: Stochastic stereo matching over scale. Int. J. Comput. Vision (IJCV) **3** (1989) 17–32
26. Chen, J.S., Medioni, G.: Parallel multiscale stereo matching using adaptive smoothing. In: Proc. European Conference on Computer Vision (ECCV). (1990) 99–103
27. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **23** (2001) 1222–1239