Advanced Robotics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tadr20

Trinocular 360-degree stereo for accurate all-round 3D reconstruction considering uncertainty

Sarthak Pathak, Takumi Hamada & Kazunori Umeda

To cite this article: Sarthak Pathak, Takumi Hamada & Kazunori Umeda (2024) Trinocular 360-degree stereo for accurate all-round 3D reconstruction considering uncertainty, Advanced Robotics, 38:15, 1038-1051, DOI: 10.1080/01691864.2024.2376022

To link to this article: https://doi.org/10.1080/01691864.2024.2376022

ADVANCED ROBOTICS
2024, VOL. 38, NO. 15, 1038–1051
https://doi.org/10.1080/01691864.2024.2376022

FULL PAPER

Trinocular 360-degree stereo for accurate all-round 3D reconstruction considering uncertainty

Sarthak Pathak, Takumi Hamada and Kazunori Umeda

Course of Precision Engineering, Faculty of Science and Engineering, Chuo University, Hachioji, Japan

ABSTRACT

This research proposes a method for accurate all-round 3D reconstruction of an indoor environment in one-shot using a system of trinocular 360-degree cameras. Binocular 360-degree stereo is unable to reconstruct in all directions due to lack of disparity along epipolar directions. Thus, a third camera along a perpendicular epipolar direction is introduced to cover for this, making the system trinocular. However, previous works with trinocular stereo did not adequately take into account the uncertainty of disparity estimation and geometric constraints around 3D reconstruction. Therefore, we propose a geometric optimization scheme considering disparity estimation uncertainty and show that this results in both higher accuracy and lesser outliers along epipolar directions, in both simulated and real environments.

ARTICLE HISTORY
Received 9 November 2023
Revised 2 May 2024
Accepted 14 June 2024

KEYWORDS
All-round 3D reconstruction; omnidirectional; stereo; trinocular

1. Introduction

Cameras are integral sensors for robotic systems. One of the main tasks they are used for is 3D reconstruction of environments. Obtaining a full color 3D reconstruction of an environment is important for disaster response, digitization, measurement, and inspection. Binocular stereo vision is a common technique for camera-based 3D reconstruction. As opposed to Structure from Motion (SfM) [1], camera positions are known in binocular stereo allowing for denser, more accurate reconstruction in a single capture from both cameras. However, in order to capture the environment in all directions, it is necessary to move the binocular stereo camera around the environment and fuse multiple 3D reconstructions together. This requires capture time, and the fusion is not always accurate. In contrast, 360-degree cameras can capture all direction in a single shot. Accordingly, it is expected that binocular 360-degree stereo [2] should be able to achieve an all-round 3D reconstruction. However, this is not possible because 3D reconstruction requires disparity i.e. a difference in position of corresponding pixels. As we approach the epipolar direction, i.e. the direction joining both cameras, the disparity approaches zero. Due to this, reconstruction around the epipolar directions is impossible. Moreover, close to the epipolar directions, the uncertainty of disparity estimation becomes quite large as compared to the disparity, leading to a loss of accuracy. As shown below in Figure 1, each pixel has an uncertainty in disparity estimation. (For simplicity, the figure shows equal uncertainty for all pixels.) This uncertainty translates to uncertainty of the 3D point position when reconstructed in space. The area inside the parallelogram represents the possible locations of the 3D point. As can be seen, this area increases greatly as the epipolar direction is approached, making 3D reconstruction difficult, leading to outlier points. Thus, binocular 360-degree vision cannot achieve all-round 3D reconstruction.

To deal with this problem, [3] introduced the concept of adding an additional camera at a direction perpendicular to the epipolar direction, making an L-shaped arrangement. The additional camera can provide a new epipolar direction that can cover for the lack of accuracy from the original. Basically, this system forms two binocular 360-degree stereo systems that are perpendicular to each other. They show that the binocular reconstruction results in huge distortions in the epipolar direction and their trinocular approach solves this to an extent. However, this approach did not attempt proper information fusion from both. They attempted a simple weighted average of both directions. In case the distortion from one direction is too high, it can show up in the final result. This approach did not attempt to preserve the geometric constraint that each reconstructed point occupies only one position in 3D space and must be projected *consistently* in all cameras, i.e. the position of the reconstructed

CONTACT Sarthak Pathak pathak@mech.chuo-u.ac.jp

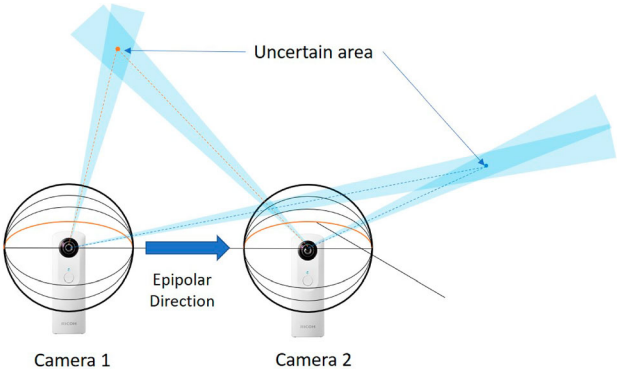


Figure 1. 3D reconstruction uncertainty in binocular 360-degree stereo vision. The camera shown in the figure is the Ricoh Theta. 360-degree images can be represented as unit spheres, as shown in the figure.

point on each image should match the pixel. Yin et al. [4] improved this result by minimizing the error of each 3D point across all images. However, the uncertainty of disparity estimation was not taken into account. Moreover, they were unable to achieve reconstruction in real environments.

The reason for the lack of accuracy in previous methods is that disparity uncertainty is difficult to calculate because the distortion present in 360-degree images makes it so that corresponding pixels lie on complex epipolar curves. Typical stereo disparity estimation methods [5] employ a tactic known as stereo rectification, in which all corresponding pixels are brought to the same horizontal image coordinate, greatly reducing the problem. Disparity uncertainty can then be estimated across the same horizontal coordinate. However, this does not work for 360-degree images, which are distorted and do not have epipolar lines. They have epipolar curves, instead, as shown Figure 2.

In this paper, we maximize reconstruction accuracy by taking disparity uncertainty into account via geometric optimization. We calculate uncertainty based on a rectification of 360-degree images in the vertical direction,

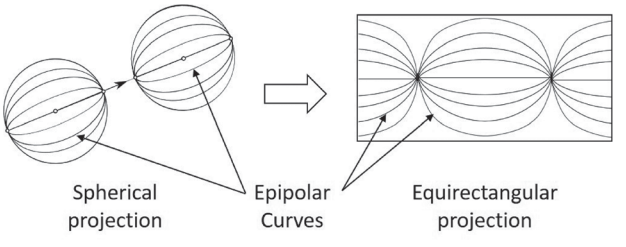


Figure 2. Epipolar lines in 360-degree images are shown on the left in spherical projection (unit sphere projection). On expanding these to equirectangular images, they take on a complicated shape making disparity estimation difficult.

that makes it easy to calculate the disparity uncertainty. We show that this results in higher accuracy as well as low distortion along the epipolar directions, minimizing the number of erroneously reconstructed 'outlier' points. In the next section, we discuss closely related literature. After that, we explain the concept of the proposed geometric optimization, diving disparity uncertainty. Following that, we show experimental evaluation, both qualitative and quantitative, in real and rendered environments. Finally, we conclude the paper and talk about future work.

2. Related work

All-round 3D reconstruction, especially for indoor environments, has been paid attention in literature in several ways. Most common methods involve using an active sensor, such as an RGB-D camera and moving it around the environment, scanning each part [6, 7]. However, such methods require alignment of various 3D scans which can induce drift errors in the final measurement. They also require time to capture the entire environment and are not suitable for real-time operation. The Velodyne LiDAR and other 360-degree LiDARs can provide real-time all-round 3D data. However, they lack color information. This has been solved by methods such as [8], which combine 360-degree cameras and 360-degree LiDARs. This forms a good solution, but is quite expensive and power hungry. Moreover, the LiDAR cannot provide the same high-density as a camera can. Other sensors like the FARO Focus [9] can provide dense measurements and color information, leading to successful use in infrastructure inspection and other practical applications. The sensor rig rotates around a single point, leading to minimal error in registering different scans. However, this is prohibitively expensive and also takes time to scan the entire structure. As opposed to all these sensors, using a pure camera based solution is attractive.

As opposed to regular perspective cameras, 360-degree cameras that can capture all-round RGB information in a single shot are suitable for this.

3D reconstruction using 360-degree cameras has been tackled in previous research. Most recent approaches such as Pano3D [10] use deep learning-based solutions for single-view 3D reconstruction. However, learning-based methods are not geometric in nature, which makes it difficult to preserve the shape, detail, and geometric scale of the environment. Thus, we choose to focus on geometric methods in this research.

As mentioned earlier, [2, 11] used binocular 360-degree stereo systems for 3D reconstruction. However, they are only applicable to two 360-degree cameras in a binocular stereo arrangement. Due to the uncertainty

difference based on angle as shown in Figure 2, 3D points located in the epipolar directions cannot be reconstructed. Moreover, [2] uses two cameras displaced in the vertical direction. Their method assumes vertical displacement. In comparison, our proposed method can be implemented with any camera arrangement via the use of the vertically rectified orientation. Since 360-degree images can be rotated to any orientation, there is no limitation on the arrangement of cameras.

Yin et al. [4] and Li [3] are applicable to three cameras and attempt to reconstruct in the epipolar directions as well. They assume that the estimated disparity is correct and do not consider the uncertainties in the epipolar directions. However, it is important to consider these uncertainties as both epipolar directions in a three camera setup are different and lead to different uncertainties of estimation. As a result, [3] did not achieve accurate reconstruction and [4] was unable to achieve performance in real environments. Our proposed method considers the uncertainties in both epipolar directions individually by use of vertical rectification.

These prior studies revealed that properly considering uncertainty is critical in fusing the information obtained from all three 360-degree cameras. As mentioned, typical stereo estimation methods for perspective cameras such as [5] are able to estimate disparity uncertainty along epipolar lines. However, this does not work for 360-degree images which have distorted epipolar curves, as shown Figure 2. When expanded to the equirectangular projection, which is commonly used to process 360-degree images, the curves take a complicated shape that makes disparity estimation difficult.

To address this issue, this paper proposes calculating uncertainty on a rectified vertical alignment of 360-degree image pairs, followed by an uncertainty based geometric optimization technique. Experimental results show superior accuracy and lower outliers in the epipolar directions. We also achieve accurate all-round 3D reconstruction in a cluttered, real environment, showing the superiority of our proposed approach.

3. Geometric optimization considering uncertainty

3.1. Setup and system overview

Our proposed method uses the same camera setup as [3]. This setup consists of three 360-degree cameras in an L-shaped arrangement as shown in Figure 3. There is no limitation on the orientations and positions of the cameras in our method as long as both epipolar directions are approximately perpendicular to each other. The proposed vertical rectification scheme can deal with any

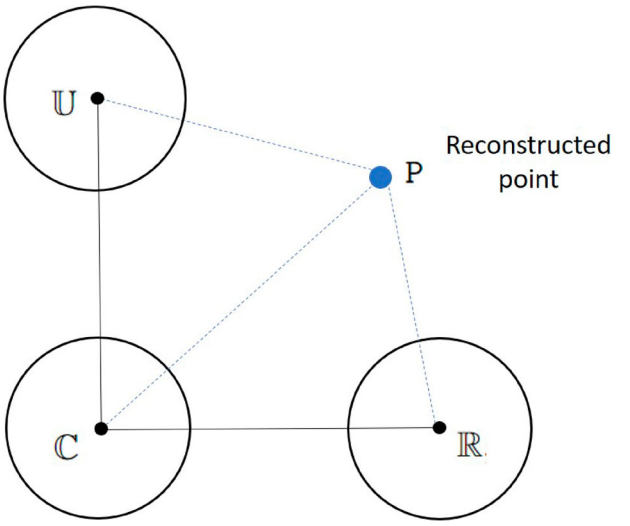


Figure 3. Our setup use three cameras in an L-shaped arrangement to form a trinocular system.

arrangement of cameras. Camera $\mathbb{C}$ is the central camera and is taken to be the origin of the coordinate system. The upper camera is Camera $\mathbb{U}$ and the camera on the right is referred to as Camera $\mathbb{R}$. The distances between Cameras $\mathbb{C}$ and $\mathbb{R}$ is the same as the distance between Camera $\mathbb{C}$ and Camera $\mathbb{U}$, i.e. they have equal baseline. The baseline can be adjusted depending on the target environment. All relative camera positions are assumed to be known/calibrated. The system is, essentially, a fusion of two binocular 360-degree stereo systems using Cameras $\mathbb{C}$ and $\mathbb{R}$, and $\mathbb{U}$ and $\mathbb{C}$.

The 3D reconstruction process proceeds as follows.

(1) First, all cameras capture the environment simultaneously.
(2) All images are rectified based in a vertical alignment, based on the known camera positions. This makes disparity and uncertainty calculation possible.
(3) Initial disparity maps between Camera pair $\mathbb{C}$ and $\mathbb{R}$, and camera pair $\mathbb{U}$ and $\mathbb{C}$ are calculated.
(4) For each point in the image captured by Camera C, distances d_{cu} and d_{uc} are triangulated using the principle of binocular stereo [2] between $\mathbb{C}$ and R, $\mathbb{U}$ and $\mathbb{C}$, respectively. Even though the same pixel in Camera$\mathbb{C}$ is triangulated in both cases, the distances will not be equal due to the effects of pixel and disparity uncertainty, and noise.
(5) Disparity uncertainty for each pixel is calculated
(6) The distance of each pixel in the image captured by Camera $\mathbb{C}$ is optimized based on the calculated uncertainty. The geometric constraint that each pixel in Camera $\mathbb{C}$ can only occupy one position in 3D space is applied and optimized. The initial guess for

the optimization is calculated as the average of the distances d_{cr} and d_{uc}.

The next few subsections will explain all the points in detail. (5) and (6) form the heart of our proposed method.

3.2. Image capture

Image capture is the first step of our proposed method. All cameras, $\mathbb{C}$, $\mathbb{U}$, and $\mathbb{R}$ capture images simultaneously (We use the same notations $\mathbb{C}$, $\mathbb{U}$, and $\mathbb{R}$ interchangably for 'Cameras' and 'Images'.). All camera positions are known or calibrated beforehand. It is important to ensure that all cameras capture under the same conditions. In real environments, each camera will also end up capturing the camera rig as well as other cameras. Since the camera positions are known, these areas are constant in all images and can be masked out beforehand. The output images are three equirectangular images in a 2:1 aspect ratio. Since most 360-degree cameras consist of two oppositely faced fisheye lenses, some may give the raw fisheye images as output. These images can be calibrated and fused to give full 360-degree images using [12].

3.3. Vertical rectification and disparity estimation

The next step in our proposed method is initial disparity estimation between $\mathbb{C}$ and $\mathbb{R}$, and $\mathbb{C}$ and $\mathbb{U}$. This disparity estimation provides an initial anchor value for the geometric optimization based on uncertainty. Typical stereo disparity estimation for perspective stereo cameras is done after stereo rectification i.e. making all the epipolar lines parallel to each other. However, as mentioned earlier and as shown in Figure 2, epipolar lines in 360-degree equirectangular images are not straight lines, but curves. However, in a special case, if two camera are displaced vertically, the epipolar lines are aligned from top to bottom and become perfectly vertical. This concept was used by Kim and Hilton [11] where they mechanically aligned two 360-degree cameras to be vertically displaced. However, this is not always possible in practice. Even if one of the cameras optical axes are slightly tilted, it can induce a large error in this alignment. Instead, we use the principle that 360-degree images contain information from all directions and can be rotated freely. Based on known camera positions and orientations, rotations can be devised in order to convert any arrangement of cameras to a vertically displaced arrangement by rotation. Earlier, this rectification was applied to binocular 3D reconstruction in [13]. Essentially, the rectification proceeds as follows:

(1) Initially, both cameras are displaced in a known arbitrary direction and have known arbitrary rotation between them. If unknown, the cameras are calibrated to find these values.
(2) One of the images is rotated to bring it to the same orientation as the other.
(3) Both images are then collectively rotated to make the displacement direction completely vertical.
(4) The images are unwarped to the equirectangular projection to make the epipolar lines vertical.

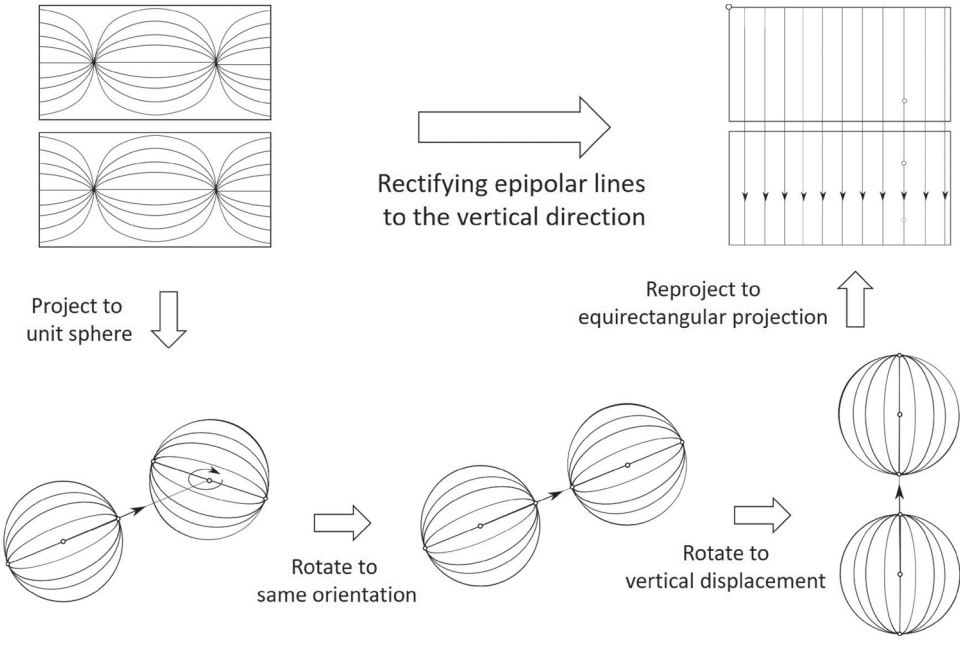


Figure 4. Rectification of arbitrarily displaced equirectangular images via multiple rotations to make the epipolar lines vertically straight.

The rectification scheme is shown in Figure 4. This rectification makes the epipolar lines vertical and makes it easy to estimate disparity. This rectification is also important for calculating disparity uncertainty, as now it can be calculated along straight epipolar lines. We apply this rectification to both image pairs, $\mathbb{C}$ and $\mathbb{R}$, and $\mathbb{U}$ and $\mathbb{C}$. Image $\mathbb{C}$, which is common to both pairs, takes two different orientations for each pair.

Once the vertical rectification is performed, the epipolar lines are formed vertically for each pair of rectified cameras. Now, disparity can be estimated using methods that are applicable to perspective camera stereo images. Our proposed method focuses on obtaining the best geometric estimate of depth, per-pixel, given a particular disparity estimate and its uncertainty. While our proposed method is applicable to any disparity estimation method, the choice of the disparity estimation method is important. Considering this, we chose to use Deepflow [14].

3.4. Initial depth estimation

Our proposed optimization (explained in Section 3.5) requires an initial value of depth for each pixel in Image $\mathbb{C}$, which is taken to be the base image and the origin of the coordinate system. In order to do that, we calculate depth values from both binocular pairs, $\mathbb{C}$ and $\mathbb{R}$, and $\mathbb{U}$ and $\mathbb{C}$.

In Section 3.3, both pairs were rectified and disparity was estimated. In this subsection, we triangulate the depth of each pixel in image $\mathbb{C}$ using both pairs to obtain distances d_{CR} and d_{UC}. The triangulation is done in the same manner as in [11, 13]. Both d_{CR} and d_{UC} are calculated from the point-of-view of Camera $\mathbb{C}$.

As a result, each pixel in image $\mathbb{C}$ has two different estimates of depth: d_{CR} and d_{UC}. As an initial guess, we simply take the average of d_{CR} and d_{UC} and call it d_{avg}. We call this the 'unoptimized' depth value and this forms a baseline for how much the performance improves after optimization. In the next section, we come to the main contribution of our work, i.e. the geometric optimization considering uncertainty.

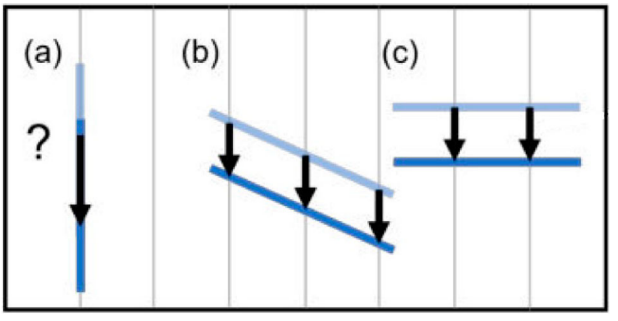


Figure 5. In order to estimate disparity, it is important to have texture perpendicular to the epipolar lines. In case of (a), texture is present but it is along the epipolar line, making it difficult to estimate disparity.

3.5. Geometric optimization considering uncertainty

In Section 3.4, initial depth for every pixel in image $\mathbb{C}$ was estimated after rectification and disparity estimation. However, every pixel in $\mathbb{C}$ has two different depth values from the two binocular systems. In this chapter, we attempt to obtain the best, geometrically consistent estimate of depth for every pixel in $\mathbb{C}$.

3.5.1. Overview

Li [3] calculated the final depth using a weighted average of both depth values obtained for the central image (Image $\mathbb{C}$, in our case). However, this violates the fundamental principle of camera-based 3D reconstruction – that every 3D pixel, when projected on an image, must coincide with the pixel it came from. Due to noise and incorrect disparity estimation, this is not strictly true. However, it can be said that the best guess of the correct position of each 3D point is that which minimizes the error between the projected 3D point and the pixel

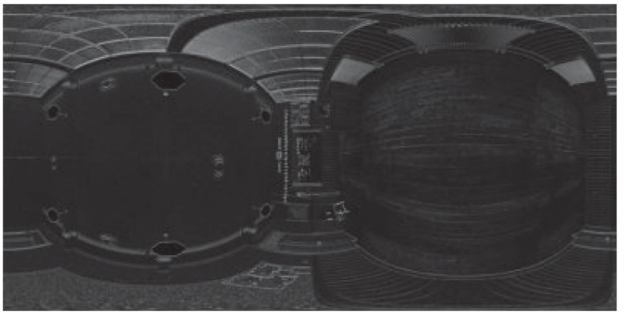


Figure 6. Calculation of Sobel filters in the horizontal direction of a rectified equirectangular image. The example is taken from one of the experiments in Section 4.

Figure 7. A virtual environment consisting of a classroom scene was chosen as our simulated experiment.

from which it came. We apply this principle to estimate the correct depth of each pixel. Further, we apply the geometric constraint that each 3D point should occupy only one position in 3D space. We decide to reproject each 3D point from image $\mathbb{C}$ to the other two images – $\mathbb{R}$ and $\mathbb{U}$, and minimize the distances to the estimates of the pixel positions obtained from disparity. This is possible because the corresponding position of each pixel in image $\mathbb{C}$ is known in image $\mathbb{R}$ and image $\mathbb{U}$ due to the disparity calculated in Section 3.3. The corresponding positions are found by going back to the rectified images and displacing the pixel by the disparity values in each image.

Thus, for each pixel $\hat{\mathbf{u}}$ in Image $\mathbb{C}$, we pose the Geometric optimization as follows:

(1) With the current depth estimate d of the pixel $\hat{\mathbf{u}}$, we obtain the 3D point P

(a) Image $\mathbb{U}$ (b) Image $\mathbb{C}$

(c) Image $\mathbb{R}$

Figure 8. Images captured in the simulated environment. (a) Image $\mathbb{U}$, (b) Image $\mathbb{C}$ and (c) Image $\mathbb{R}$.

(a) Experimental environment (b) Ricoh Theta Z1

Figure 9. The real environment chosen for experiments. (a) Experimental environment and (b) Ricoh Theta Z1.

(2) We reproject $\mathbb{P}$ to Image $\mathbb{R}$ and Image $\mathbb{U}$ and obtain the reprojected image points $\hat{\mathbf{p}}_R$ and $\hat{\mathbf{p}}_U$, respectively.
(3) Using the disparity values calculated in Section 3.4 we back-calculate the rectification and find its ideal position in Image $\mathbb{R}$ and Image $\mathbb{U}$ as $\hat{\mathbf{u}}_R$ and $\hat{\mathbf{u}}_U$.
(4) The best depth estimate d of pixel $\hat{\mathbf{u}}$ is the one that minimizes the distance between $\hat{\mathbf{u}}_R$ and $\hat{\mathbf{p}}_R$, and $\hat{\mathbf{u}}_U$ and $\hat{\mathbf{p}}_U$, respectively. This is found by minimization.

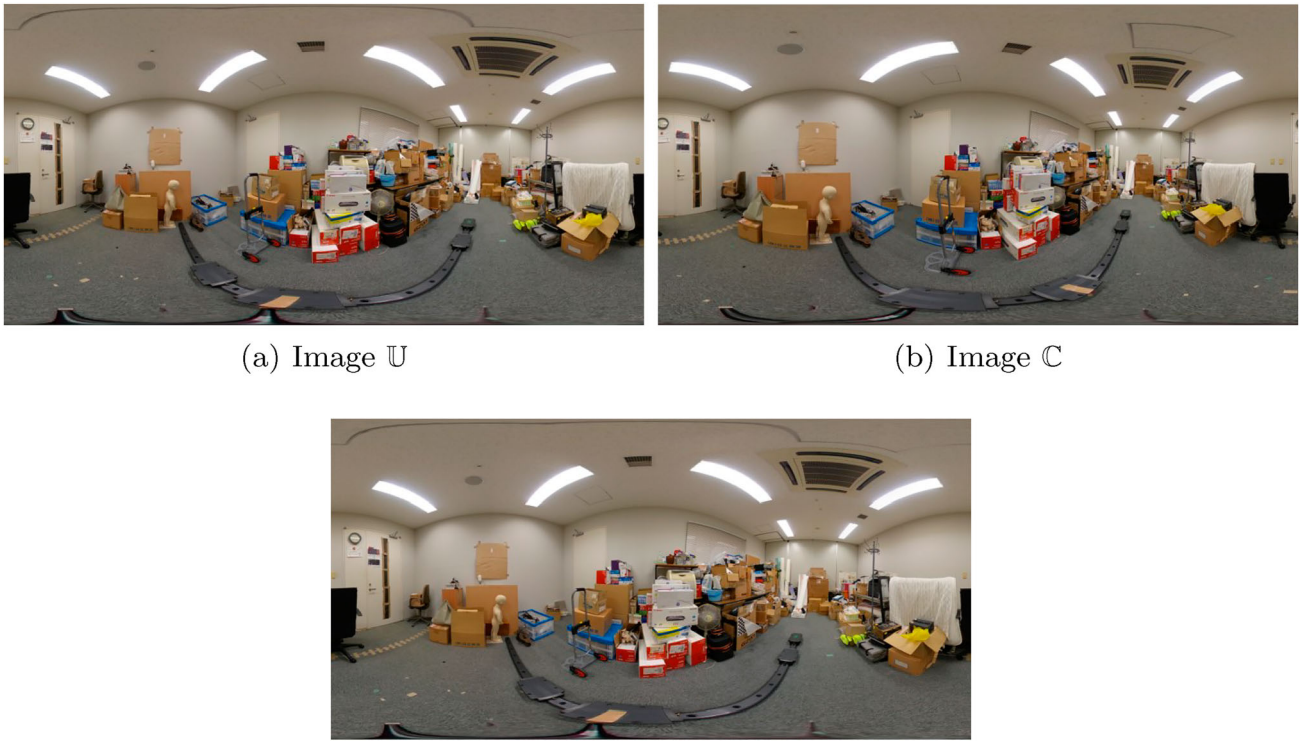
The minimization in (4) should take into consideration the disparity uncertainty of each pixel. This is described next.

3.5.2. Calculation of uncertainty and posing the optimization problem

In order to maximize accuracy, we take disparity uncertainty into account. Disparity uncertainty arises due to incorrect estimation of disparity. This is dependant on several factors such as the presence of texture, lighting condition differences between the two images, image noise, etc. The most important of these is the presence of image textures. Most disparity estimation methods calculate disparity by comparing local image information along epipolar lines. Textureless regions have very little information, making them uncertain to match. However, the presence of textures is not enough. The textures must be present in a way to elicit information *along epipolar lines*. Since epipolar lines for equirectangular images are complicated curves as shown in Figure 2, calculating the texture information is quite difficult.

In order to solve this issue, we use the vertical rectification described in Section 3.4. Once the epipolar lines are rectified to a vertical direction, we check for texture along them. As shown in Figure 5, texture must be present perpendicular to the rectified epipolar line in order to enable reliable calculation of disparity. We choose to quantify this uncertainty via a Sobel filter. In the rectified state, we choose the horizontal direction of the sobel filter as a measure of the 'certainty' of disparity estimation. Figure 6 shows an example of calculating the Sobel filter in rectified state of a 360-degree equirectangular image. Since there are two rectified states, one for each binocular pair $\mathbb{C}$–$\mathbb{R}$ and $\mathbb{U}$–$\mathbb{C}$, two Sobel filter calculations of Image $\mathbb{C}$ are done.

In this research, we chose to estimate uncertainty in the rectified state using a Sobel filter. The Sobel filter can be replaced with any filter that is able to capture the uncertainty of disparity estimation. The core of our method is in considering the geometric uncertainty of disparity estimation in the epipolar direction in order



(a) Image $\mathbb{U}$ (b) Image $\mathbb{C}$

(c) Image $\mathbb{R}$

Figure 10. Images captured in the real environment. (a) Image $\mathbb{U}$, (b) Image $\mathbb{C}$ and (c) Image $\mathbb{R}$.

to produce the best possible estimate of distance for each pixel.

The choice of the filter depends on whether its response can capture the uncertainty of the disparity estimation. For Deepflow [14], a Sobel filter that approximates the local gradient is suitable. This is because Deepflow [14] is based on variational optical flow estimation that penalises gradients and performs optimization in a coarse-to-fine manner to estimate optical flow. Thus, it can be said that regions with higher gradients i.e. higher Sobel filter responses are estimated with higher accuracy.

Previously, we described the optimization to estimate the correct depth d for each pixel $\hat{\mathbf{u}}$ in Image $\mathbb{C}$ as a minimization between the reprojected points $\hat{\mathbf{p}}_R$, $\hat{\mathbf{p}}_U$, and the corresponding points obtained by disparity $\hat{\mathbf{u}}_R$, $\hat{\mathbf{u}}_U$. We consider the Sobel filter response in the rectified state as a measure of 'certainty' of disparity estimation. Hence, we weigh the optimization of each pixel with the Sobel filter responses ω_{CR} and ω_{CU}.

Moreover, in order to calculate the correct distances between $\hat{\mathbf{p}}_R$, $\hat{\mathbf{p}}_U$, and $\hat{\mathbf{u}}_R$, $\hat{\mathbf{u}}_U$, it is important to consider the fact that they are unit vector pixels on the surface of a sphere. They do not move in Euclidian space, but on a spherical Riemannian manifold of unit radius. Thus, the actual distance between them should be be calculated as the 'geodesic' distance i.e. the distance along the surface of the sphere.

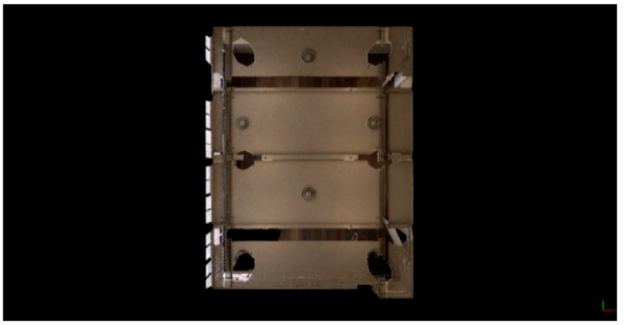
Thus, considering the disparity uncertainty and the geodesic distance along with sphere, the final optimization problem to obtain the ideal depth $d(\hat{\mathbf{u}})$ of

(a) Groundtruth

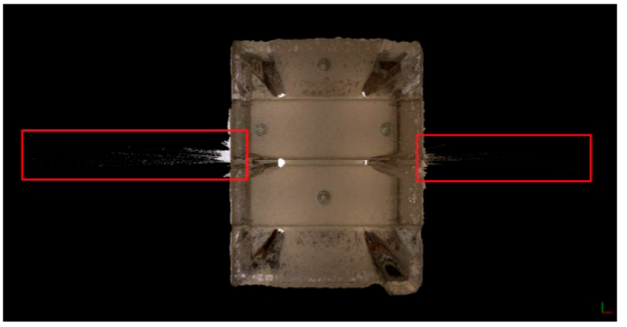
(b) Unoptimized result

(c) Result of the proposed method

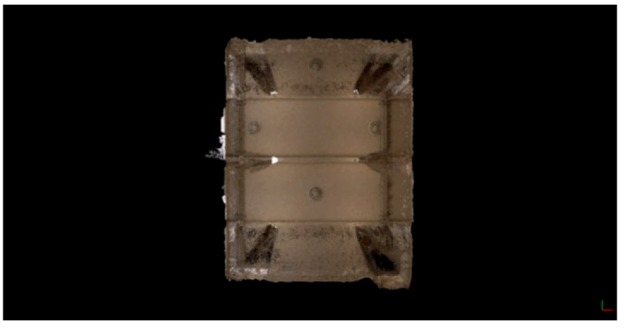
Figure 11. Results of 3D reconstruction in the simulated environment. (a) Groundtruth, (b) Unoptimized result and (c) Result of the proposed method.



(a) Groundtruth



(b) Unoptimized result



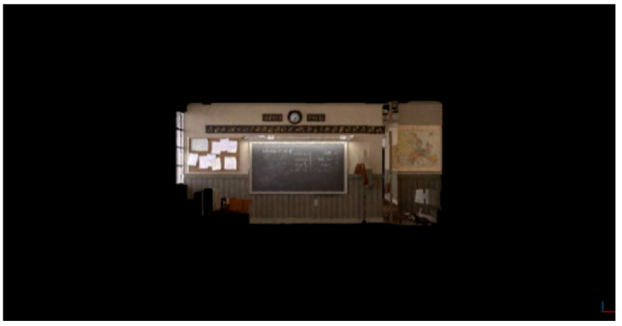
(c) Result of the proposed method

Figure 12. Results of 3D reconstruction in the simulated environment: top view. (a) Groundtruth, (b) Unoptimized result and (c) Result of the proposed method.

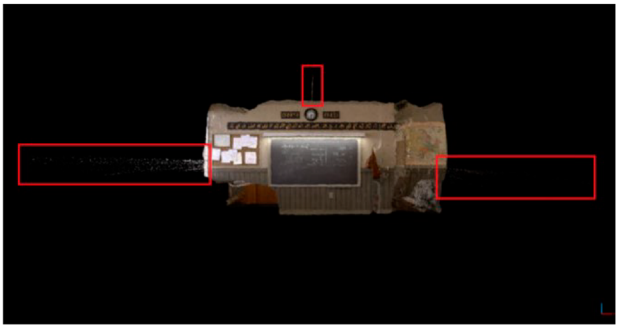
pixel $\hat{\mathbf{u}}$ in $\mathbb{C}$ is a minimization of the reprojection error, posed as,

$$d(\hat{\mathbf{u}}) = \underset{\forall(\hat{\mathbf{u}})}{\operatorname{argmin}} \left(\omega_{CR}(\hat{\mathbf{u}}_R - \hat{\mathbf{p}}_R)^2 + \omega_{UR}(\hat{\mathbf{u}}_U - \hat{\mathbf{p}}_U)^2 \right). \tag{1}$$

This minimizes the reprojection error between the calculated pixels and reprojected pixels of Image $\mathbb{C}$ on images $\mathbb{R}$ and $\mathbb{U}$, each weighted by the certainty of disparity estimation. The more certain disparity pair will be given importance over the other. The value of d at the end of the minimization is taken to be the best depth estimate of pixel $\hat{\mathbf{u}}$. The minimization can be done using any non-linear least squares method. As mentioned earlier, the initial guess for the optimization is taken to be the average of the depth values calculated for the binocular systems $\mathbb{C}$–$\mathbb{R}$ and $\mathbb{U}$–$\mathbb{C}$, i.e. the 'unoptimized' state.



(a) Groundtruth



(b) Unoptimized result

(c) Result of the proposed method

Figure 13. Results of 3D reconstruction in the simulated environment: side view. (a) Groundtruth, (b) Unoptimized result and (c) Result of the proposed method.

4. Experimental evaluation

4.1. Experimental conditions

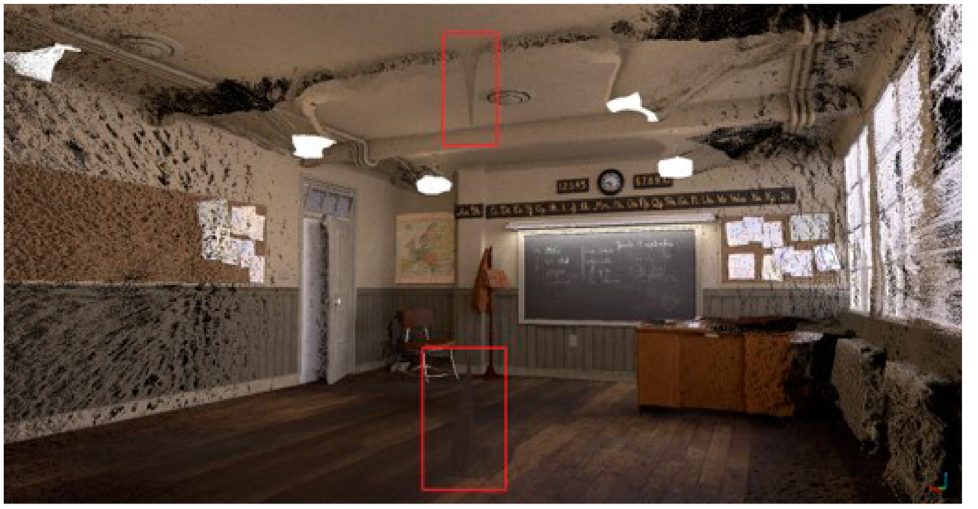
In order to evaluate the performance of our proposed method, we conducted experiments in a simulated and a real environment. The method used for disparity estimation was [14], a deep-learning based optical flow estimation method. The algorithm used for the non-linear least-squares optimization was the Levenberg-Marquardt [15] method. In order to illustrate the increase in accuracy when considering epipolar and geometric uncertainty when using 3D information obtained from all three views, we compared our method with the unweighted average method in [3], i.e. the 'unoptimized' state of the depth calculation from both 360-degree images. In addition, to illustrate the differences with learning-based monocular depth estimation methods, we also performed qualitative comparison with Pano3D [10].

4.1.1. Simulated environment

The simulated environment was used for the reason that it is difficult to estimate accurate all-round depth in a real environment. A simulated environment provides accurate groundtruth and allows for quantitative testing. The simulated environment was created using Blender [16], an open-source 3DCG software commonly used for CG production in animation and movies. A virtual 'classroom' environment was created and realistic 360-degree trinocular images were rendered at a resolution of 5000 × 2500 pixels and processed using the proposed method. Except for the image capture, all processing was done exactly the way it was on real, captured images. The cameras were 0.4m apart in an L-shaped orientation as described in the proposed method. The environment is shown in Figure 7 and the captured images are shown in Figure 8.

4.1.2. Real environment

For real-experiments, we chose an experimental room in the university with a lot of clutter. The Ricoh Theta Z1 Spherical camera was used to capture the environment from known camera positions. The baseline was the same as that of the simulated environment at 0.4 m. The image resolution used was the full resolution of the camera at 6720 × 3360 pixels. The capture was actually done using a single camera which was moved to the other two positions after the first capture. All three camera positions are at the same height, forming a horizontal 'L' shape in the room. The environment and the camera are shown in Figure 9 and the captured images are shown in Figure 10. Since



(a) Unoptimized result

(b) Result of the proposed method

Figure 14. Results of 3D reconstruction in the simulated environment: inside view. (a) Unoptimized result and (b) Result of the proposed method.

groundtruth information was not available, only qualitative comparisons to the baseline 'unoptimized' state are shown.

4.2. Experimental results and discussions

4.2.1. Simulated environment

The results of 3D reconstruction are shown in Figure 11. The groundtruth, unoptimized state, and the results of the proposed method are shown in order. Figures 12–14 show the results from the top and side views of the classroom, respectively. It can be seen that the overall shape of the classroom is well reconstructed by all methods, but the result in the unoptimized state has many outliers and distortion in the epipolar region, both inside and outside indicating a significantly degraded reconstruction. Particularly, distortion can be seen from the inside view of the room in Figures 14 and 15. Meanwhile, the results of the proposed method show significantly reduced outliers.

Error maps of the results of the proposed method are shown in Figure 17, ranging from 0 percent (green) and 10 percent (red). The number of outliers and mean absolute errors are shown in Figure 16. We judged an outlier to be an error of more than 10 m. The number of outliers reduced from 2109 to just 84 in the proposed method. All of them were in the epipolar directions. The mean absolute errors reduced overall from 0.0545 m to 0.0449 m. In the epipolar direction (taken within a range of 30 degrees from the epipolar direction) the mean absolute error reduced from 0.0806 m to 0.0529 m, representing a reduction of approximately 34.3 percent. A reduction in both, the number of outliers and the mean absolute error indicates that the proposed method worked well.

(a) Unoptimized result

(b) Result of the proposed method

Figure 15. Results of 3D reconstruction in the simulated environment: another inside view. (a) Unoptimized result and (b) Result of the proposed method.

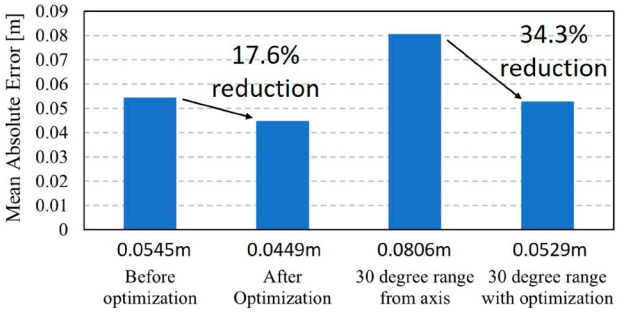


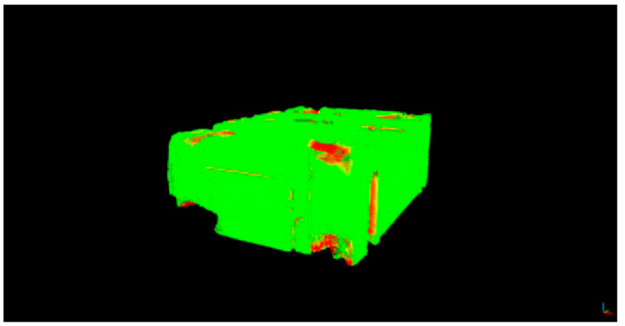
Figure 16. Reduction in the number of outliers and mean average error.

An important factor in the measurement is the baseline length. In this experiment, the maximum distance from the camera to the wall was around 6 m. Therefore, we decided that a baseline of 0.4 m was optimal. This is based on the calculation that a resolution of 5000 × 2500 pixels would yield a the measurable distance of a maximum of 500 m, at which point the disparity would be 1 pixel. Therefore, we excluded points with a measurement distance of 500 m or more from the accuracy evaluation.

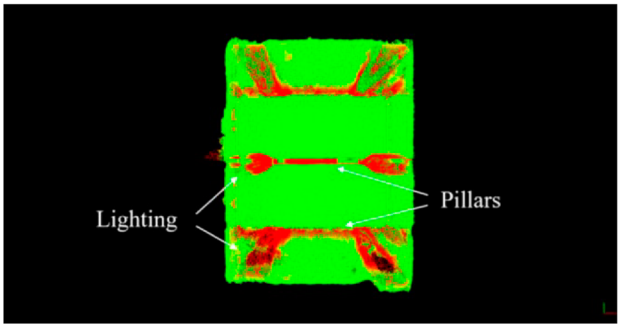
However, some remaining problems included rounding at the corners of the classroom and at the intersections of walls and other planes. This is probably due to degraded optical flow in presence of textureless discontinuities, i.e. a lack of textures in both epipolar directions.

4.2.2. Real environment

Figures 18–20 show the results of restoring a room using images taken by the Ricoh Theta Z1. It can be seen that several errors occur in the area close to the epipolar line. The proposed method greatly reduced the number of outliers and improved accuracy, but the distortion in the

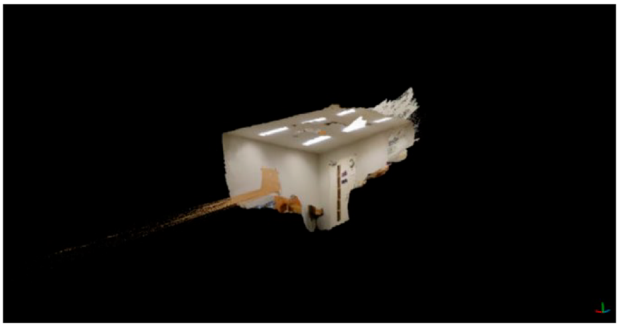


(a) Error map

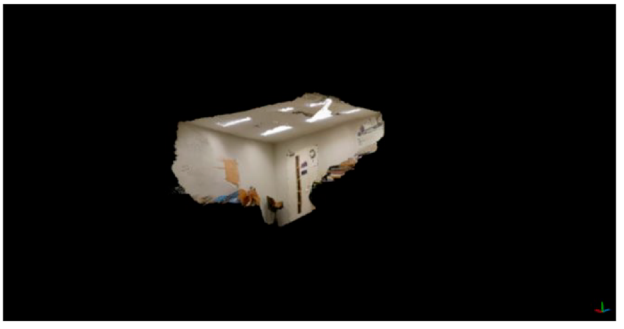


(b) Error map (top view)

Figure 17. Error maps of the 3D Reconstruction result ranging from 0 percent (in green) to 10 percent (in red). Some errors remain near the lights and pillars due to difficult of disparity estimation in the presence of sharp discontinuities. (a) Error map and (b) Error map (top view).

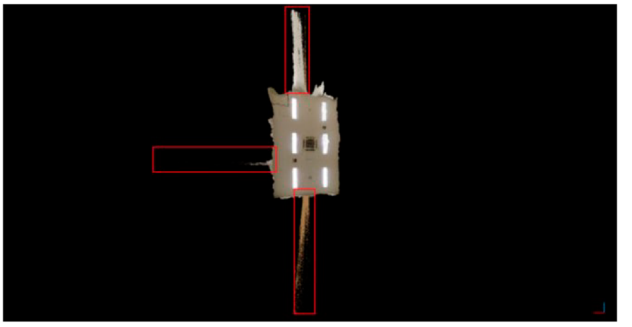


(a) Unoptimized result showing outliers



(b) Results of the proposed method

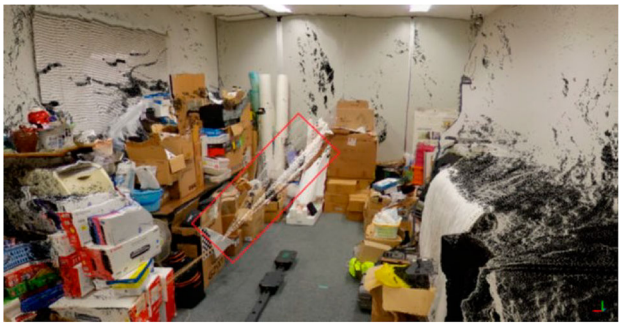
Figure 18. Results in the real environment. (a) Unoptimized result showing outliers and (b) Results of the proposed method.



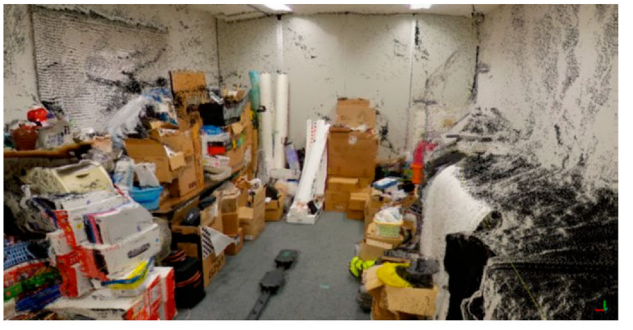
(a) Unoptimized result showing outliers (top view)

(b) Results of the proposed method (top view)

Figure 19. Results in the real environment (top view). (a) Unoptimized result showing outliers (top view) and (b) Results of the proposed method (top view).



(a) Unoptimized result showing outliers (view from inside)



(b) Results of the proposed method (view from inside)

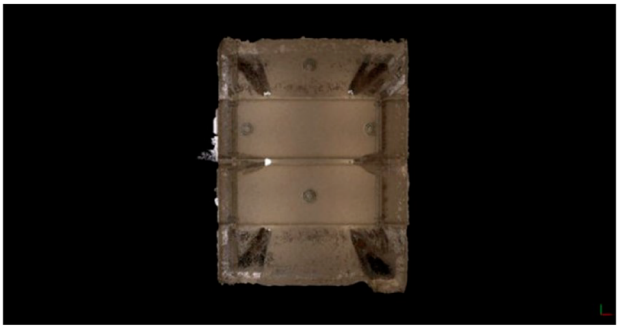
Figure 20. Results in the real environment (view from inside). (a) Unoptimized result showing outliers (view from inside) and (b) Results of the proposed method (view from inside).

corners of the room indicated by the arrows was a little larger after optimization. The shape of the room cannot be seen clearly at several areas due to the fact that there was a lot of clutter in the room and the reconstruction was conducted from the inside. In Figure 20, it can be seen that outliers were observed in the interior of the room in the unoptimized state and they were removed using the proposed method. Inspite of the clutter, our proposed method was able to recover the shape of the room from the inside.

4.2.3. Comparison with learning-based monocular depth estimation

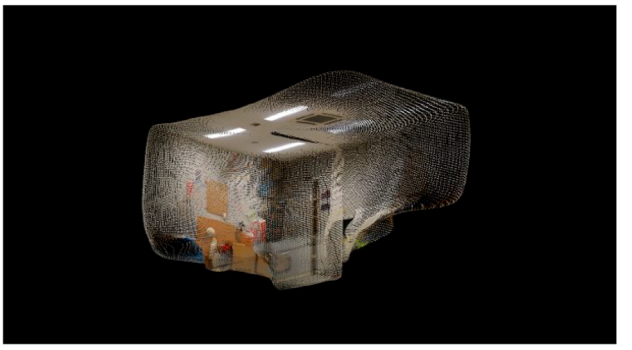
In order to justify the focus on geometry based methods, we also compared the results of our proposed method to those obtained from Pano3D [10] in both simulated and real environments. The results are shown Figures 21 and 22. It can be seen that in both the real and the simulated environment, Pano3D [10] was unable to preserve the geometric details of the room, as predicted. The overall shape was distorted as compared to our proposed method.

(a) Results of Learning-based Monocular Depth Estimation by Pano3D [10] in the simulated environment (top view). The results show lack of detail and uneven shape preservation.

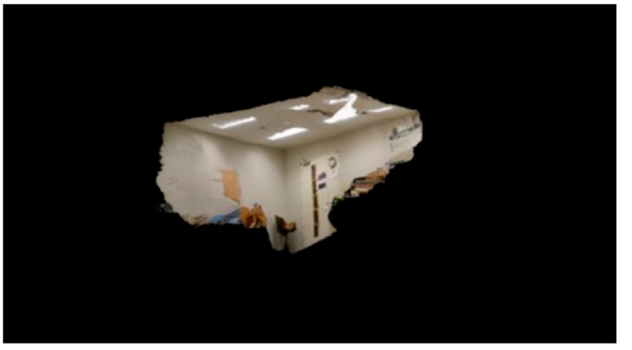


(b) Results of the proposed method in the simulated environment, showing preservation of detail and shape (top view).

Figure 21. Comparison with Learning-based Monocular Depth Estimation Method Pano3D [10] in the simulated environment. (a) Results of Learning-based Monocular Depth Estimation by Pano3D [10] in the simulated environment (top view). The results show lack of detail and uneven shape preservation and (b) Results of the proposed method in the simulated environment, showing preservation of detail and shape (top view).



(a) Results of Learning-based Monocular Depth Estimation by Pano3D [10] in the simulated environment (top view). The results show lack of detail and uneven shape preservation.



(b) Results of the proposed method in the simulated environment, showing preservation of detail and shape (top view).

Figure 22. Comparison with Learning-based Monocular Depth Estimation Method Pano3D [10] in the real environment. (a) Results of Learning-based Monocular Depth Estimation by Pano3D [10] in the simulated environment (top view). The results show lack of detail and uneven shape preservation and (b) Results of the proposed method in the simulated environment, showing preservation of detail and shape (top view).

5. Conclusion and future work

In this research, we proposed a method for accurate, all-round 3D reconstruction via trinocular 360-degree cameras, considering uncertainty in a geometric optimization to maximize accuracy. The proposed method estimated the disparity uncertainty and applied the constraint that each reconstructed point should be projected in all images at geometrically consisted positions. This was effective in reducing outliers and distortion in the epipolar directions and minimizing the mean absolute error of measurement, as shown in quantitative and qualitative evaluation experiments.

In future, we will consider several directions for improving the results. The consistency of the RGB information of the pixels can also be considered. Moreover, it would also be effective to consider the information of multiple pixels instead of comparing one pixel at a time. Moreover, a trinocular setup can also be used to find intersecting epipolar lines and improve the disparity estimation itself.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Sarthak Pathak received his B.Tech., and M.Tech. degrees in the Department of Engineering Design, Indian Institute of Technology Madras, India, in 2014. He received his Ph.D. in the Department of Precision Engineering, the University of Tokyo, Japan, in 2017. After working as a Postdoctoral Research Fellow and Project Assistant Professor at the same, he is currently an Assistant Professor in the Department of Precision Mechanics at Chuo University in Tokyo, Japan. His research interests are robot vision, specifically, localization and 3D reconstruction, especially using 360-degree cameras.

Takumi Hamada received his B.Eng., and M.Eng., degrees from the Department of Precision Mechanics at Chuo University in 2021 and 2023, respectively. He is currently employed as an engineer at Nissan Motor Corporation, Kanagawa, Japan. His research interests include 3D reconstruction using 360-degree cameras.

Kazunori Umeda received B.Eng., M.Eng., and Ph.D. degrees in Precision Machinery Engineering from the University of Tokyo, Japan, in 1989, 1991, and 1994, respectively. He became a Lecturer of Precision Mechanics at Chuo University, Japan in 1994, and is currently a Professor since 2006. He was a visiting worker at the National Research Council of Canada from 2003 to 2004. His research interests include robot vision, 3D vision, and human interface using vision. He is a member of RSJ, IEEE, JSPE, JSME,SICE, IEICE, etc.

ORCID

Sarthak Pathak http://orcid.org/0000-0002-5271-1782

References

[1] Pagani A, Stricker D. Structure from motion using full spherical panoramic cameras. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). IEEE; 2011. p. 375–382.

[2] Li S. Binocular spherical stereo. IEEE Trans Intell Transp Syst. 2008;9(4):589–600. doi: 10.1109/TITS.2008.2006736

[3] Li S. Trinocular spherical stereo. 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE; 2006. p. 4786–4791.

[4] Yin W, Pathak S, Moro A, et al. Accurate all-round 3D measurement using trinocular spherical stereo via weighted reprojection error minimization. 2019 IEEE International Symposium on Multimedia (ISM). IEEE; 2019. p. 86–867.

[5] Chen L, Wang W, Mordohai P. Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 17235–17244. Vancouver, Canada.

[6] Hermans A, Floros G, Leibe B. Dense 3D semantic mapping of indoor scenes from RGB-D images. 2014 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2014. p. 2631–2638.

[7] Wang J, Huang S, Zhao L, et al. High quality 3D reconstruction of indoor environments using RGB-D sensors. 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE; 2017. p. 1739–1744.

[8] Vlaminck M, Luong H, Goeman W, et al. 3D scene reconstruction using omnidirectional vision and lidar: a hybrid approach. Sensors. 2016;16(11):1923. doi: 10.3390/s16111923

[9] Charron N, McLaughlin E, Phillips S, et al. Automated bridge inspection using mobile ground robotics. J Struct Eng. 2019;145(11):04019137. doi: 10.1061/(ASCE)ST.1943-541X.0002404

[10] Albanis G, Zioulis N, Drakoulis P, et al. Pano3d: a holistic benchmark and a solid baseline for 360deg depth estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. p. 3727–3737.

[11] Kim H, Hilton A. 3D scene reconstruction from multiple spherical stereo pairs. Int J Comput Vis. 2013;104(1):94–116. doi: 10.1007/s11263-013-0616-1

[12] Scaramuzza D, Martinelli A, Siegwart R. A toolbox for easily calibrating omnidirectional cameras. 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE; 2006. p. 5695–5701.

[13] Pathak S, Moro A, Yamashita A, et al. Dense 3D reconstruction from two spherical images via optical flow-based equirectangular epipolar rectification. 2016 IEEE International Conference on Imaging Systems and Techniques (IST). IEEE; 2016. p. 140–145.

[14] Weinzaepfel P, Revaud J, Harchaoui Z, et al. DeepFlow: large displacement optical flow with deep matching. Proceedings of the IEEE International Conference on Computer Vision; 2013 Dec. p. 1385–1392. Portland, OR, USA.

[15] Lourakis M. LevMar: Levenberg-Marquardt nonlinear least squares algorithms in C/C++. 2004 Jul. Available from: http://www.ics.forth.gr/lourakis/levmar/+.

[16] Blender Foundation. Blender – a 3D modelling and rendering package. Amsterdam: Stichting Blender Foundation; 2023. Available from: http://www.blender.org.