Efficient Hundreds-baseline Stereo by Counting Interest Points for Moving Omni-directional Multi-camera System

Tomokazu Sato and Naokazu Yokoya

Corresponding author: Tomokazu Sato (tomoka-s@is.naist.jp)

Affiliation: Graduate School of Information Science, Nara Institute of Science and Technology, Japan

Address: 8916-5 Takayama, Ikoma, Nara, Japan

Phone:+81-743-72-5293

FAX: +81-743-72-5299

This manuscript is submitted to Journal of Visual Communication and Image Representation (JVCI), special issue on Multi-Camera Imaging, Coding and Innovative Display: Techniques and Systems

Abstract

In this article, we propose an efficient method for estimating a depth map from long-baseline image sequences captured by a calibrated moving multi-camera system. Our concept for estimating a depth map is very simple; we integrate the counting of the total number of interest points (TNIP) in images with the original framework of multiple baseline stereo. Even by using a simple algorithm, the depth can be determined without computing similarity measures such as SSD (sum of squared differences) and NCC (normalized cross correlation) that have been used for conventional stereo matching. The proposed stereo algorithm is computationally efficient and robust for distortions and occlusions and has high affinity with omni-directional and multi-camera imaging. Although expected trade-off between accuracy and efficiency is confirmed for a naive TNIP-based method, a hybrid approach that uses both TNIP and SSD improve this with realizing high accurate and efficient depth estimation. We have experimentally verified the validity and feasibility of the TNIP-based stereo algorithm for both synthetic and real outdoor scenes.

Keywords

multi-baseline stereo depth estimation interest point omni-directional camera system multi-camera system.

1. Introduction

Depth map estimation from images is a very important topic in the field of computer vision because depth information can be used in several different applications such as 3-D modeling, object recognition, surveillance, and novel view synthesis. In the past decades, a lot of methods for stereo algorithm are developed by many researchers, and most of these works were designed for a pair of standard camera units [1]. On the other hand, like the Google Street View, we can now easily acquire an omni-directional image sequence for large outdoor environments by moving a vehicle where the camera is mounted. However, for such an omni-directional image stream, most of conventional works designed for two-frame images do not work well due to large image distortion and large baseline. In this paper, in order to realize efficient and accurate depth estimation for omni-directional image sequence, we extend the conventional multi-baseline stereo framework that is proposed by Okutomi and Kanade [2]. This method has good feature in that an arbitrary number of images can be simultaneously used for depth estimation. This increases the accuracy of depth estimation and decreases the ambiguity in stereo matching. Using recent developments in camera calibration techniques, the multiple-baseline stereo framework have been employed for a freely moving video camera [3, 4, 5, 6, 7]. A freely moving video camera is suitable for the 3-D modeling of a large-scale environment because it easily makes a long-distance baseline between cameras. However, when we employ the long-baseline omni-direcitonal images, following weaknesses of original mult-baseline stereo becomes the critical problem in practice.

(1) Image distortion: In omni-directional video sequence, image patterns around the physical 3-D point is easily distorted and resolutions of these patterns are not uniform due to both the characteristic of omni-directional vision and the large motion of the camera system. Because depth information should be estimated for any directions, rectification techniques cannot resolve this problem.

(2) Occlusion: In large-baseline stereo in outdoor environment, there are much occluders than standard short-baseline stereo. When a point on an object where the depth is to be estimated is occluded by other objects in a part of an input video, the occluder gives a negative score to the score function of the multiple-baseline stereo: SSSD (sum of SSD). This negative score prevents the algorithm from obtaining a correct estimation of the depth map around occlusions.

(3) Computational cost: Although the utilization of multiple input images increases the accuracy of depth estimation, it consumes a large amount of memory and computational resources. Some patches for the distortion and the occlusion problems need additional computational cost.

In order to solve these problems, we propose a novel approach that estimates depths by using interest points, as shown in Fig. 1, that are corners and cross points of edges in images. The framework of our depth estimation method is basically the same as the original multiple-baseline stereo except for a newly employed objective function: TNIP (total number of interest points). The concept is based on the very simple assumption that the corners of objects and cross points of texture edges in the 3-D space (3-D interest points) will appear in video images as 2-D interest points at the projected positions of the 3-D interest points. By searching a depth that maximizes the

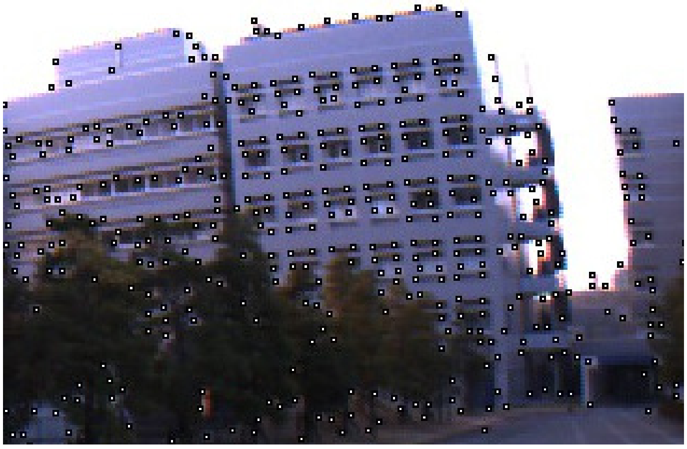


Figure 1: Example of interest points.

total number of 2-D interest points under epipolar constraint, the depth can be determined as the position of a 3-D interest point. It should be noted that the proposed method assumes that camera parameters of the input videos are pre-calibrated and the camera calibration problem is beyond the scope of this research.

By using the objective function TNIP for depth estimation, the problems mentioned earlier can be solved: (1) Detected position of the 2-D interest point is ideally not affect by image rotation and distortion. (2) The score function TNIP is not significantly affected by occluders and the position of corners indicates the unique position in 3-D space. (3) The computational cost of depth estimation is very cheap because the depth can be determined by only counting the interest points. However, depths for non-interest points cannot be estimated by TNIP; this is not a critical problem for 3-D modeling and some other applications because 3-D interest points contain the corners of the 3-D models. Further it should be noted that the TNIP-based method estimates the depth for the 3-D corners rather than that for the target pixel. Thus, the accuracy of the depths obtained from the raw TNIP function is a little lower than that obtained by SSSD; however, TNIP drastically decreases the computational cost. In this research, in order to resolve the weaknesses of the raw TNIP function, we also suggest a hybrid approach in which both SSSD and TNIP are used. In the hybrid approach, first, TNIP is used to roughly and quickly determine the depth for each interest point. For a limited searching range by TNIP, the depth value is then re-searched by SSSD with very small window size.

The reminder of this paper is organized as follows. First, related stereo algorithms and 3-D reconstruction methods are reviewed in Section 2, and the contribution of the proposed method is also explained. In Section 3, the original multi-baseline stereo method for a moving video camera is described. Then, the new score function TNIP for multiple baseline stereo is proposed in Section 4. Each process for estimating a dense depth map is detailed in Section 5. Experimental results with simulation and a real scene are used to demonstrate the validity and feasibility of the proposed method in Section 6. Finally, Section 7 presents the conclusion and outlines of future studies.

2. Related works

2.1. Multi-view reconstruction

In the field of traditional stereo reconstruction, 3-D information has been estimated as depth maps by assuming camera parameters are pre-calibrated.

Although they have conventionally been designed for binocular and trinocular stereo imaging, recent works tend to use multi-view images [8, 9, 10, 11, 12, 13]. In these multi-view approaches, in addition to the traditional depth map based 3-D representation [2, 12, 13], voxel based [8, 6, 10] and polygon mesh based [11] 3-D representation are employed for 3-D reconstruction. However there are many combinations for 3-D representation and modeling approaches in these conventional works, one of the common problems for the multi-view stereo reconstruction is how to corresponds pixels between multiple images. In order to correspond pixels between images, the photo-consistency measure have commonly been used. The photo-consistency measure is a similarity measure that correlates the pixels on multiple images based on variance of image pixels for projected position of the unique 3-D position. The key difference between these multi-view works and the proposed method is that we do not need to utilize any intensity-based similarity measure for making correspondences between images. Instead of similarity measures, the spatial-consistency of the positions of feature points is employed, and the method efficiently tests the spatial consistency by simply counting the interest points along epipolar lines.

2.2. Feature-based 3-D reconstruction

Another work that closely related to our approach is conventional feature-based stereo [14, 15, 16]. The feature-based stereo method uses feature points in images, such as intensity edges on epipolar lines, as positions of matching candidates in image pairs. In this approach, for these matching candidates, a similarity measure such as SSD or NCC is computed and the corresponding points between images are determined based on pattern similarity. In the same manner as conventional feature-based stereo algorithms, the proposed method also employs feature points to realize the efficient and robust determination of corresponding points. However, our algorithm basically determines corresponding points without using intensity-based similarity measures.

Similar to the case of feature-based stereo algorithms, EPI(epipolar plane image)-based 3-D reconstruction [17] uses motion of the edge features along the epipolar line to recover the 3-D information. In this method, video is first captured using a video camera that moves along a direction vertical to the optical axis at a constant speed. The line images of the corresponding epipolar plane are then collected and expanded vertically to generate the EPI. In the EPI, corresponding points can be determined easily because they lie along a line in the EPI. In conventional studies, these lines are detected by Hough transformation. Okutomi et. al [18] applied the EPI-based method for rotating camera motion. This method is used for an object on a turning table and it can detect the sin curve in the EPI image instead of the lines. The problem faced in the EPI-based methods [17, 18] is that only a steady camera motion is allowed. Although some deviation from the steady camera motion can be compensated, it is difficult to process the images that include camera motion along the optical axis.

On the other hand, structure from motion (SFM) is a 3-D reconstruction method that uses the motion of feature points [19]. In this approach, based on the 2-D motion of image features in a video sequence, not only the 3-D positions of image features but also the camera position and posture parameters can be simultaneously estimated. In this study, the SFM is assumed as one of the camera calibration methods. However, in the SFM approach, good feature points are usually selected for tracking in order to ensure a certain accuracy and efficiency. Feature points on repeatable textures and natural objects are often discarded as outliers. Thus, the 3-D positions acquired by the SFM are generally very sparse.

2.3. Enhancement of multiple baseline stereo

As mentioned in Section 1, the original multiple baseline stereo method has some problems. Several methods have been proposed to resolve the occlusion problem in the multiple baseline stereo method. Okutomi et. al [7] employed an adaptive window approach [20] to improve the accuracy around the occluding boundary. In this method, they also attempted to remove blocking effect in the depth map that is often generated by the adaptive-window-based method. Kang et. al [21] and Sato et. al [6] proposed a similar concept idea: the objective function SSSD is computed as the sum of selected SSD values that are smaller than the median of the SSD values. On the other hand, Sanfourche et. al [22] employed the photo-consistency-based objective function that measures the variance of intensity for corresponding points instead of the original SSSD function. In this method, the photo-consistency is computed by shifting the projected pixels to the neighboring pixels to avoid the effect of camera calibration errors. Note that all these conventional studies are categorized into area-based and pixel-based approaches that use pixel values to determine the depths. The proposed method is a feature-based approach and it realizes the fast and robust estimation of the depth map using a completely different approach as compared to conventional approaches.

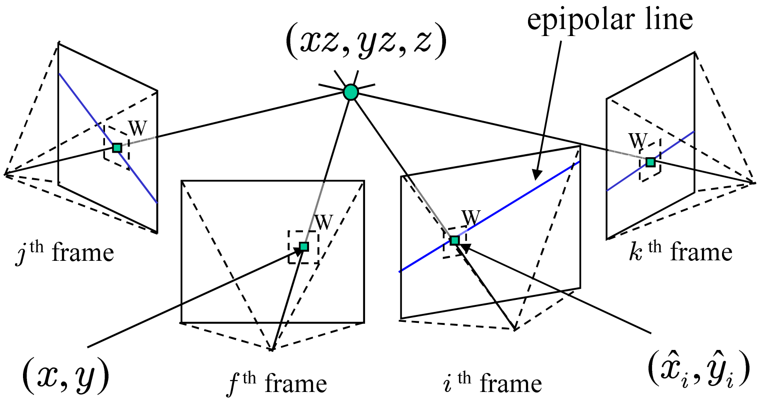


Figure 2: 3-D position of (x, y) with depth z and its projected line to each frame.

3. Multiple baseline stereo by using SSSD

In this section, we first define the coordinate systems of a general moving camera. We then summarize the principle of the original multiple baseline stereo method [2] using SSSD.

3.1. Definition of coordinate systems for moving camera

In the multiple baseline stereo, as shown in Fig. 2, the depth z of a pixel (x, y) in the f^{th} frame is estimated by using images from the j^{th} to k^{th} frames $(j \leq f \leq k)$. In the following, for simplicity, it is assumed that the focal length is 1 and the lens distortion effect has already been corrected by known intrinsic parameters. In this case, the 3-D position of the pixel (x, y) with depth z is represented as (xz, yz, z) in the camera coordinate system of the f^{th} frame. The 3-D position (xz, yz, z) is projected to the pixel $(\hat{x}_i, \hat{y}_i)$ in the image of the i^{th} frame by the following expression.

$$\begin{pmatrix} a\hat{x}_i \\ a\hat{y}_i \\ a \\ 1 \end{pmatrix} = \mathbf{M}_{fi} \begin{pmatrix} xz \\ yz \\ z \\ 1 \end{pmatrix}, \tag{1}$$

where a is a parameter and $\mathbf{M}_{fi}$ denotes a 4×4 transformation matrix from the camera coordinate system of the f^{th} frame to that of the i^{th} frame. In the multiple baseline stereo, as shown in Fig. 2, the position $(\hat{x}_i, \hat{y}_i)$ is constrained on the epipolar line, which is the projection of the 3-D line connecting the position (xz, yz, z) and the center of projection in the f^{th} frame.

3.2. Depth estimation using SSSD

In the traditional multiple baseline stereo, the depth z of a pixel (x, y) is determined by using the similarity measure SSD. The SSD is computed as the sum of the squared differences between two image patterns that have a certain size W. The SSD for the pixel (x, y) in the f^{th} frame and the pixel $(\hat{x}_i, \hat{y}_i)$ in the i^{th} frame is defined using the image intensity I as follows.

$$SSD_{fixy}(z) = \sum_{(u,v) \subseteq W} \left\{ I_f(x + u, y + v) - I_i(\hat{x}_i + u, \hat{y}_i + v) \right\}^2. \tag{2}$$

In order to evaluate the error in the depth z for all the input images, the SSD is summed up as follows:

$$SSSD_{fxy}(z) = \sum_{i=j}^{k} SSD_{fixy}(z). \tag{3}$$

The depth z is determined for each frame so as to minimize the SSSD function. Generally, to find a global minimum of the SSSD, the depth z should be searched for the entire depth range along a 3-D line from a reference pixel (x, y).

If the pixel (x, y) in the f^{th} frame is occluded by other objects in the i^{th} frame, the cost $SSSD_{fxy}(z)$ for the true depth z is increased by the occluder because $SSD_{fixy}(z)$ gives a large error. Thus, to obtain a correct depth at such an occluded part, some other computationally expensive extensions should be applied to the original multiple baseline stereo. For example, a modified SSSD can be computed by summing up only the lower halves of SSDs [6, 21]. However, the computational cost problem remains unresolved.

4. Multiple baseline stereo by counting interest points

In this section, a new score function TNIP is defined to estimate the depth z of a pixel (x, y) using the multiple baseline stereo framework. Generally, feature points in a 3-D space, such as corners of objects and cross points of texture edges, appear as 2-D feature points in images at projected positions of the 3-D feature points. These 2-D feature points can be easily detected by interest operators such as Harris's [23] and Moravec's [24] operators. If there is a 2-D feature point in image, there is some probability of 3-D corner existence on the 3-D line connecting this 2-D feature point and camera's projection center. In the TNIP based method, basically, the 3-D position where this probability becomes the maximum is searched for along the epipolar lines as shown in Figure 2. By searching a depth that maximizes the total number of 2-D interest points under epipolar constraint, the depth can be determined as the position of a 3-D interest point.

In this study, the depth z of a pixel (x, y) is determined so as to maximize the TNIP score function that is defined as follows.

$$TNIP_{fxy}(z) = \sum_{i=j}^{k} \sum_{(u,v) \subseteq W} H_i(\hat{x}_i + u, \hat{y}_i + v). \tag{4}$$

$$H_i(u, v) = \begin{cases} 1 \ ; \text{ interest point exists at} \\ \qquad (u, v) \text{ in } i^{th} \text{ frame.} \\ 0 \ ; \text{ otherwise} \end{cases} \tag{5}$$

The TNIP score represents the total number of interest points that exist within the $(\hat{x}_i, \hat{y}_i)$ centered windows W for all the frames. It should be noted that the size of W should be appropriately small because interest points are not detected at positions far from the projected positions of (xz, yz, z) when there exists a feature point in the 3-D space.

By using the TNIP instead of the SSSD function in the multiple baseline stereo, the computational time can be drastically decreased because the time consuming process of comparing intensity patterns can be eliminated from the depth estimation. Moreover, the TNIP has another good feature in that it is not significantly influenced by occluders because it counts only positive scores. These claims will be experimentally verified in later sections.

5. Depth estimation from an image sequence

This section describes the processes of depth estimation. In the proposed method, after detecting the interest points in all the input images, the depths of all the interest points are determined by the multiple baseline stereo framework with the TNIP score function. The outliers of the estimated depths are then eliminated by using their confidences defined by considering the consistency among the results in multiple frames.

5.1. Depth estimation for interest points

The depths of all the interest points detected in the video images are computed by maximizing the TNIP score function that is defined in Section 4. The depth z is searched along a 3-D line from each target pixel so as to maximize the TNIP score in a given range of depth. In order to find the best depth value that maximizes TNIP, all the depths within the given range should be tested. For realizing an efficient search, the skip value l for the depth z is adaptively determined in this study. Concretely, the 3-D line segment that connects (xz, yz, z) and $(x(z+l), y(z+l), z+l)$ is projected to each frame, and the skip value l for depth z is adjusted so that the maximum length of the projected line segment becomes a certain length $L(L \leq 1.0)$ [pixel]. It should be noted that no intensity images are required to compute TNIP-scores after once all the feature points on images are detected Only the 2-D positions of interest points and camera parameters should be stored to compute the TNIP-scores. This implies that the proposed method requires only one-eighth the memory space required by SSSD to compute depth, assuming that 8-bit grayscale images are used in SSSD. Although the

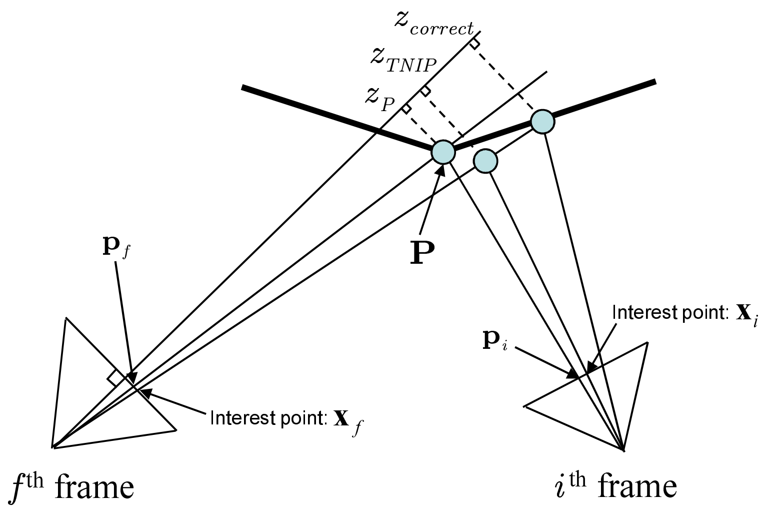


Figure 3: Cause of estimation error in TNIP.

TNIP-based method can estimate depths with a low computational cost, the raw TNIP function has a weakness in that the accuracy of the estimated depth is slightly worse than that in case of SSSD. Below, we describe the reason and solution for this weakness.

As shown in Fig. 3, the TNIP-based method estimates the depth value by using interest points that are detected around the projected position $\mathbf{p}_i(j \leq i \leq k)$ of the 3-D interest point $\mathbf{P}$. However, the detected position $\mathbf{x}_i$ of the interest point $\mathbf{P}$ in the i^{th} frame does not always coincide with the projected position $\mathbf{p}_i$ due to feature detection errors. For the target frame f, there also exists a detection error for the pixel $\mathbf{x}_f$. Thus, the searching line for the depth of $\mathbf{x}_f$ may not cross the 3-D point $\mathbf{P}$. As shown in Fig. 3, in this case, although the correct depth for the pixel $\mathbf{x}_f$ is $z_{correct}$, the TNIP score is maximized at z_{TNIP} that is almost equal to the depth z_P of the 3-D position $\mathbf{P}$. This is caused by a characteristic of the TNIP score function; the TNIP-based method estimates the depth for the 3-D corners rather than the depth for the target pixel. In contrast, the SSSD-based methods do not have such a problem. This characteristic of the TNIP may not be a problem for certain applications such as collision avoidance and environment recognition where robustness and efficiency are more important than accuracy.

On the other hand, for applications where the accuracy of the depth remains an important factor, one solution is to refine the estimated depth using the SSSD. More concretely, after TNIP-based depth estimation, a limited range ($z_{TNIP} - Cl < z < z_{TNIP} + Cl$) can be re-scanned by SSSD. If we employ this hybrid approach, although the advantages in terms of memory requirement are lost, the computational efficiency is maintained because the searching range of depth in re-scanning is quite limited in the refinement process. Moreover, because the TNIP function is not significantly affected by occlusions, the hybrid approach can robustly and accurately estimate depths.

5.2. Elimination of outliers

In this process, unreliable depths are eliminated by a cross validation approach for multiple image inputs. Fig. 4 shows an example setting of two cameras. The depth z_f of the pixel $\mathbf{x}$ is evaluated by a consistency check of the estimated depths. First, for the i^{th} frame, the projected position $\hat{\mathbf{x}}_i$ of the 3-D position $\mathbf{S}_f$ that corresponds to the depth z_f is computed. Inversely, for the f^{th} frame, the projected position $\mathbf{x}'$ of the 3-D position $\mathbf{S}_i$ that corresponds to the depth z_i is computed. The consistency for the depths z_f and z_i is then evaluated by the distance $d_i(\mathbf{x}) = |\mathbf{x} - \mathbf{x}'|$ in the

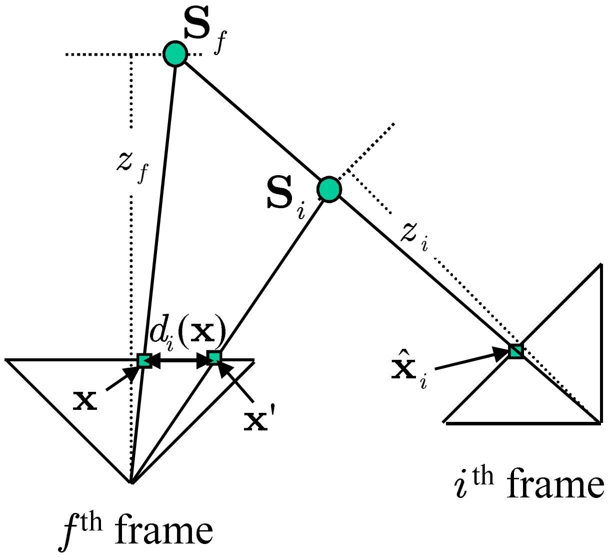


Figure 4: Elimination of outliers.

target frame. In this research, the confidence $R(\mathbf{x})$ for the depth z_f of the pixel $\mathbf{x}$ is defined as follows using distance d_i.

$$R(\mathbf{x}) = \frac{\sum_{i=j}^{k}\{0; d_i(\mathbf{x}) > T, 1; d_i(\mathbf{x}) \leq T)}{k-j+1}, \tag{6}$$

where T is a threshold for the distance d_i that judges whether or not the depth z_f is consistent with the depth z_i. The confidence $R(\mathbf{x})$ indicates a rate of consistent depth pairs. If all the depths are correctly estimated, $R(\mathbf{x})$ has a maximum value of 1. In the proposed method, the depth z_f of the pixel $\mathbf{x}$ whose confidence $R(\mathbf{x})$ is lower than a given threshold U is regarded as an outlier, and it is deleted. Note that for the TNIP-based algorithm, a feature point exists at the position $\hat{\mathbf{x}}_i$ in each frame, and for these feature points, the depth z_i is directly available. However, when no feature point exists at the position $\hat{\mathbf{x}}_i$ or the hybrid approach is employed, the depth z_i is unknown and $d_i(\mathbf{x})$ cannot be computed. In this case, z_i is first computed by interpolation using the estimated depths of nearby feature points, and they are then used to compute $d_i(\mathbf{x})$.

6. Experiments

We have conducted two types of experiments. One is concerned with the comparison of SSSD, TNIP, and the hybrid (HYBRID) method by means of a computer simulation. The other is conducted for depth map estimation for a real outdoor environment. For all the experiments, the Harris interest operator [23] is employed as an interest point detector. Although the SIFT detector [25] has often been employed as an interest operator in many recent researches, this operator tends to detect non-corner points due to its scale independent characteristic and the points detected by SIFT do not satisfy the assumption for TNIP. Thus, the Harris operator, one of the standard corner detectors, is employed in this experiment.

6.1. Quantitative evaluation in computer simulation

In this section, we first describe the configuration of the computer simulation. After determining the best window size for the SSSD, TNIP, and HYBRID approaches, the accuracy and computational efficiency are compared.

6.1.1. Setup of computer simulation

In the experiment, two textured planes are located in a virtual environment, and a virtual camera captures an image sequence by moving the camera around these planes. Two types of texture patterns are used for the planes, as shown in Fig. 5. The layout of the planes and the motion path of the virtual camera are illustrated in Fig. 6. A total of 91 input images, some of which are shown in Fig. 7, are captured by the moving camera whose motion draws a quarter circle, as shown in Fig. 6. Due to the motion of the

Table 1: Parameters used in the simulation.

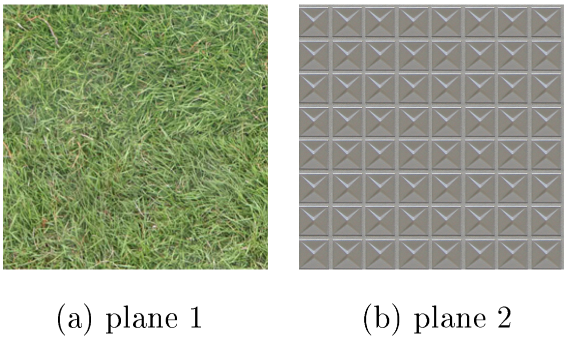
(a) for depth estimation.

Searching range of depth [mm]	3,000–35,000
Maximum depth skip on image L [pixel]	1.0
Re-scanning range coefficient C	10

(b) for outlier elimination.

Threshold of distance T [pixel]	1.0
Threshold of confidence U	0.4

camera, plane 1 is occluded by plane 2 after half of the input images, and both textures are apparently distorted by the camera motion. In order to consider camera calibration errors related to intrinsic and extrinsic camera parameters, a Gaussian noise with standard deviation σ is added to the projected positions of the 3-D points and these positions are sampled to pixels. The other parameters that are used for this experiment are listed in Table 1.



(a) plane 1 (b) plane 2

Figure 5: Textures of planes.

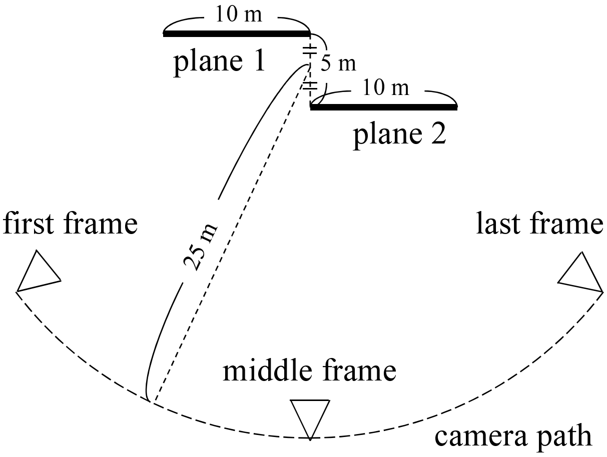
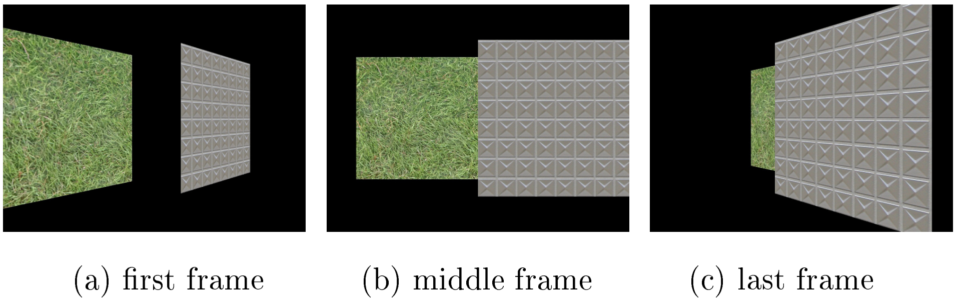


Figure 6: Layout of planes and camera path in simulation.



(a) first frame (b) middle frame (c) last frame

Figure 7: Sampled frames from 91 input images.

6.1.2. Determination of window size

In this experiment, the best window size W in Eqs. (2) and (4) is determined using the same conditions as those in the computer simulation described in the previous section. In order to determine the window size W, we evaluated the rate of inaccurate depth estimation based on the average re-projection error that is defined as follows.

$$E_p = \frac{1}{N}\sum_{i=1}^{N} |\hat{\mathbf{x}}_{ip} - \bar{\mathbf{x}}_{ip}|, \tag{7}$$

where p is a pixel index and N is the number of images used for depth estimation and it is set as 91 in this experiment. $\hat{\mathbf{x}}_{ip}$ is a projected position of the estimated depth z in the i^{th} frame for the pixel p, and $\bar{\mathbf{x}}_{ip}$ is the projected position of the ground truth. In this experiment, the automatic elimination of the outliers described in Section 5.2 is not performed to show the raw characteristic of each method, and if E_p is over 1.0 pixel for the pixel p, the estimated depth for the pixel p is judged to be inaccurate.

Fig. 8 shows the rate of inaccurate depth estimation for various window sizes and various noise levels. As shown in this figure, although the best window size of TNIP is 3×3 pixels, the rate of inaccurate depth estimation is greater than that of SSSD for 7×7 pixels, the best size for SSSD. As described in Section 5.1, it is difficult for TNIP to estimate the depth with high accuracy due to the detecting errors of feature positions in the target frame.

On the other hand, in the HYBRID, the depths are refined after TNIP-based estimation using the SSSD function. For the HYBRID, the horizontal axis in Fig. 8 indicates the window size of SSSD in the refinement process,

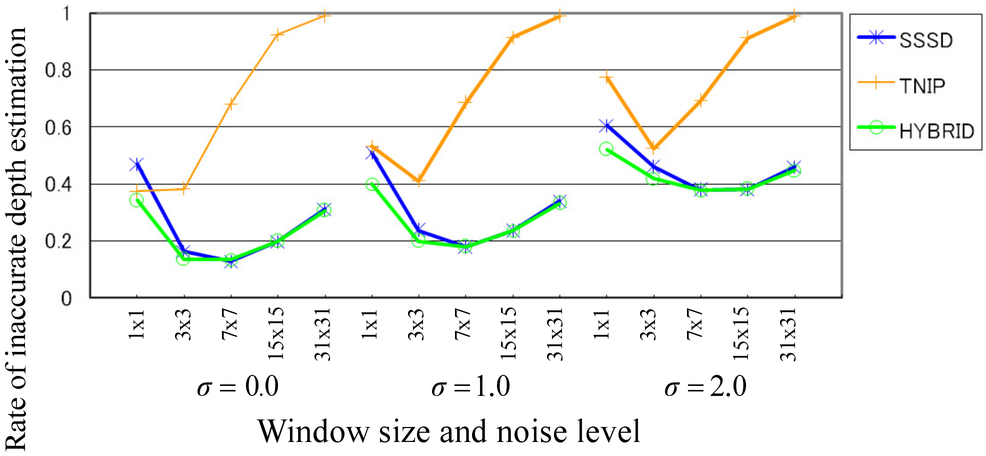


Figure 8: Rate of inaccurate depth estimation ($E \geq 1.0$ pixel) for various window sizes and noise levels.

and the window size of TNIP for the initial estimate is fixed as 3 × 3 pixels. From this figure, it is confirmed that 7 × 7 pixels is the best size for the HYBRID and the rate of inaccurate depth estimation for this approach HYBIRD is almost the same as that in the case of SSSD.

Table 2 indicates the average time required to estimate the depth of a single pixel using all the 91 input images with respect to different window sizes. The implementations of these methods are the same except for the objective function. The computation time is measured by using a PC (CPU: Pentium 4 Xeon 3.20 GHz dual core, Memory: 2 GB). From this table, we can confirm that the computational costs of the TNIP (3 × 3 window) and HYBRID (7 × 7 window) approaches are approximately 9 times and 5 times lesser than that of SSSD (15 × 15 window). Although a more detailed analysis is

Table 2: Average computational time required for estimating the depth of a single pixel [ms].

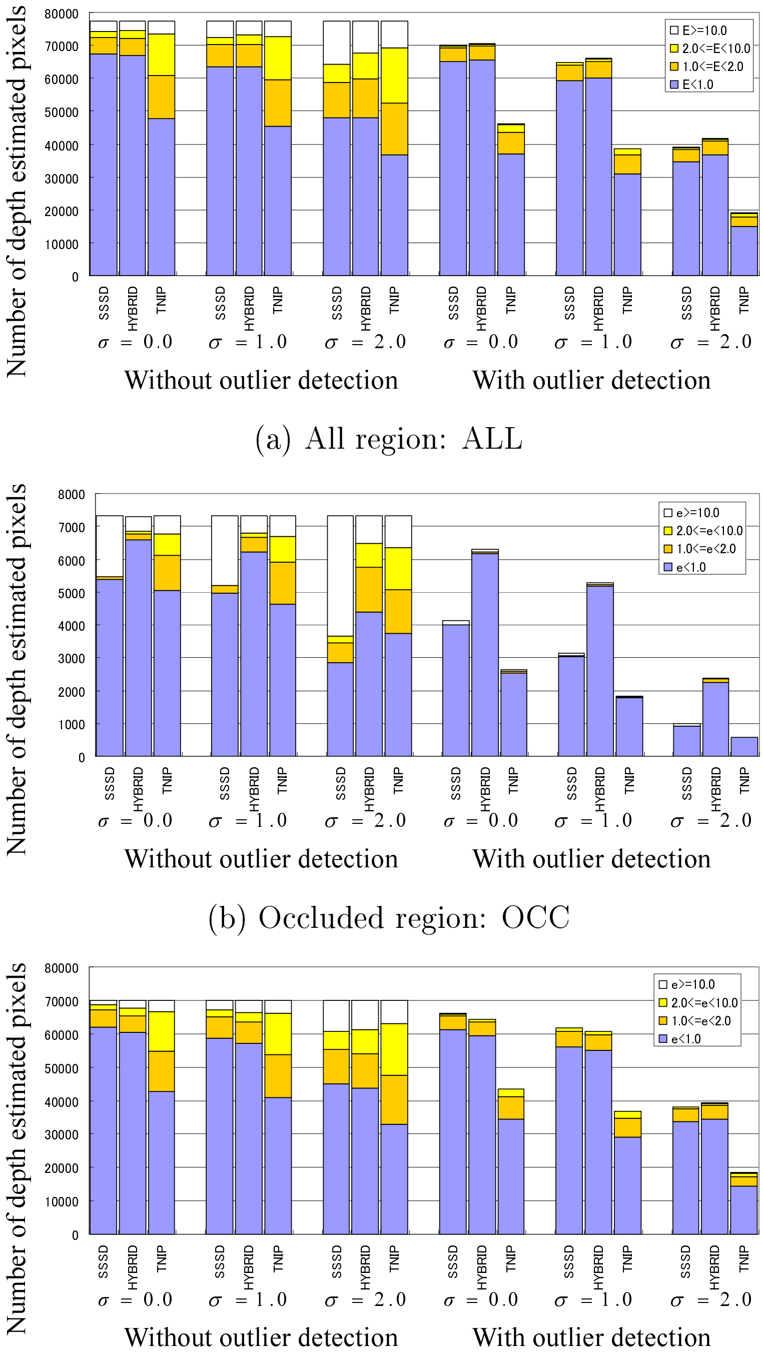
W	1×1	3×3	7×7	15×15	31×31
SSSD	13.6	25.2	**86.3**	353.9	1530
TNIP	9.0	**9.8**	11.2	13.0	21.2
HYB.	10.3	11.5	**16.7**	40.1	141.3

described in the following section, it is evident that the HYBRID approach realizes efficient depth determination with almost same the accuracy as the conventional SSSD-based method.

6.1.3. Accuracy comparison of TNIP, SSSD, and HYBRID

In this experiment, the TNIP, SSSD, and HYBRID methods are compared using the best window size for each method. In order to analyze the characteristic of each method for the occlusions, the entire image region (ALL) is divided into the occluded region (OCC) and other region (NOR). In this experiment, a region where the pixel in the target frame is occluded in more than half of the reference images is classified as OCC.

Fig. 9 shows a stacked bar chart that indicates the breakdown of estimated errors. The vertical axis in this figure indicates the number of estimated depths and each chart is separated by the level of average re-projection errors. For the horizontal axis, a combination of outlier elimination (with or without), noise level σ, and objective functions (SSSD, TNIP, and HYBRID)



(a) All region: ALL

(b) Occluded region: OCC

(c) Outside of occluded region: NOR

Figure 9: Analysis of estimation errors for each region.

are specified. As shown in Fig. 9(a), without outlier elimination, the rate of accurate depth estimation ($E < 1.0$ pixel) for the region ALL is almost the same as that in the case of SSSD and HYBRID, and that of TNIP is worse in this case. For large errors ($E \geq 10.0$ pixels), although the rate is almost at the same level for all the methods when the noise level is low, if the noise level is high ($\sigma = 2.0$ pixels), larger errors of TNIP are suppressed as compared to the others. This result validates one very important characteristic of TNIP; As explained in Section 5.1, the TNIP estimates the depth for the 3-D corners rather than the depth for the target pixel. By this reason, TNIP cannot estimate precise depth for the target pixel. However, TNIP can estimate more roughly correct depths than that of SSSD especially with high noise level. Thus, in this case, the hybrid approach can realize more accurate depth estimation than SSSD by refining the result using TNIP. With outlier elimination, for all the methods, most inaccurate results ($E \geq 2.0$ pixels) are eliminated. This demonstrates the effectiveness of outlier elimination, as described in Section 5.2.

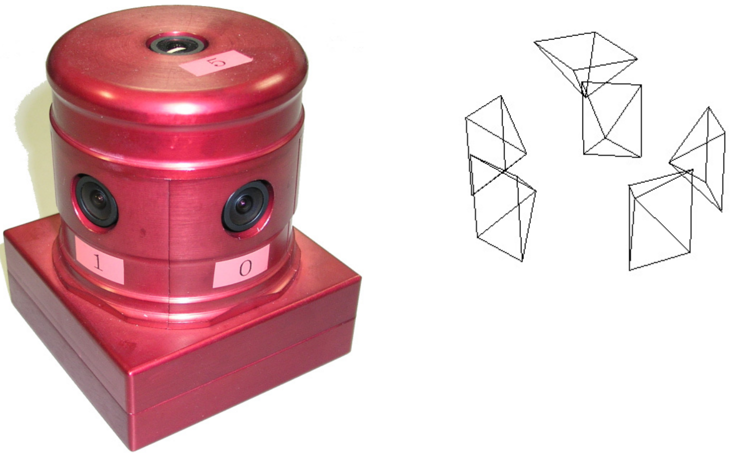
Figs. 9(b) and (c) show the error rates for the occluded region (OCC) and the other region (NOR), respectively. Note that scale of the vertical axis for OCC is different from that of other graphs because the occupancy rate of the OCC region for the ACC region is approximately 0.1. From the comparison of (b) and (c), it can be confirmed that the error rate for SSSD in the OCC region is much higher than that in the NOR region. In contrast, the error rates of the TNIP and HYBRID for the OCC region are not drastically different from those for the NOR region. This verifies the robustness of the proposed method for occluded regions.

From these experimental results, it is confirmed that the new objective function TNIP is robust for occlusions; however its raw accuracy is lower than that of the SSSD. The hybrid approach that uses both the TNIP and the SSSD can realize robust and accurate depth estimation with a lower computational cost. Please also note that TNIP based method can involve most of improvements for SSSD based multi-baseline stereo because TNIP employs the same framework with SSSD based multi-baseline stereo.

6.2. Depth estimation in an outdoor environment

In this experiment, an outdoor environment is captured by an omnidirectional multi-camera system (OMS): Pointgrey Ladybug. Fig. 10(a) shows a photograph of Ladybug, and (b) shows the view volume of each camera unit that is illustrated based on the camera calibration result. This camera system has six radially located camera units and captures six synchronized image sequences at 15 fps (resolution of each camera: 768×1024 pixels).

First, the outdoor environment was captured by the OMS as 3,000 images (500 frames). Fig. 11 shows a sampled frame of six input image sequences. Intrinsic camera parameters including geometric relations among fixed camera units are calibrated in advance by using a marker board and a 3-D laser measure [26]. Extrinsic camera parameters of the input image sequences are estimated using bundle adjustment by tracking both a small number of feature landmarks of known 3-D positions and a large number of natural features of unknown 3-D positions in input images across adjacent camera units [27]. Fig. 12 shows the recovered camera path that is used as an input for depth estimation. The curved line and pyramids denote the motion path of a camera unit and its posture at every 20 frames, respectively. The length



(a) appearance (b) view volume

Figure 10: Omnidirectional multi-camera system, Ladybug.

of the camera path is approximately 29 m. The accuracy of the estimated camera path is evaluated as 50 mm and 0.07° with regard to the camera position and posture, respectively [27].

By using input images and estimated camera parameters, omnidirectional depth maps are actually estimated by the hybrid method. After the detection of interest points for all the frames of six input image sequences by using the Harris operator, the depths of all the interest points are estimated using the TNIP score function. In this experiment, 1,750 interest points are detected on average in a single input image (10,500 points per frame). Interest points in the $(f-100)^{th}$ to the $(f+100)^{th}$ frames at every 2 frames (606 images, 101 frames) are used to estimate the depth data for the f^{th} frame. The size of the window W for TNIP (for initial estimation) and SSSD (for refinment) were set as 3×3 and 7×7, respectively, according to the result of the simulation described in Section 6.1.2. The searching range to find a maximum TNIP in

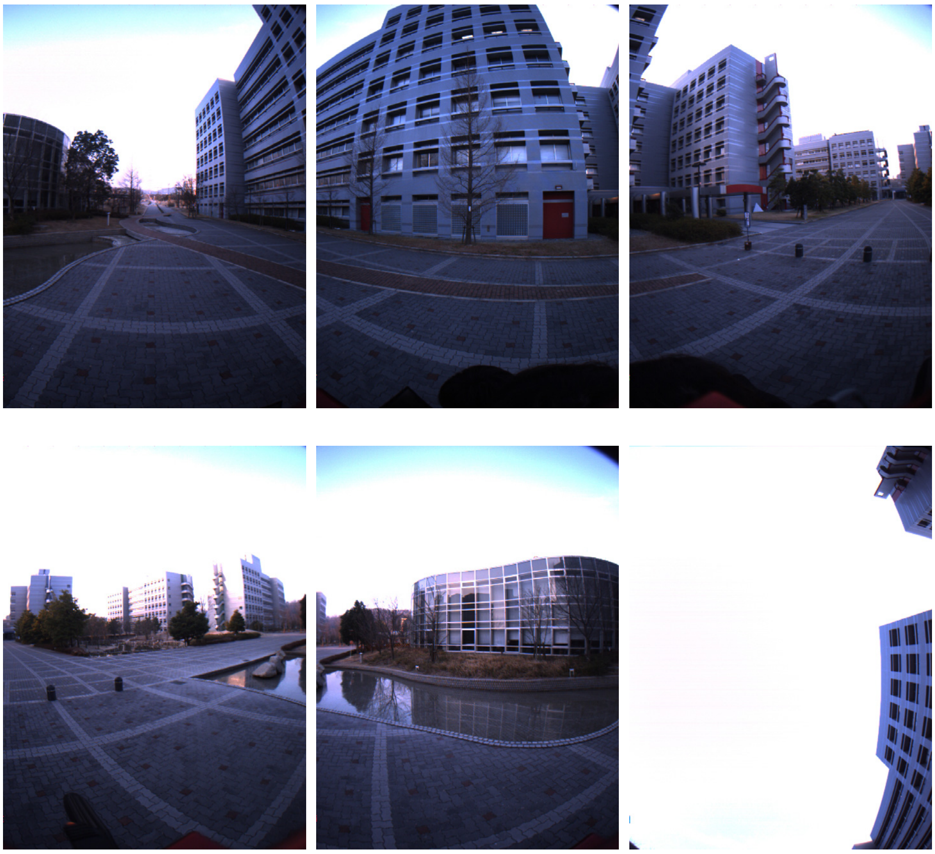
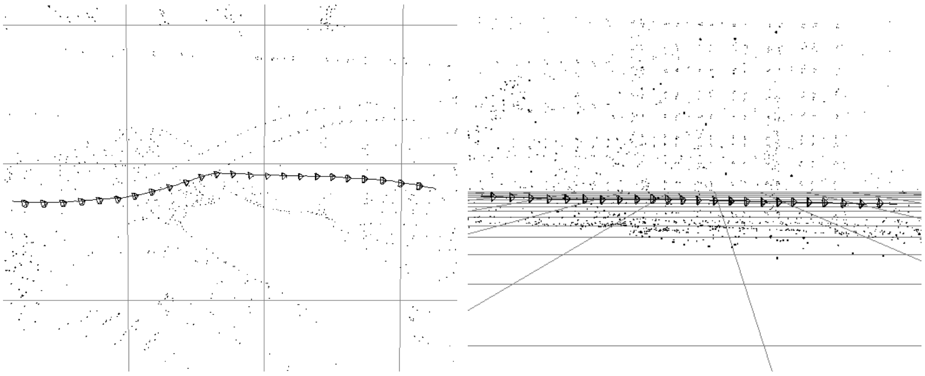


Figure 11: Sampled frame of input image sequences.



(a) top view (b) side view

Figure 12: Camera path of OMS used for input (29 m).

this stage is 1,000 mm (near) to 80,000 mm (far).

Next, low confidence depths are eliminated. The thresholds for T and U for outlier detection were set as 2.0 pixels and 0.3, respectively. In this experiment, approximately half of the estimated depths are rejected as outliers. As a result, on average, 700 depths for each image (4,500 depths per frame) were acquired. This is much larger than 88 depths for each image (530 depths per frame) that are computed at the structure from motion process for extrinsic camera parameter estimation in this experiment.

Fig. 13 shows the results of depth estimation for the images shown in Fig. 11. In this figure, the depth values are indicated by the intensity. Fig. 14 indicates the TNIP and SSSD scores for the six randomly selected interest points shown in Fig. 13. The solid vertical line in the graph for SSSD indicates the depth value given by the TNIP and the dotted line indicates the refined depth value determined by the SSSD. It can be confirmed from

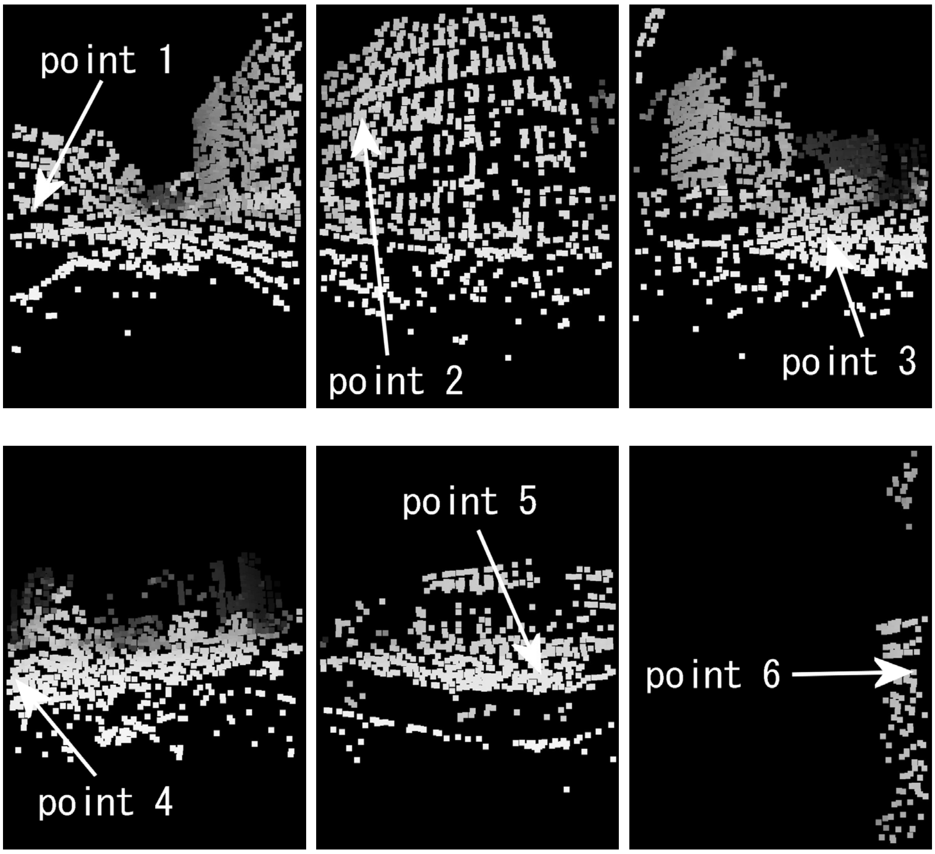
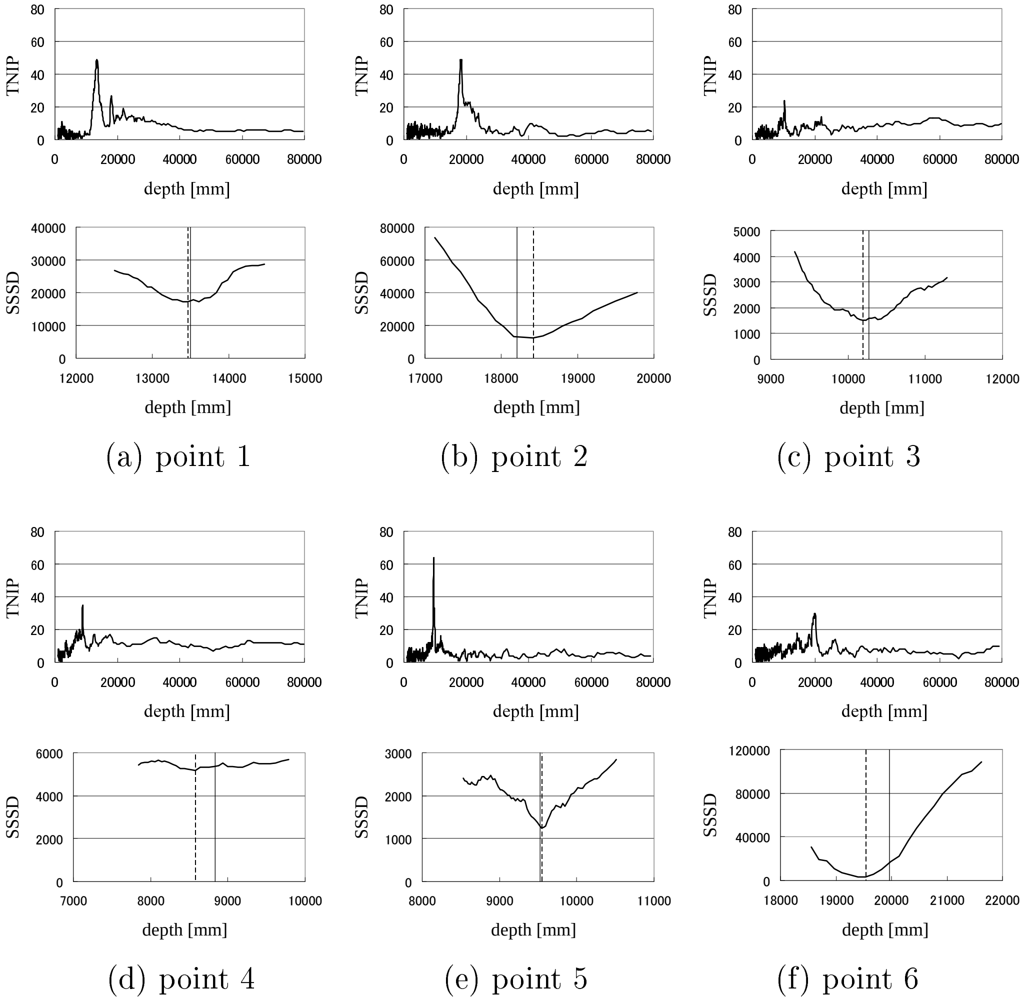


Figure 13: Result of depth estimation for the interest points shown in Fig. 11.

Fig. 14 that each TNIP plot has a single apparent peak at a certain depth value and there are no other comparable peaks. This clearly shows that depth estimation can be easily preformed for these interest points. In the refinement process by SSSD, except for point 4, there exists an apparent local minimum for each point and the TNIP gives a good initial value for refinement. The graph of SSSD for point 4 is comparatively flat because point 4 indicates a low-contrast texture on the ground. Even for such a texture, TNIP has a peak and the limitation of the searching range is thus considered to be preformed appropriately.

Finally, omnidirectional dense depth maps are generated using depth interpolation. For depth interpolation, first, 2-D feature points in omnidirectional images are triangulated by Delaunay's triangulation method [28]. After triangulation, the depth of each pixel is determined by computing the depth for 3-D plane for each triangle in 3-D space. Fig. 15 shows a panoramic image that is generated from the six input images shown in Fig. 11. Fig. 16 shows the corresponding dense depth map. By comparing these figures, it is observed that the depth map is correctly computed for most parts of the input image. However, some incorrect depths are also observed around the boundaries between the buildings and the sky. These incorrect results are caused by depth interpolation over different objects. In order to improve the result, region information in the input images should be considered as in the case of [29, 30].



(a) point 1

(b) point 2

(c) point 3

(d) point 4

(e) point 5

(f) point 6

Figure 14: TNIP and SSSD scores for depth searching.

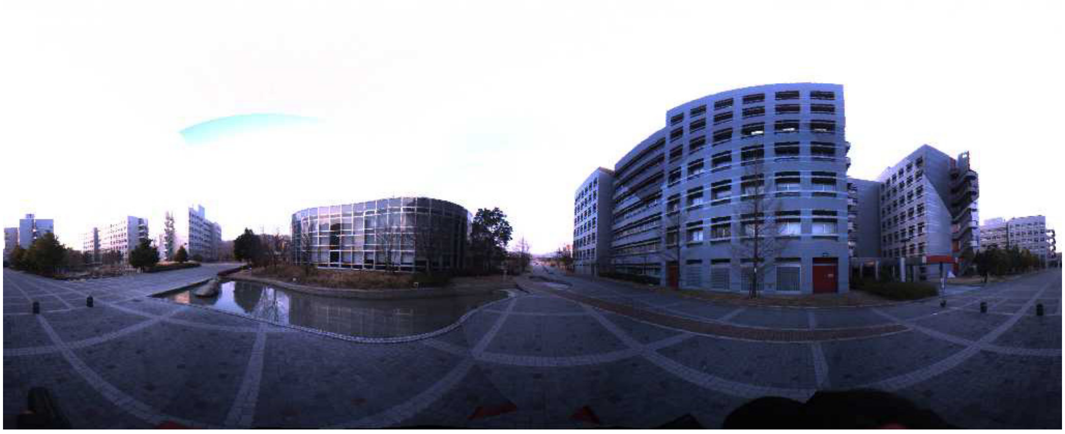


Figure 15: Panoramic image generated from the six images acquired by the OMS.

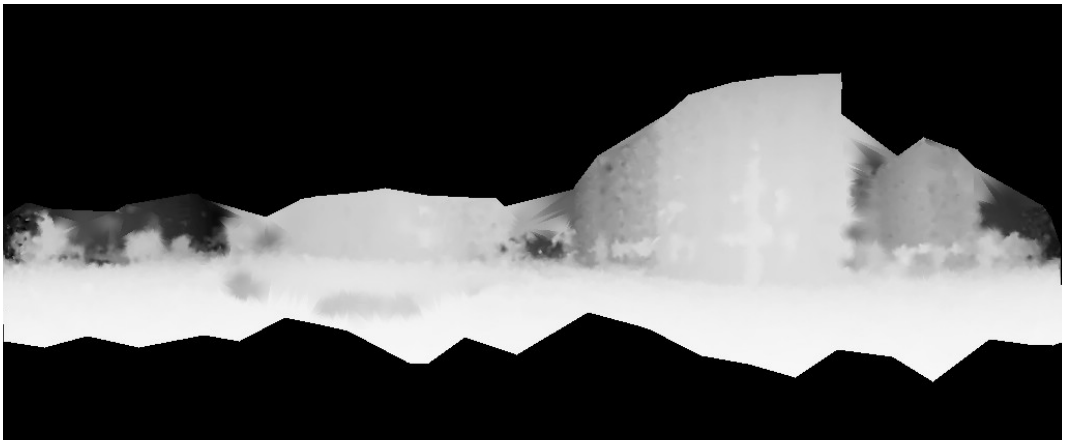


Figure 16: Generated dense depth map.

7. Conclusion

In this paper, a novel multi-baseline stereo for a moving camera has been proposed, where the depth can be determined by only counting the number of interest points. The raw TNIP function has a weakness in that highly accurate depth estimation is difficult due to feature detection errors in the target frame. In order to resolve this problem, we also propose the hybrid approach that refines the estimated depth using the SSSD function for a limited searching range. The proposed method is robust against occlusions and distortions, and its computational cost is also cheaper than that of the method based on the traditional SSSD function. We have experimentally verified our claims by using both synthetic and real image sequences. In future studies, the estimated depth maps will be integrated to reconstruct a 3-D model of a large outdoor environment.

References

[1] D. Scharstein and R. Szeliski: "A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithm," Int. J. of Computer Vision, Vol. 47, No. 3, pp. 7–42, 2002.

[2] M. Okutomi and T. Kanade: "A Multiple-baseline Stereo," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.15, No.4, pp. 353–363, 1993.

[3] S. B. Kang, J. A. Webb, C. Zitnick and T. Kanade: "A Multibaseline Stereo System with Active Illumination and Real-time Image Acquisition," Proc. Int. Conf. on Computer Vision, pp. 88–93, 1995.

[4] S. B. Kang and R. Szeliski: "3-D Scene Data Recovery using Omnidirectional Multibaseline Stereo," Int. Journal of Computer Vision, Vol.25, No.2, pp. 167–183, 1997.

[5] W. Zheng, Y. Kanatsugu, Y. Shishikui and Y. Tanaka: "Robust Depth-map Estimation from Image Sequences with Precise Camera Operation Parameters," Proc. Int. Conf. on Image Processing, Vol.II, pp. 764–767, 2000.

[6] T. Sato, M. Kanbara, N. Yokoya and H. Takemura: "Dense 3-D Reconstruction of an Outdoor Scene by Hundreds-baseline Stereo Using a Hand-held Video Camera," Int. Journal of Computer Vision, Vol.47, Nos.1–3, pp. 119–129, 2002.

[7] M. Okutomi, Y. Katayama and S. Oka: "A Simple Stereo Algorithm to Recover Precise Object Boundaries and Smooth Surface," Int. Journal of Computer Vision, Vol.47, Nos.1–3, pp. 261–273, 2002.

[8] K. Kutulakos and S. Seitz: "A Theory of Shape by Space Carving," Int. J. of Computer Vision, Vol. 38, No. 3, pp. 199–218, 2000.

[9] S. Seitz, B. Curless, J. Diebel, D. Scharstein and R. Szeliski: "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," Proc. Int. Conf. on Computer Vision and Pattern Recognition, pp. 519–528, 2006.

[10] A. Hornung and L. Kobbelt: "Robust and Efficient Photo-consistency Estimation for Volumetric 3D Reconstruction," Proc. European Conf. on Computer Vision, 2006.

[11] Y. Furukawa and J. Ponce: "Accurate, dense, and robust multiview stereopsis," Proc. Int. Conf. on Computer Vision and Pattern Recognition, pp. 1–8, 2007.

[12] M. Goesele, N. Snavely, B. Curless, H. Hoppe and S. Seitz: "Multi-View Stereo for Community Photo Collections," Proc. Int. Conf. on Computer Vision, pp. 14–20, 2007.

[13] Y. Furukawa, S. Seitz, B. Curless and R. Szeliski: "Manhattan-world Stereo," Proc. Int. Conf. on Computer Vision and Pattern Recognition, 2009.

[14] H. H. Baker: "Edge Based Stereo Correlation," Proc. Image Understanding Workshop, pp. 168–175, 1980.

[15] W. E. L. Grimson: "Computational Experiments with a Feature-based Stereo Algorithm," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.7, No.1, pp. 17–34, 1985.

[16] S. Pollard, M. Pilu, S. Hayes and A. Lorusso: "View Synthesis by Trinocular Edge Matching and Transfer," Image and Vision Computing, Vol.18, No.9, pp. 739–748, 2000.

[17] R. C. Bolles, H. H. Baker and D. H. Marimont: "Epipolar-plane Image Analysis: An Approach to Determining Structure from Motion," Int. Journal of Computer Vision, Vol.1, No.1, pp. 7–55, 1987.

[18] M. Okutomi and S. Sugimoto: "Shape Recovery of Rotating Object Using Weighted Voting of Spatio-temporal Images," Proc. Int. Conf. on Pattern Recognition, Vol.1, pp. 790–793, 2000.

[19] B. Triggs, P. McLauchlan, R. Hartley and A. Fitzgibbon: "Bundle Adjustment - a Modern Synthesis," Proc. Int. Workshop on Vision Algorithms, pp. 298–372, 1999.

[20] M. Okutomi and T. Kanade: "A Locally Adaptive Window for Signal Matching," Int. Journal of Computer Vision, Vol.7, No.2, pp. 143–162, 1992.

[21] S. B. Kang, R. Szeliski and J. Chai: "Handling Occlusions in Dense Multi-view Stereo," IEEE Conf. on Computer Vision and Pattern Recognition, Vol.1, pp. 103–110, 2001.

[22] M. Sanfourche, G. L. Benerais and F. Champagnat: "On the Choice of the Correlation Term for Multi-baseline Stereo-vision," Proc. British Machine Vision Conference, Vol.II, pp. 697–706, 2004.

[23] C. Harris and M. Stephens: "A Combined Corner and Edge Detector," Proc. Alvey Vision Conf., pp. 147–151, 1988.

[24] H. Moravec: "Towards Automatic Visual Obstacle Avoidance," Proc. Int. Joint Conf. on Artificial Intelligence, p. 584, 1977.

[25] D. G. Lowe: "Distinctive Image Features from Scale Invariant Keypoints," Int. Journal of Computer Vision, Vol.60, No.2, pp. 91–110, 2004.

[26] S. Ikeda, T. Sato and N. Yokoya: "High-resolution Panoramic Movie Generation from Video Streams Acquired by an Omnidirectional Multi-camera System," Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent System, pp. 155–160, 2003.

[27] T. Sato, S. Ikeda and N. Yokoya: "Extrinsic Camera Parameter Recovery from Multiple Image Sequences Captured by an Omni-directional Multi-camera System," Proc. European Conf. on Computer Vision, Vol.2, pp. 326–340, 2004.

[28] P. Heckbert Ed.: Graphics Gems IV, pp. 47–59, Academic Press, 1994.

[29] D.D. Morris and T. Kanade: "Image-consistent Surface Triangulation," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Vol.1, pp. 332–338, 2000.

[30] A. Nakatsuji, Y. Sugaya and K. Kanatani: “Mesh Optimization using an Inconsistency Detection Template,” Proc. Int. Conf. on Computer Vision, Vol.2, pp. 1148–1153, 2005.