論文 / 著書情報
Article / Book Information

Title	Terrain mapping under extreme light conditions with direct stereo matching method through aggregating matching costs by weight
Authors	Jianhua Li, Gen Endo, Edwardo F. Fukushima
Citation	Advanced Robotics, Volume 30, 13, pp. 861-876
Pub. date	2016, 3
Note	This is an electronic version of an article published in ADVANCED ROBOTICS, Volume 30, 13, pp. 861-876, 2016. ADVANCED ROBOTICS is available online at: www.tandfonline.com/Article DOI; http://dx.doi.org/10.1080/01691864.2016.1155483.
Note	このファイルは著者（最終）版です。 This file is author (final) version.

To appear in *Advanced Robotics*
Vol. 00, No. 00, February 2016, 1–19

FULL PAPER

Terrain Mapping under Extreme Light Conditions with Direct Stereo Matching Method through Aggregating Matching Costs by Weight

Jianhua Li*, Gen Endo and Edwardo F. Fukushima

Department of Mechanical and Aerospace Engineering,
Tokyo Institute of Technology, Tokyo 152-8552, Japan

(*v1.0 released February 2016*)

One of the biggest problems in applying stereo vision techniques in field robotics is how to acquire 3D terrain maps under extreme light conditions. Through multiple exposures, the dynamic range of images can be increased. In this paper, instead of using existing lighting enhancement methods such as exposure fusion to increase the texture of 2D image, we propose that the matching costs of the images grabbed with multiple exposures are directly summed by weight. Compared with the previous methods such as exposure fusion, with the proposed method, it is not necessary to fuse the 2D images captured with multiple exposures, and for each pixel of the matching image, the local information in its local window can be better retained. Since it is possible that the camera is moved between exposures when the images are grabbed, the images captured with multiple exposures are aligned to the image acquired with auto exposure. In order to evaluate the performance of the proposed method, two different stereo matching algorithms were used: a local window-based method and semi-global method. Through experiments in laboratory and outdoors with a stereo vision camera fixed on a tripod and held in the hand, it was verified that the proposed method consistently allowed more valid points to be obtained and the 3D model of terrain can be built more accurately. Especially when the local window-based method was used, the proposed method performed much better.

Keywords: multiple exposures; extreme light conditions; stereo matching; stereo vision camera

Index to information contained in this guide

*Corresponding author. Email: li.j.an@m.titech.ac.jp

1. Introduction

Humanitarian demining is the action of clearing mines and unexploded ordnance from an area of land to allow the local population to safely return to live there. In manual demining operations, a human deminer systematically scans the ground with a mine metal detector. This process is time consuming, expensive and can be dangerous to the deminers. The actual clearing of a minefield is a very risky task even for highly trained professionals. Most automatic demining systems tend to explode the ordnances without defusing them. However, this method is not totally safe and always followed by a careful manual inspection afterwards. A robotic system to assist human deminers, named Gryphon, has been developed at Tokyo Institute of Technology since 2002. The Gryphon platform consists of an all-terrain vehicle, mounted with a robotic manipulator that carries a mine metal detector, and it is able to autonomously scan the interested area. The control system of the Gryphon platform was developed to be as easy as possible to be used, because it is not feasible to use highly trained engineers in the field. This leads to the fact that Gryphon platform was intended to be used by personnel with a minimum basic training, working for governments of humanitarian agencies. The consequence for the fact is that the entire platform must be cheap, robust, and easy to maintain. Since no assumption of the grounds shape can be made a priori, it is necessary that system is able to perform 3D acquisition of the scene. Considering robustness to various field conditions, cost, precision and processing speed, a Bumblebee stereo vision camera from Point Grey Research was selected. The 3D terrain model is built with the stereo vision camera and the Gryphon automatically scans areas with moving mine sensors at a constant distance from the ground.

The luminance of an object is measured in lux or candelas per square meter ($1lux = 1cd/m^2$). The range of the luminance of an object is called the dynamic range and defined as the ratio of the maximum luminance value to the minimum luminance value within the specimen. Intensity values of scenes in the real world can have a very broad dynamic range. From outdoor shade to outdoor sunlight, the scene luminance could be changed from 100 lux to 100,000 lux. For a digital camera, it is possible that the image acquired with auto exposure saturates in some areas while keeping others visibly underexposed. This is particularly true for scenes that have areas of both low and high illumination. In field, for some lighting conditions, the stereo correspondence algorithm is unable to find enough features to perform a depth analysis. The scan can become totally impossible or even more, the manipulator is wrongly positioned and hits obstacles or mines. The problem was reproduced in laboratory and reported in [1]. The depth map calculated from the stereo pair lacked so many features that only a small fraction of it was actually used to compute the 3D information. The limitation comes from the camera's dynamic range, which represents the limits of luminance range that a given device can capture. The dynamic range of cameras is limited by the charge-coupled devices (CCD), analog-to-digital conversion (ADC) and film characteristics [2]. In [3], the authors pointed that sunlit scenes and the scenes with shiny materials and artificial light sources, often have extreme differences in radiance values that are impossible to capture without either under-exposing or saturating the film. To cover the full dynamic range in such a scene, a series of photographs can be taken with different exposures. Through multiple exposures, the dynamic range of images is increased. By using a reduced exposure time, one may sacrifice lowlight detail in exchange for improved detail in areas of high illumination, and this is demonstrated in the short exposure image of Figure 3. Similarly, by increasing exposure time, a better representation of lowlight areas may be gotten, at the cost of losing information in areas of high illumination and an example of this is shown in the long exposure image of Figure 3.

In [4], a system architecture was introduced for terrain mapping using a stereo vision camera. As presented in Figure 1, traditionally, the 2D images captured with multiple exposures are fused with exposure fusion [5]. With the resulting fused images the disparity image is computed through stereo matching and the 3D terrain map is reconstructed. In our previous work [6], instead of using existing lighting enhancement methods such as exposure fusion to increase the

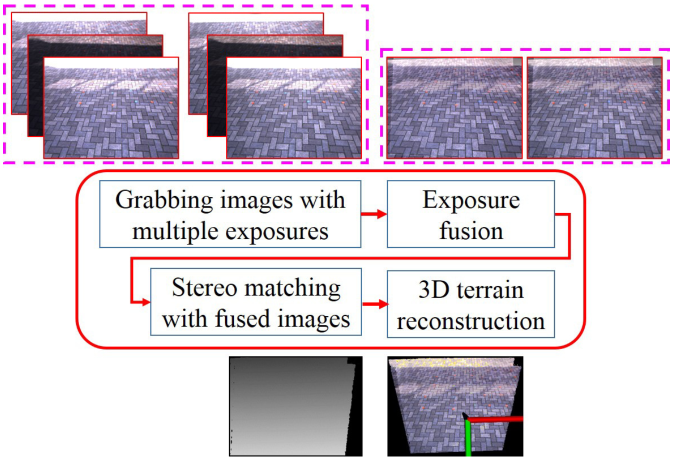

Grabbing images with multiple exposures
Exposure fusion
Stereo matching with fused images
3D terrain reconstruction

Figure 1. The previous system architecture for 3D terrain mapping using a stereo vision camera.

texture of the 2D image, the stereo matching was directly done using the images grabbed with multiple exposures.

In the real application, it is possible that the camera is moved when the images are grabbed with multiple exposures. So it is crucial to properly align the input images before fusing a high dynamic range image. Image registration is the process of overlaying images (two or more) of the same scene taken at different times, from different viewpoints, and/or by different sensors [7]. There are two major image alignment algorithms: pixel-based methods and feature-based alignment methods. Feature-based approaches have the advantage of being more robust against scene movement, and are potentially faster if implemented the right way [8]. In [9], after capturing high dynamic range images from a set of photographs taken at different exposures, the key-points or feature-points in these images were searched. The key-points were used to find matrices, which transform a set of images to a single coordinate system.

In [10], a taxonomy of dense, two-frame stereo methods was presented. This taxonomy is designed to assess the different components and design decisions made in individual stereo algorithms. The stereo algorithms generally perform (subsets of) the following four steps [10]:

- Matching cost computation;
- Cost aggregation;
- Disparity computation or optimization;
- Disparity refinement.

In this paper, we focus on the second step ”Cost aggregation” and improve our former method [6]. The images grabbed with short and long exposures are aligned to the image captured with auto exposure. The matching costs of the resulting registered images and the image grabbed with auto exposure are directly summed by weight. In order to evaluate the performance of our proposed method, two different stereo matching algorithms were used: a local, window-based method and semi-global method. Through experiments in laboratory and outdoors with a stereo vision camera fixed on a tripod and held in the hand, it was verified that with the proposed method, more valid 3D points could be obtained and the terrain maps could be reconstructed more accurately. Especially when the local window-based method was used, the proposed method performed much better. The remainder of the paper is structured as follows. Section 2 introduces the proposed method in detail and the experimental results are presented in Section 3. In section 4 we present conclusion.

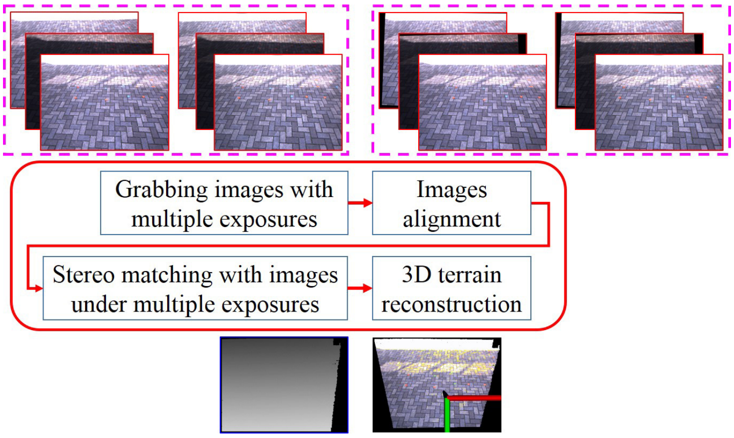

Grabbing images with multiple exposures
Images alignment
Stereo matching with images under multiple exposures
3D terrain reconstruction

Figure 2. The proposed system architecture for 3D terrain mapping using a stereo vision camera.

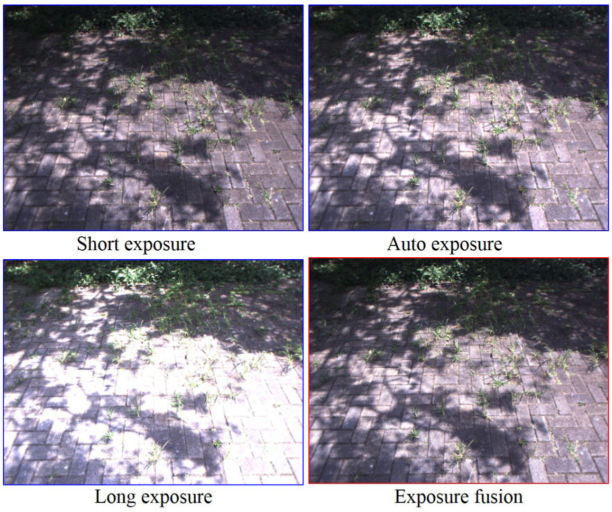


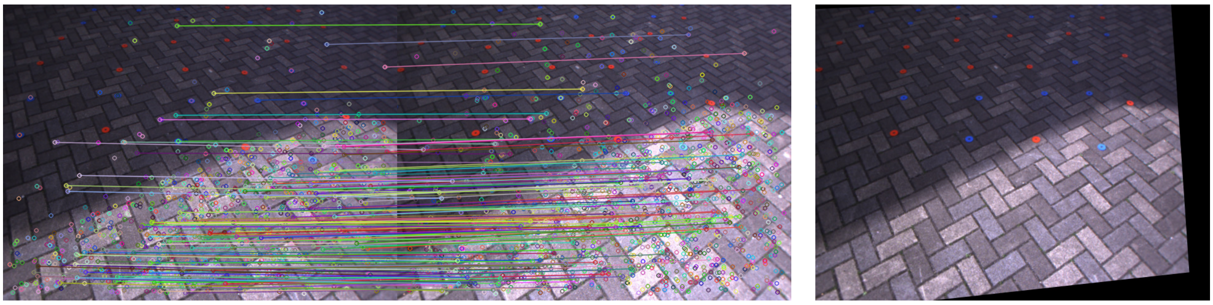
Figure 3. The images were grabbed with multiple exposures. With exposure fusion, these images were fused.

2. Proposed system architecture for 3D terrain mapping with a stereo vision camera

The proposed system architecture for 3D terrain mapping with a stereo vision camera is presented in Figure 2 and introduced in detail in this section. The images grabbed with multiple exposures are aligned to the image captured with auto exposure. Compared with the previous system architecture presented in Figure 1, with the proposed system architecture, stereo matching is directly done with the resulting registered images and the exposure fusion is not needed.

2.1 *Acquiring images with multiple exposures*

In order to acquire the images with multiple exposures, it is important to properly set the exposure parameters of camera, alternating between a long exposure to capture the shadows and a short exposure to capture the highlights. Using a method described in [11], the shutter times for short and long exposures are set. We assume that the brightness value is in the range of 0 to 255. For the short exposure, it is required that fewer than p_{short} (e.g. 1%) of the pixels in the image are bright which have values above B_{short} (e.g. 217). If there are too many bright pixels, the exposure time is decreased for the subsequent short exposures. Similarly, for the long exposure it is required that fewer than p_{long} (e.g. 1%) of pixels are dark which have values less



(a) Key points detected with SURF and matched pairs of key points.

(b) Aligned image.

Figure 4. The key points of the images grabbed in short (or long) exposure and auto exposure are detected with SURF. After matching the descriptor vectors of the key points, the key points pairs are obtained. RANSAC is performed to estimate the homography matrix and the image acquired with short (or long) exposure is aligned to the image acquired with auto exposure according to the homography matrix.

than B_{long} (e.g. 38), otherwise the exposure time is increased for the subsequent long exposures. The camera gain is kept as low as possible to minimize noise, only raising it when the camera shutter time setting is not available for the camera [11]. One example is shown in Figure 3, and the images grabbed with multiple exposures are presented.

2.2 *Image alignment*

In the real application, since it is possible that the camera is moved between exposures when the images are grabbed, it is important to register the images. With a method similar to [12, 13], the correspondence relationship between the images grabbed with short (or long) exposure and auto exposure are calculated with image alignment algorithm and the image acquired with short (or long) exposure is aligned to the image acquired with auto exposure. As shown in Figure 4, the key points of the images grabbed in short (or long) exposure and auto exposure are detected with Speeded Up Robust Features (SURF) [14]. After matching the descriptor vectors of the key points, the key points pairs are obtained and denoted as (x_i, y_i) in the image with short (or long) exposure and $(x_i^{'}, y_i)^{'}$ in the image with auto exposure respectively. They are related with (1), where $\boldsymbol{H}$ is an arbitrary 3x3 matrix and itself homogeneous [8]. Random sample consensus (RANSAC) [15] is performed to estimate the homography matrix $\boldsymbol{H}$ through solving the optimization problem (2), where N_R is the number of the key points pairs which are used to estimate the parameter $\boldsymbol{H}$. According to the homography matrix $\boldsymbol{H}$, the image acquired with short (or long) exposure is aligned to the image acquired with auto exposure.

$$\begin{bmatrix} x_i^{'} \\ y_i^{'} \\ 1 \end{bmatrix} \sim \boldsymbol{H} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}, \tag{1}$$

where

$$\boldsymbol{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}.$$

$$\min \sum_{i=1}^{N_R} (x_i^{'} - \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + h_{33}})^2 + (y_i^{'} - \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + h_{33}})^2 \tag{2}$$

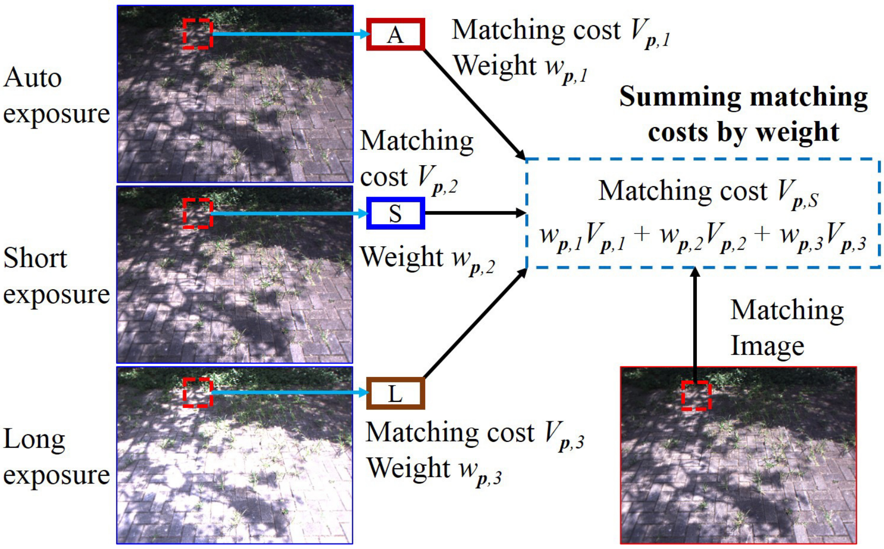


Figure 5. For each pixel $\boldsymbol{p}$, the matching costs $V_{\boldsymbol{p},1}$, $V_{\boldsymbol{p},2}$ and $V_{\boldsymbol{p},3}$ for the images captured with auto, short and long exposures can be calculated respectively. For the matching image, its matching cost $V_{\boldsymbol{p},S}$ is directly summed by weight.

2.3 *Proposed method of aggregating matching cost: matching costs of the images grabbed with multiple exposures are directly summed by weight*

In this paper, matching cost was defined based on intensity (luminance) instead of colour, which is stored as 8-bit unsigned integers in the range of 0 to 255. It is simple to extend this matching cost to colour by computing the costs for each colour channel separately and then summing the matching costs over all channels. Matching cost can be calculated with the methods such as Absolute Difference (AD), Squared Difference (SD), Census Transform (CT) and so on. As shown in Figure 5, for the image of the $k-th$ image (auto, short and long exposures in sequence, $k = 1, 2, 3$), for each pixel $\boldsymbol{p}$ ($\boldsymbol{p} = (x, y)$), its matching cost is defined to be $V_{\boldsymbol{p},k}$. The matching cost of the matching image is defined as $V_{\boldsymbol{p},S}$.

For the grayscale image of the $k - th$ image, the intensity of the pixel $\boldsymbol{p}$ is defined as $I(\boldsymbol{p}, k)$ ($0 \leq I(\boldsymbol{p}, k) \leq 255$) and the exposure quality $\phi^e_{\boldsymbol{p},k}$ is calculated with (3) based on how close it is to 127.5 [5]. σ_e was set to be 0.2 in this paper. The exposure quality weight $w^e_{\boldsymbol{p},k}$ is computed with (4).

$$\phi^e_{\boldsymbol{p},k} = exp(-\frac{(I(\boldsymbol{p}, k) - 127.5)^2}{2(255\sigma_e)^2}) \tag{3}$$

$$w^e_{\boldsymbol{p},k} = \frac{\phi^e_{\boldsymbol{p},k}}{\sum_{k=1}^{3} \phi^e_{\boldsymbol{p},k}} \tag{4}$$

For each pixel $\boldsymbol{p}$, $N(\boldsymbol{p})$ is the set of pixels surrounding it in its neighborhood with the window size of L_w x L_h pixels, where L_w and L_h are the width and height of the window in pixels respectively. For the grayscale image of the $k - th$ image, the intensity difference of the pixel $\boldsymbol{p}$ is defined as $S^c_{\boldsymbol{p},k}$ and calculated with (5) by comparing its intensity with the pixels in its local neighborhood $N(\boldsymbol{p})$. The intensity diversity $\phi^c_{\boldsymbol{p},k}$ is calculated with (6) and σ_c was set to be 0.2 in this paper. The intensity diversity weight $w^c_{\boldsymbol{p},k}$ is computed with (7).

$$S^c_{\boldsymbol{p},k} = \sum_{\boldsymbol{q} \in N(\boldsymbol{p})} f(I(\boldsymbol{p}, k), I(\boldsymbol{q}, k)), \tag{5}$$

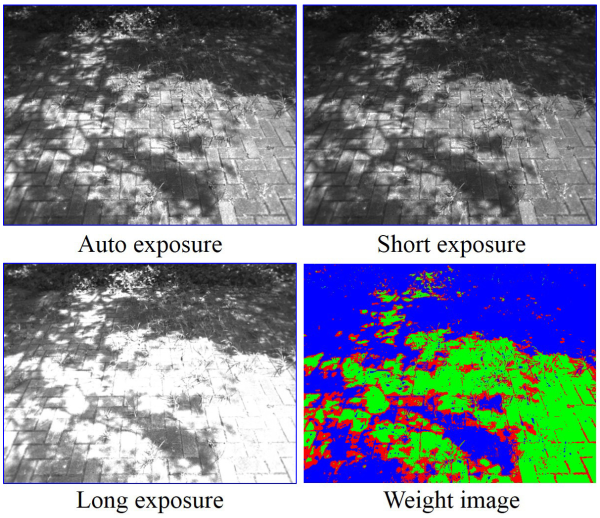


Figure 6. For each pixel $\boldsymbol{p}$ of the images grabbed with auto, short and long exposures, its weight $w_{\boldsymbol{p},k}$ was calculated with the proposed method. The intensity diversity weight of each pixel was calculated with a window of 15x15 pixels. In the weight image, for each pixel $\boldsymbol{p}$, the red colour means that the auto exposure image has the biggest weight, the green colour means that the short exposure image has the biggest weight and the blue colour means that the long exposure image has the biggest weight. It is noticed that the pixel which is well exposured has the biggest weight.

where

$$f(x,y) = \begin{cases} 1 \text{ if } x < y, \\ 0 \text{ else.} \end{cases}$$

$$\phi^c_{\boldsymbol{p},k} = exp(-\frac{(S^c_{\boldsymbol{p},k} - 0.5L_c)^2}{2(L_c\sigma_c)^2}), \tag{6}$$

where

$$L_c = L_w L_h - 1.$$

$$w^c_{\boldsymbol{p},k} = \frac{\phi^c_{\boldsymbol{p},k}}{\sum_{k=1}^3 \phi^c_{\boldsymbol{p},k}} \tag{7}$$

For each pixel $\boldsymbol{p}$ of the $k-th$ image, its weight $w_{\boldsymbol{p},k}$ is calculated with (8) through summing the exposure quality weight $w^e_{\boldsymbol{p},k}$ and intensity diversity weight $w^c_{\boldsymbol{p},k}$. λ_c was set to be 0.1 in this paper. As shown in Figure 5, for each pixel $\boldsymbol{p}$, the matching cost $V_{\boldsymbol{p},S}$ of the matching image is calculated with (9) based on the matching costs $V_{\boldsymbol{p},1}$, $V_{\boldsymbol{p},2}$, $V_{\boldsymbol{p},3}$ and the weights $w_{\boldsymbol{p},1}$, $w_{\boldsymbol{p},2}$, $w_{\boldsymbol{p},3}$.

$$w_{\boldsymbol{p},k} = w^e_{\boldsymbol{p},k} + \lambda_c w^c_{\boldsymbol{p},k} \tag{8}$$

$$V_{\boldsymbol{p},S} = \sum_{k=1}^{3} w_{\boldsymbol{p},k} V_{\boldsymbol{p},k} \tag{9}$$

One example is shown in Figure 6. For each pixel $\boldsymbol{p}$ of the images grabbed with auto, short and long exposures, its weight $w_{\boldsymbol{p},k}$ was calculated with the proposed method. The intensity diversity weight of each pixel was calculated with a window of 15x15 pixels. In the weight image, for each pixel $\boldsymbol{p}$, the red colour means that the auto exposure image has the biggest weight, the green colour means that the short exposure image has the biggest weight and the blue colour means that the long exposure image has the biggest weight. From this figure, it is noticed that the pixel which is well exposured has the biggest weight.

3. Experimental results

In order to evaluate the performance of the stereo matching algorithm described in this paper, a Bumblebee XB3 stereo vision camera from Point Grey Research was used and the experiments were done in laboratory and outdoors. The Bumblebee XB3 stereo vision camera is a 3-sensor multi-baseline IEEE-1394b (800Mb/s) stereo vision camera designed for improved accuracy and pre-calibrated for lens distortions and camera misalignments. The colour images grabbed with stereo vision camera were converted to gray images, with which the matching costs of the matching image were calculated. The stereo matching with image size of 640x480 pixels and disparity range of 100 pixels was done with local window-based method and SGM respectively.

3.1 *Stereo matching methods*

Census transform was used to calculate the matching cost in this paper. It is able to deal with radiometric changes since it is a non-parametric local transform which relies on the relative ordering of local intensity values and not on the intensity values themselves [16]. The census transform encodes the local neighborhood (e.g. window with a window size of 11x11 pixels) around each pixel into a bit cost that only stores whether the compared neighboring pixel has a lower value than the center pixel or not. For the matching image (left image or right image of a stereo pair), its matching cost between two pixels in the matching image and reference image of a stereo pair is the Hamming distance of their census transform in their local windows.

Using the image alignment method described in this paper, the images grabbed with short and long exposures were aligned to the image grabbed with auto exposure. For the image grabbed with auto exposure and the registered images of the photographs captured with short and long exposures, for each pixel $\boldsymbol{p}$, its matching costs $V_{\boldsymbol{p},1}$, $V_{\boldsymbol{p},2}$ and $V_{\boldsymbol{p},3}$ can be computed. The image grabbed with auto exposure and the registered images of the photographs captured with short and long exposures were fused with exposure fusion [5], and for the resulting fused image, the matching cost of each pixel $\boldsymbol{p}$ is defined as $V_{\boldsymbol{p},E}$. For each pixel $\boldsymbol{p}$, the matching cost of the matching image is defined as $V_{\boldsymbol{p},M}$. In order to evaluate the performance of the proposed method, the matching cost $V_{\boldsymbol{p},M}$ is calculated with the following four methods.

- Auto exposure. $V_{\boldsymbol{p},M}$ is set to be $V_{\boldsymbol{p},1}$.
- Exposure fusion. $V_{\boldsymbol{p},M}$ is set to be $V_{\boldsymbol{p},E}$.
- Multiple images. $V_{\boldsymbol{p},M}$ is set to be $V_{\boldsymbol{p},D}$, which is calculated with (10) through directly summing the matching costs $V_{\boldsymbol{p},1}$, $V_{\boldsymbol{p},2}$ and $V_{\boldsymbol{p},3}$. .

$$V_{\boldsymbol{p},D} = \sum_{k=1}^{3} V_{\boldsymbol{p},k} \tag{10}$$

- Multiple images by weight. $V_{\boldsymbol{p},M}$ is set to be $V_{\boldsymbol{p},S}$, which is calculated with (9) using the proposed method. The intensity diversity weight of each pixel was calculated with same window size of census transform.

Since the performance of a matching cost depends on the algorithm which uses it, two different stereo algorithms were used: a local window-based method [16] and semi-global matching (SGM) [17]. For the local window-based method, after computing the matching costs, the disparity with the lowest matching cost was selected with winner-takes-all. The SGM is adopted as the optimization technique to stereo matching for it is more advantageous since it delivers denser results with far fewer outliers. Many applications have proved that SGM is of high quality and can reconstruct thin or small objects. In this paper, the stereo matching was done with SGM by summing the matching costs in four directions (up, down, left and right).

The uniqueness check invalidates disparities if the minimum cost is not unique. With the method described in [18], the sub-pixel disparity refinement is obtained through interpolating the three matching costs (the winning cost value and its neighbors). The occlusions and mismatches are distinguished by the left/right consistency check, which invalidates disparities if the disparity with the left/right images stereo matching and its corresponding disparity of with the right/left images stereo matching differ by more than one pixel.

3.2 *Experiments of mapping a flat terrain*

Experiments were done to reconstruct the 3D map of a flat terrain. Since the terrain for mapping is almost flat, in order to evaluate the stereo matching result, the best-fit plane was estimated with RANSAC. The distance of a point to the estimated plane is defined as $d_t (d_t \geq 0)$. Points with the distance d_t smaller than D_T were considered to be valid. D_T was set to be 40.0 mm in this paper. The number of valid points is defined to be N_V and it is the most important criteria to evaluate the performance of the stereo matching methods. The average distance to the estimated plane is define as d_V and calculated with (11) using the valid points.

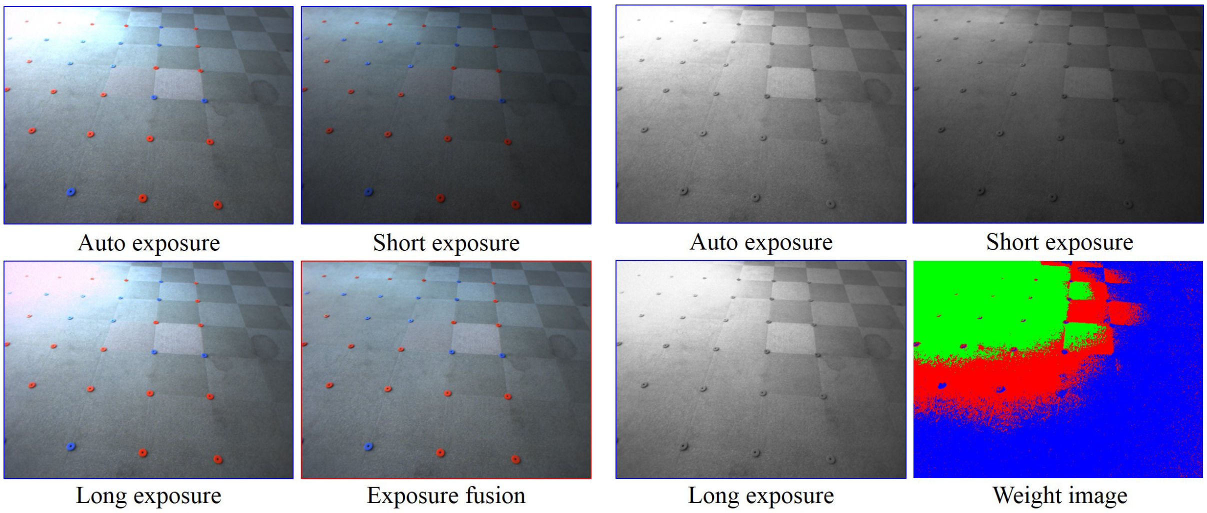
$$d_V = \frac{1}{N_V} \sum_{d_t \leq D_T} d_t \tag{11}$$

3.2.1 Experiment in laboratory with stereo vision camera fixed on a tripod

Experiments were done in laboratory with the stereo vision camera fixed on a tripod and the grabbed images of the right camera are presented in Figure 7(a). The image grabbed with auto, short and long exposures were fused with exposure fusion and the resulting fused image is shown in Figure 7(a). The colour images were converted to grayscale images as shown in Figure 7(b). As an example, using the method described in this paper, the intensity diversity weight of each pixel was calculated with a window of 15x15 pixels and the resulting weight image is shown in Figure 7(b).

First, stereo matching was done with the local window-based method. For example, with a window of 15x15 pixels, the disparity images calculated with four different methods are shown in Figure 8(a). As it is overexposure and texture-less in the top left of the image grabbed with auto exposure, with the method “auto exposure”, the disparity values for the pixels in this part were not calculated. With the window size changed from 7x7 to 23x23 pixels, Figure 9 shows the valid point number N_V and the average distance to the estimated plane d_V. It shows that compared with the methods “Auto exposure” and “Exposure fusion”, with the methods “Multiple images” and “Multiple images by weight”, more valid points can be obtained and the average point to estimated plane distance becomes smaller. Especially when the window size is small, the methods “Multiple images” and “Multiple images by weight” performed much better.

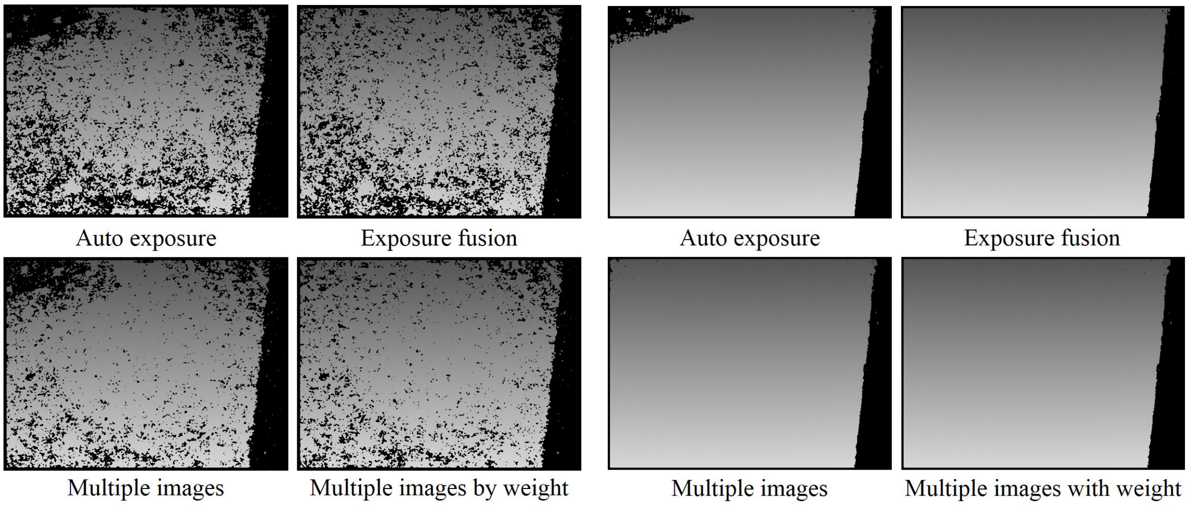
Next, stereo matching was done with SGM. For example, with a window of 11x11 pixels, the disparity images computed with four different methods are shown in Figure 8(b). Since it is overexposure in the top left of the image grabbed with auto exposure, even when SGM was used, the disparity values for the pixels in this part still were not calculated with the method “Auto exposure”. With the window size changed from 7x7 to 15x15 pixels, Figure 10 presents the valid point number N_V and the average distance to the estimated plane d_V. It shows that the valid point numbers of the methods “Exposure fusion”, “Multiple images” and “Multiple images by weight” are quite close and bigger than the result of the method “Auto exposure”. However, the proposed method “Multiple images by weight” performs best with the smallest



(a) The images were grabbed with multiple exposures. With exposure fusion, these images were fused.

(b) Using the method described in this paper, with the intensity diversity weight of each pixel was calculated with a window of 15x15 pixels, the weight image was calculated.

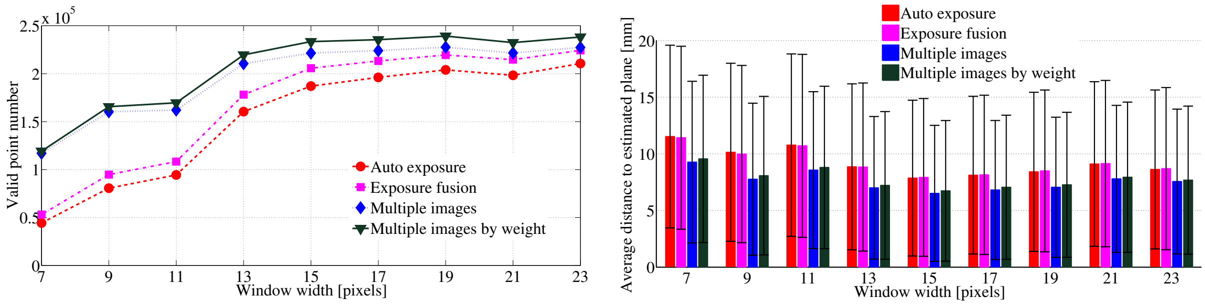
Figure 7. The images were acquired in laboratory with the stereo vision camera fixed on a tripod.



(a) With the window size of 15x15 pixels, the disparity images were calculated with local window-based method.

(b) With the window size of 11x11 pixels, the disparity images were calculated with SGM.

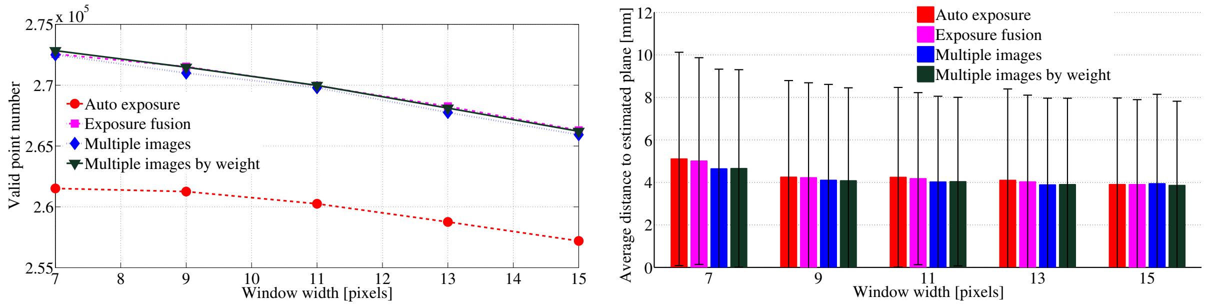
Figure 8. For the images shown in Figure 7, using four methods to calculate the matching costs of the matching images, the disparity images were calculated.



(a) Valid point number N_V.

(b) Average distance to the estimated plane d_V.

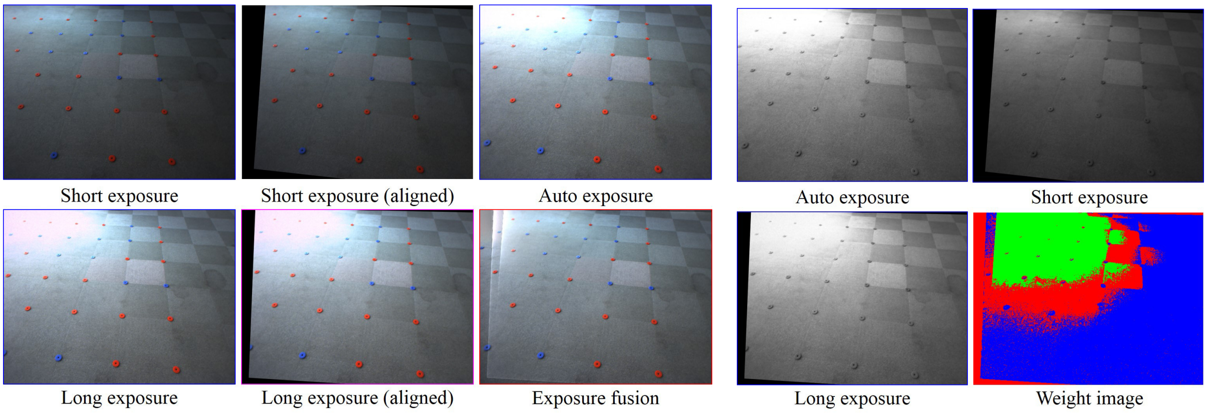
Figure 9. For the images shown in Figure 7, using four methods to calculate the matching costs of the matching images, the stereo matching was computed with local window-based method. The window size is changed from 7x7 to 23x23 pixels.



(a) Valid point number N_V.

(b) Average distance to the estimated plane d_V.

Figure 10. For the images shown in Figure 7, using four methods to calculate the matching costs of the matching images, the stereo matching was done with SGM. The window size is changed from 7x7 to 15x15 pixels.



(a) The images were grabbed with multiple exposures. The images grabbed with short and long exposures were aligned to the image captured with auto exposure. With exposure fusion, the image grabbed with auto exposure and the registered images of the photographs captured with short and long exposures were fused.

(b) Using the method described in this paper, with the intensity diversity weight of each pixel was calculated with a window of 15x15 pixels, the weight image was calculated.

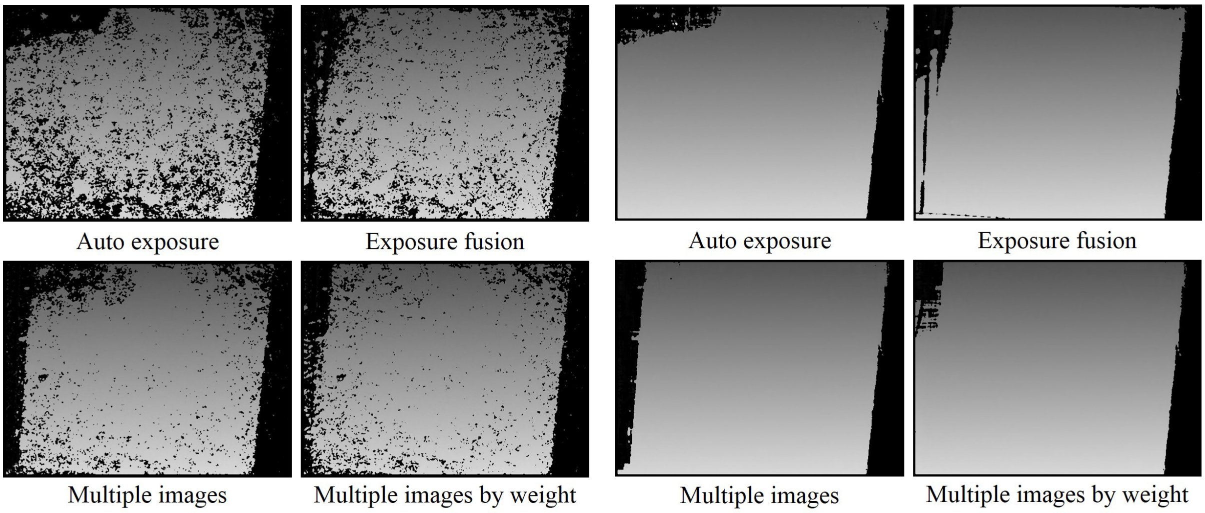
Figure 11. The images were acquired in laboratory with the stereo vision camera held in the hand.

average distance to the estimated plane d_V.

3.2.2 Experiments in laboratory with the stereo vision camera held in the hand

The experiments were done in laboratory with the stereo vision camera held in the hand and the grabbed images of the right camera are shown in Figure 11(a). With the image alignment mothod described in this paper, the key points of the images grabbed with short (long) and auto exposures were detected with SURF and matched with RANSAC. The images grabbed with short and long exposures were aligned to the image captured with auto exposure and the aligned images are shown in Figure 11(a). The image grabbed with auto exposure and the registered images of the photographs captured with short and long exposures were fused with exposure fusion and the resulting fused image is shown in Figure 11(a). The grayscale images are shown in Figure 11(b). As an example, the intensity diversity weight was calculated with a window of 15x15 pixels and the resulting weight image is shown in Figure 11(b).

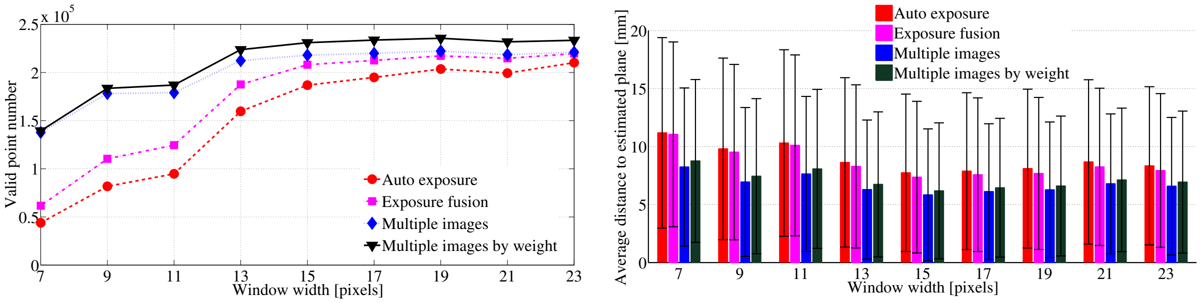
First, stereo matching was done with local window-based method. For example, with a window of 15x15 pixels, the disparity images calculated with four different methods are shown in Figure 12(a). With the window size changed from 7x7 to 23x23 pixels, Figure 13 presents the valid point number N_V and the average distance to the estimated plane d_V. It shows that compared with the methods "Auto exposure" and "Exposure fusion", with the methods "Multiple images" and "Multiple images by weight", more valid points are obtained and the average point to estimated plane distance becomes smaller. Especially when the window size is small, the methods "Multiple images" and "Multiple images by weight" performed much better.



(a) With the window size of 15x15 pixels, the disparity images were calculated with local window-based method.

(b) With the window size of 11x11 pixels, the disparity images were calculated with SGM.

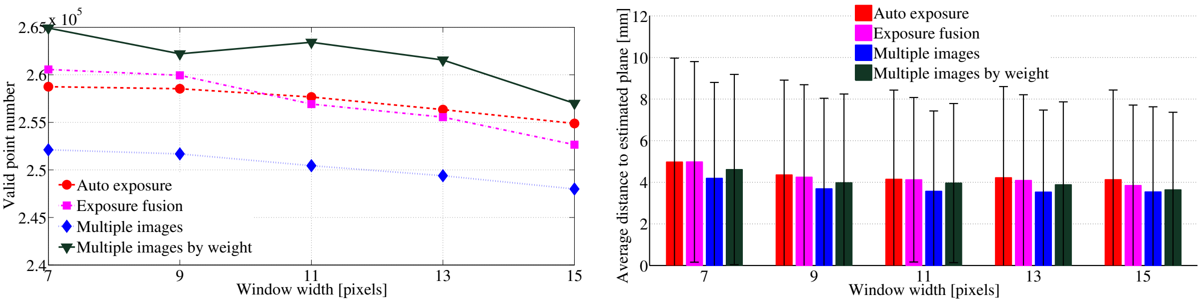
Figure 12. For the images shown in Figure 11, using four methods to calculate the matching costs of the matching images, the disparity images were calculated.



(a) Valid point number N_V.

(b) Average distance to the estimated plane d_V.

Figure 13. For the images shown in Figure 11, using four methods to calculate the matching costs of the matching images, the stereo matching was computed with local window-based method. The window size is changed from 7x7 to 23x23 pixels.

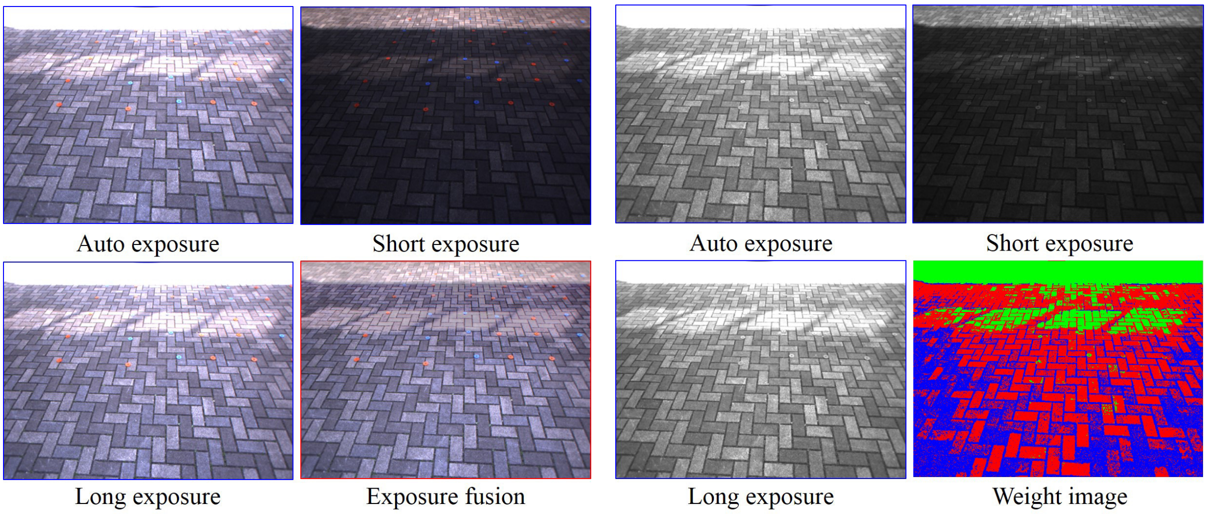


(a) Valid point number N_V.

(b) Average distance to the estimated plane d_V.

Figure 14. For the images shown in Figure 11, using four methods to calculate the matching costs of the matching images, the stereo matching was done with SGM. The window size is changed from 7x7 to 15x15 pixels.

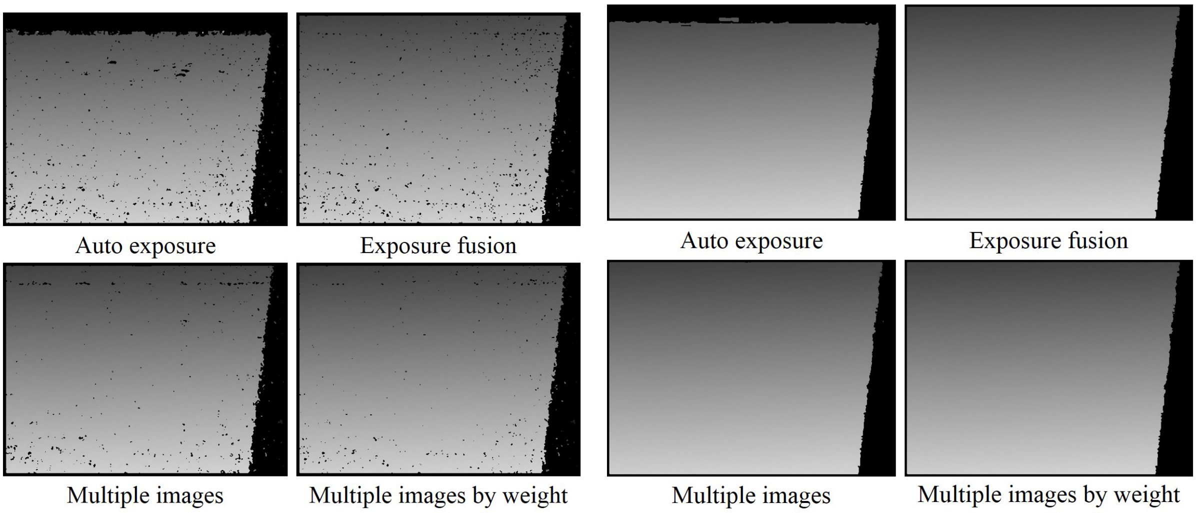
Next, stereo matching was done with SGM. For example, with a window of 11x11 pixels, the disparity images computed with four different methods are shown in Figure 12(b). With the window size changed from 7x7 to 15x15 pixels, Figure 14 illustrates the valid point number N_V and the average distance to the estimated plane d_V. It shows that the proposed method "Multiple images by weight" performs best with biggest N_V and its d_V is smaller than the results of the methods "Auto exposure" and "Exposure fusion".



(a) The images were grabbed with multiple exposures. With exposure fusion, these images were fused.

(b) With the intensity diversity weight was calculated with a window of 15x15 pixels, the weight image was calculated.

Figure 15. The images were acquired outdoors with the stereo vision camera fixed on a tripod.



(a) With the window size of 15x15 pixels, the disparity images were calculated with local window-based method.

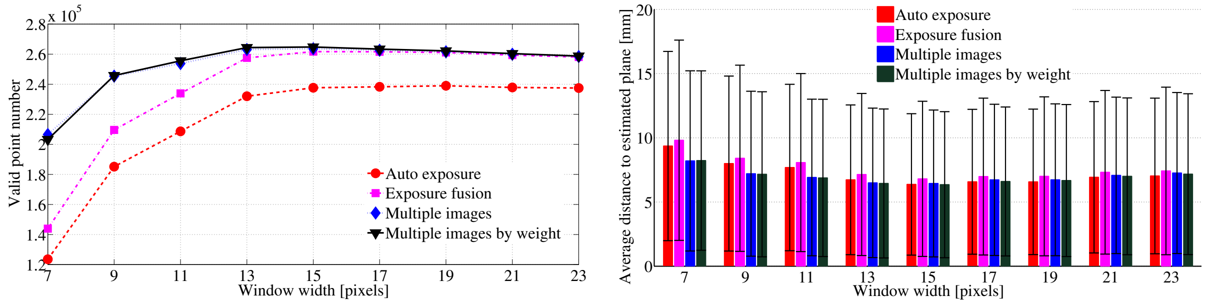
(b) With the window size of 11x11 pixels, the disparity images were calculated with SGM.

Figure 16. For the images shown in Figure 15, using four methods to calculate the matching costs of the matching images, the disparity images were calculated.

3.2.3 Outdoor experiments with the stereo vision camera fixed on a tripod

The experiments were done outdoors with the stereo vision camera fixed on a tripod. The grabbed images of the right camera and the fused image are shown in Figure 15 (a). The grayscale images are presented in Figure 15(b). As an example, the intensity diversity weight was calculated with a window of 15x15 pixels and the resulting weight image is shown in Figure 15(b).

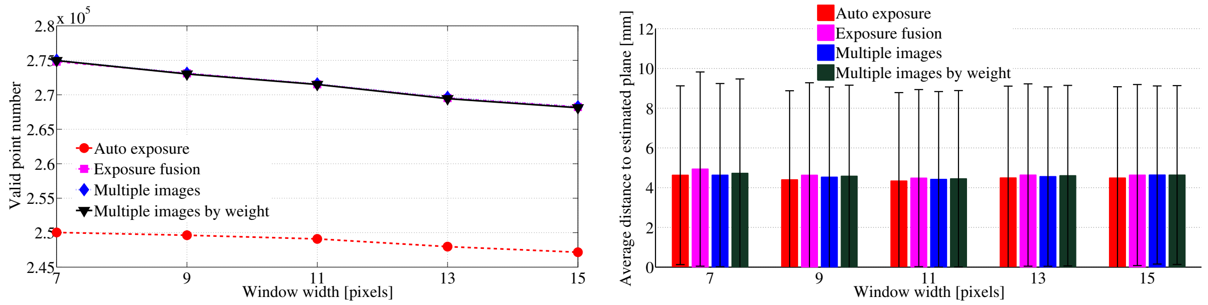
First, stereo matching was done with local window-based method. For example, with a window of 15x15 pixels, the disparity images computed with four different methods are shown in Figure 16(a). Since it is overexposure in the top of the image grabbed with auto exposure, the disparity values for the pixels in this part were not calculated with the method “Auto exposure”. With the window size changed from 7x7 to 23x23 pixels, Figure 17 presents the valid point number N_V and the average distance to the estimated plane d_V. It depicts that compared with the method “Exposure fusion”, with the methods “Multiple images” and “Multiple images by weight”, more valid points are obtained and the average point to estimated plane distance



(a) Valid point number N_V.

(b) Average distance to the estimated plane d_V.

Figure 17. For the images shown in Figure 15, using four methods to calculate the matching costs of the matching images, the stereo matching was computed with local window-based method. The window size is changed from 7x7 to 23x23 pixels.



(a) Valid point number N_V.

(b) Average distance to the estimated plane d_V.

Figure 18. For the images shown in Figure 15, using four methods to calculate the matching costs of the matching images, the stereo matching was done with SGM. The window size is changed from 7x7 to 15x15 pixels.

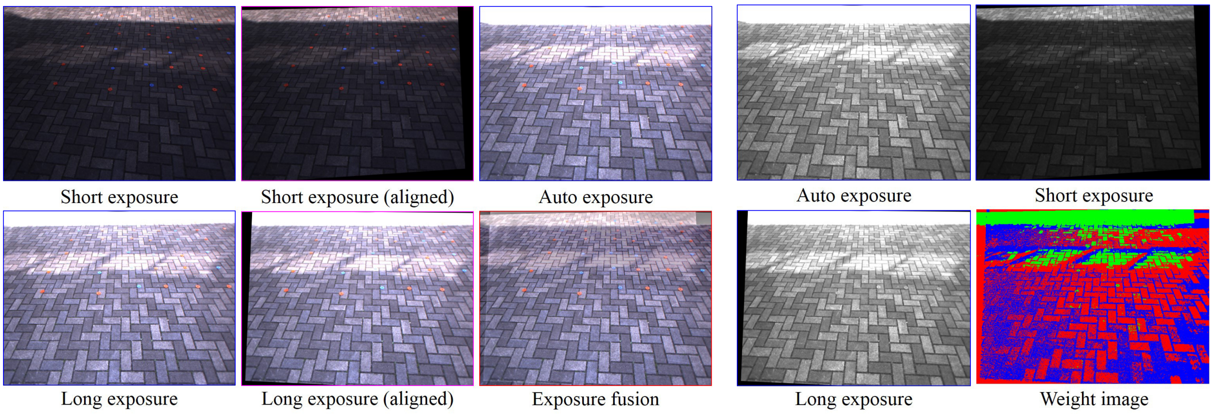
becomes smaller. Especially when the window size is small, the methods “Multiple images” and “Multiple images by weight” performed much better. It is noticed that in some window sizes it appears that the average distance to the estimated plane d_V calculated with the method “Auto exposure” is smallest and the method “Auto exposure” performs “best”. However, compared with other methods, its N_V is smallest. Besides, since the disparity values for the pixels in the top of the image were not calculated with the method “Auto exposure” as shown in Figure 16(a), the top part of the image was not used to calculated its d_V. For these reasons, the method “Auto exposure” actually performs worst.

Next, stereo matching was done with SGM. For example, with a window of 11x11 pixels, the disparity images calculated with four different methods are shown in Figure 16(b). As it is overexposure in the top of the image grabbed with auto exposure, even when SGM was used, the disparity values for the pixels in this part still were not calculated with the method “Auto exposure”. With the window size changed from 7x7 to 15x15 pixels, Figure 18 presents the valid point number N_V and the average distance to the estimated plane d_V. It shows that the valid point numbers of the methods “Exposure fusion”, “Multiple images” and “Multiple images by weight” are quite close and bigger than the result of the method “Auto exposure”. However, the methods “Multiple images” and “Multiple images by weight” perform better than the method “Exposure fusion” with a smaller average distances to the estimated plane.

3.2.4 Outdoor experiments with the stereo vision camera held in the hand

The experiments were done outdoors with the stereo vision camera held in the hand and the grabbed images of the right camera are shown in Figure 19(a). The images grabbed with short and long exposures were aligned to the image captured with auto exposure. The aligned images and resulting fused image are shown in Figure 19(a). The grayscale images are shown in Figure 19(b). As an example, the intensity diversity weight was calculated with a window of 15x15 pixels and the weight image is shown in Figure 19(b).

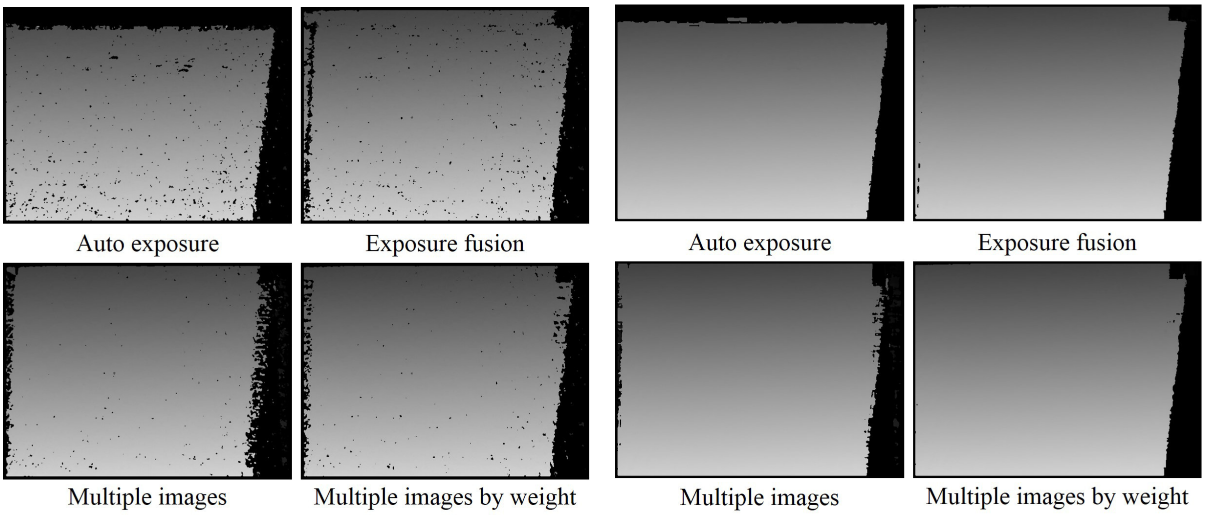
First, stereo matching was done with local window-based method. For example, with a win-



(a) The images were grabbed with multiple exposures. The images grabbed with short and long exposures were aligned to the image captured with auto exposure. With exposure fusion, the image grabbed with auto exposure and the registered images of the photographs captured with short and long exposures were fused.

(b) With the intensity diversity weight was calculated with a window of 15x15 pixels, the weight image was calculated.

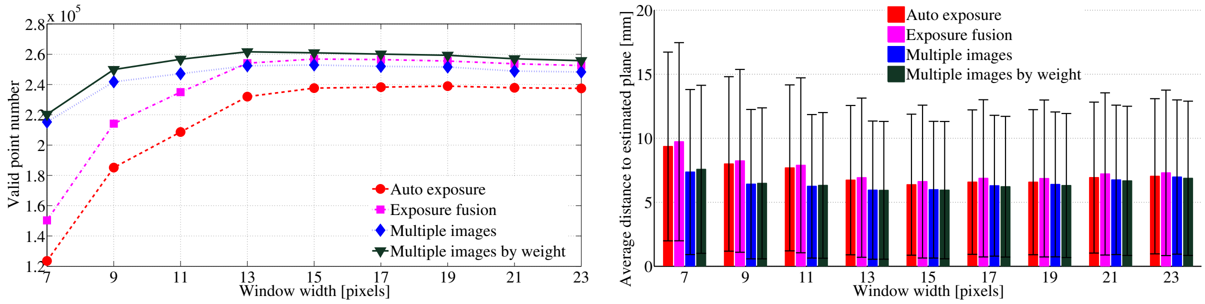
Figure 19. The images were acquired outdoors with the stereo vision camera held in the hand.



(a) With the window size of 15x15 pixels, the disparity images were calculated with local window-based method.

(b) With the window size of 11x11 pixels, the disparity images were calculated with SGM.

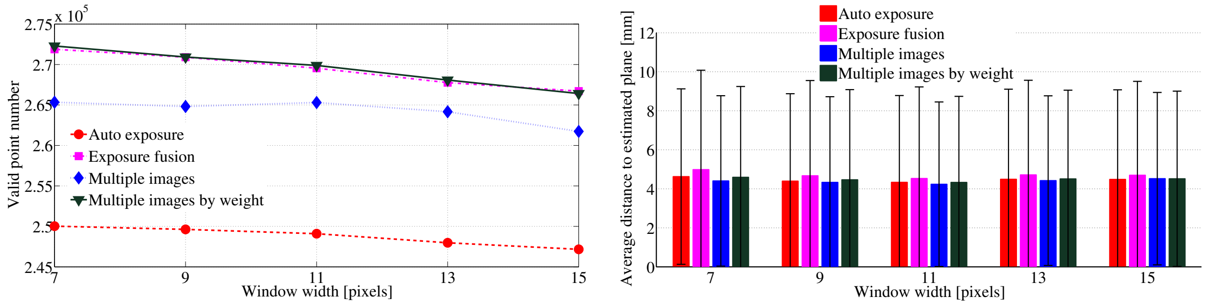
Figure 20. For the images shown in Figure 19, using four methods to calculate the matching costs of the matching images, the disparity images were calculated.



(a) Valid point number N_V.

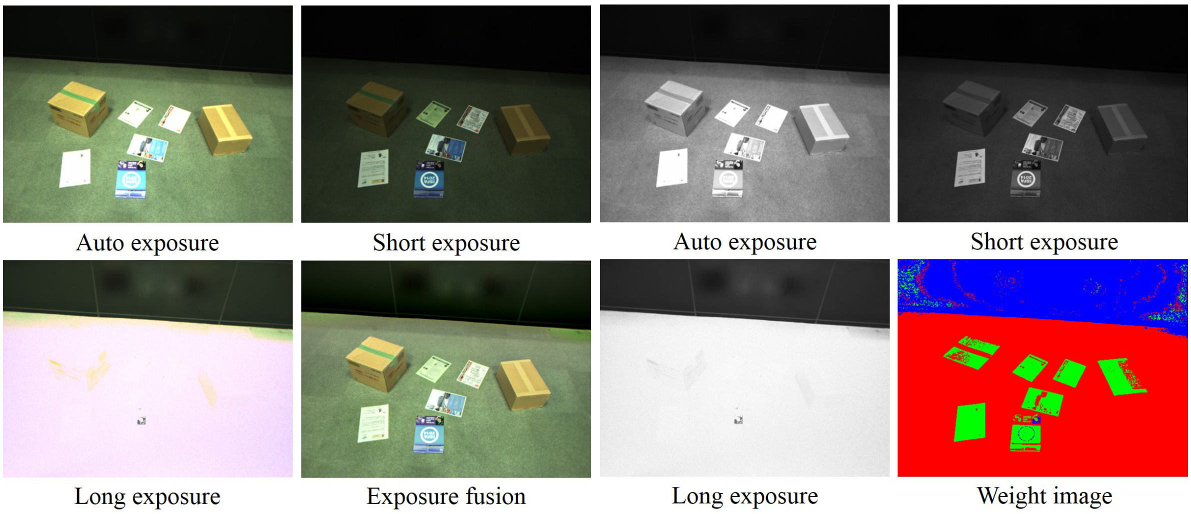
(b) Average distance to the estimated plane d_V.

Figure 21. For the images shown in Figure 19, using four methods to calculate the matching costs of the matching images, the stereo matching was computed with local window-based method. The window size is changed from 7x7 to 23x23 pixels.



(a) Valid point number N_V. (b) Average distance to the estimated plane d_V.

Figure 22. For the images shown in Figure 19, using four methods to calculate the matching costs of the matching images, the stereo matching was done with SGM. The window size is changed from 7x7 to 15x15 pixels.



(a) The images were grabbed with multiple exposures. With exposure fusion, these images were fused.

(b) With the intensity diversity weight was calculated with a window of 15x15 pixels, the weight image was calculated.

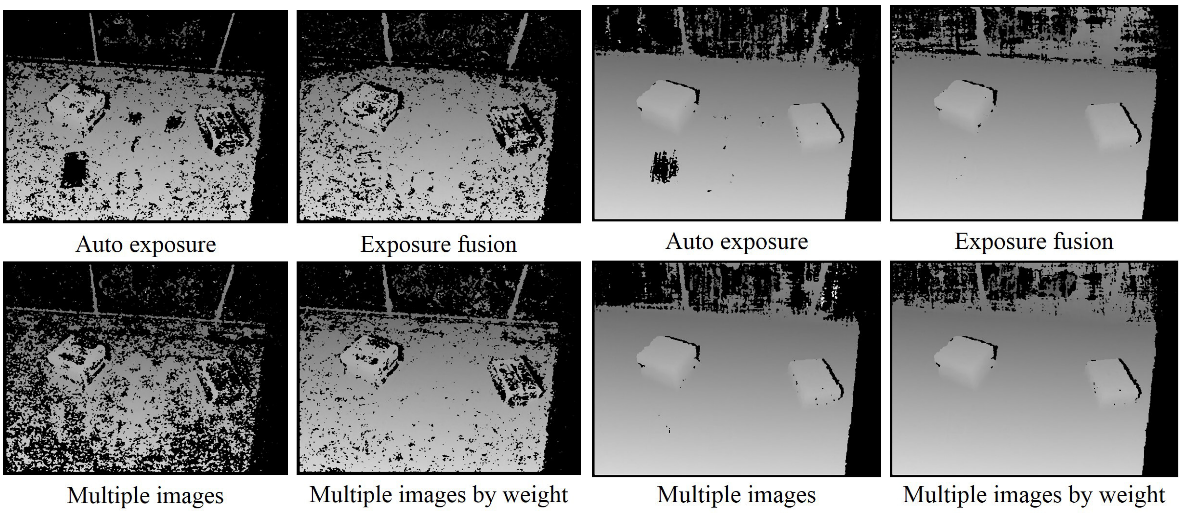
Figure 23. The images were acquired for environment perception.

dow of 15x15 pixels, the disparity images calculated with four different methods are shown in Figure 20(a). With the window size changed from 7x7 to 23x23 pixels, Figure 21 shows the valid point number N_V and the average distance to the estimated plane d_V. It illustrates that with the proposed method “Multiple images by weight”, more valid points can be obtained and the average point to estimated plane distance becomes smaller compared with the methods “Auto exposure” and “Exposure fusion”. Especially when the window size is small, the proposed method “Multiple images by weight” performed much better.

Next, stereo matching was done with SGM. For example, with a window of 11x11 pixels, the disparity images computed with four different methods are shown in Figure 20(b). With the window size changed from 7x7 to 15x15 pixels, Figure 22 presents the valid point number N_V and the average distance to the estimated plane d_V. It shows that the N_V of the methods “Exposure fusion” and “Multiple images by weight” are quite close and bigger than the results of the methods “Auto exposure” and “Multiple images”. However, the proposed method “Multiple images by weight” performs better than the method “Exposure fusion” with the smaller average distance to the estimated plane d_V.

3.3 *Experiment of environment perception*

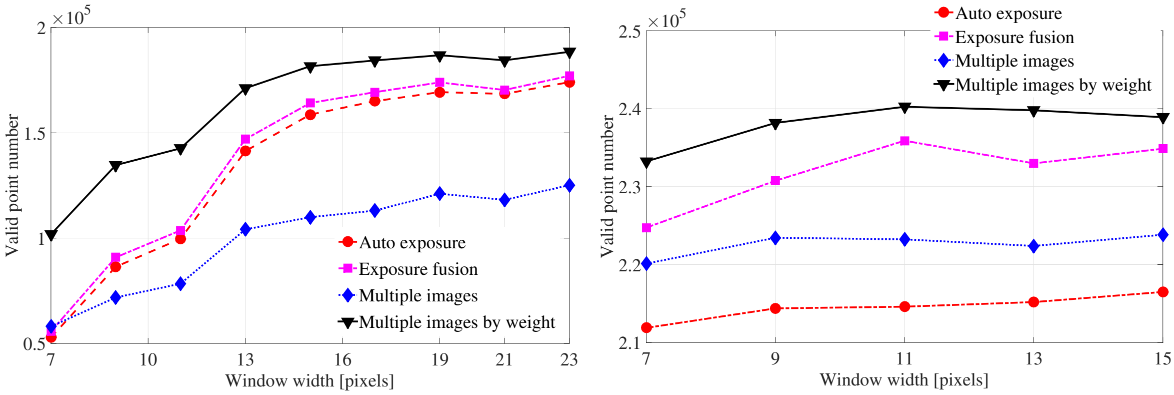
The experiments were done for environment perception. The grabbed images of the right camera and the resulting fused image are shown in Figure 23(a). The grayscale images are shown in



(a) With the window size of 15x15 pixels, the disparity images were calculated with local window-based method.

(b) With the window size of 11x11 pixels, the disparity images were calculated with SGM.

Figure 24. For the images shown in Figure 23, using four methods to calculate the matching costs of the matching images, the disparity images were calculated.



(a) The stereo matching was computed with local window-based method and the window size is changed from 7x7 to 23x23 pixels.

(b) The stereo matching was done with SGM and the window size is changed from 7x7 to 15x15 pixels.

Figure 25. For the images shown in Figure 23, using four methods to calculate the matching costs of the matching images, the valid point number N_V.

Figure 23(b). As an example, the intensity diversity weight was calculated with a window of 15x15 pixels and the resulting weight image is shown in Figure 23(b).

First, stereo matching was done with local window-based method. For example, with a window of 15x15 pixels, the disparity images calculated with four different methods are shown in Figure 24(a). With the window size changed from 7x7 to 23x23 pixels, Figure 25(a) shows the valid point number. It illustrates that with proposed method “Multiple images by weight”, more valid points can be obtained, especially when the window size is small.

Next, stereo matching was done with SGM. For example, with a window of 11x11 pixels, the disparity images computed with four different methods are shown in Figure 24(b). With the window size changed from 7x7 to 15x15 pixels, the valid point number is shown in Figure 25(b). From Figure 25(b), it shows that the proposed method “Multiple images by weight” performs better than other methods with more valid points.

3.4 *Discussion*

Compared with the method “Exposure fusion”, for each pixel of the matching image, the local information in its local window acquired from the images grabbed with auto, short and long exposures is better retained when the matching cost is aggregated with the methods “Multiple images” and “Multiple images by weight”. With the proposed method “Multiple images by weight”, the matching cost value obtained from the pixel which is well exposed and has significantly different intensity values in its local window becomes dominant and the useful information in the local window can be well retained. For this reason, the experimental results show that compared with the methods “Auto exposure” and “Exposure fusion”, the proposed method consistently allowed more valid points to be obtained and the 3D terrain model can be built more accurately. Since the matching cost is not smoothed when it is computed with the method “Multiple images”, the method “Multiple images” performed worse than the proposed method with less valid points.

4. Conclusion

In order to apply stereo vision techniques in field robotics to acquire 3D terrain maps in extreme light conditions, a series of photographs are taken with multiple exposures. Since it is possible that the camera is moved when the images are grabbed with multiple exposures, the images acquired with short and long exposures are aligned to the image grabbed with auto exposure. A stereo matching algorithm, the matching costs of the images grabbed with multiple exposures are directly summed by weight, is proposed in this paper. Compared with the traditional methods such as “Exposure fusion”, with the proposed method, it is not needed to fuse the images grabbed with multiple exposures, and for each pixel of the matching image, the local information in its local window acquired from the images grabbed with multiple exposures can be better retained. Experiments were done in laboratory and outdoors with a stereo vision camera fixed on a tripod and held in the hand, and the stereo matching were done with a local window-based method and SGM. The experiments were also done for environment perception. Through the experiments, it was verified that compared with the methods “Auto exposure” and “Exposure fusion”, the proposed method consistently allowed more valid points to be obtained and the 3D terrain model can be built more accurately. Especially when the local window-based method was used, compared with other methods, the proposed method performed much better. Field experiments are planned to be conducted with the Gryphon system in Angola in the near future to further evaluate the proposed method. The proposed methods can be used in other applications.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 25303012.

References

[1] Alex Masuo Kaneko, Marco Marino, Edwardo F. Fukushima, “Humanitarian Demining Robot Gryphon: New Vision Techniques and Optimization Methods”, IEEE/RSJ Int. Conf. on Intelligent Robotics and Systems (IROS), Taipei, Taiwan, pp.228-233, October, 2010.

[2] Mark A. Robertson, Sean Borman and Robert L. Stevenson, “Dynamic range improvement through multiple exposures”, Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on, Kobe, Japan, pp.159-163, October, 1999.

[3] Paul E. Debevec and Jitendra Malik, “Recovering high dynamic range radiance maps from photographs”. Proceedings of the 24th annual conference on Computer graphics and interactive techniques, Los Angeles, CA, USA, pp.369-378, August, 1997.

[4] R. Correal, G. Pajares and J.J. Ruz, "Automatic expert system for 3D terrain reconstruction based on stereo vision and histogram matching", Expert Systems with Applications, Vol. 41, Issue 4, pp.2043-2051, 2014.

[5] Mertens Tom, Kautz Jan and Van Reeth Frank, "Exposure Fusion", Computer Graphics and Applications, 2007. 15th Pacific Conference on, Hawaii, USA, pp.382-390, October, 2007.

[6] Jianhua Li, Alex M. Kaneko and Edwardo F. Fukushima, "Proposal of Terrain Mapping under Extreme Light Conditions Using Direct Stereo Matching Methods", IEEE/SICE International Symposium on System Integration, Tokyo, Japan, pp.153-158, December, 2014.

[7] Barbara Zitova and Jan Flusser, "Image registration methods: a survey". Image and Vision Computing, Vol. 21, Issue 11, pp.977-1000, October, 2003.

[8] Richard Szeliski, "Image alignment and stitching: a tutorial", Foundations and Trends in Computer Graphics and Vision archive, Vol. 2, Issue 1, pp.1-104, January, 2006.

[9] Anna Tomaszewska and Radoslaw Mantiuk, "Image Registration for Multi-exposure High Dynamic Range Image Acquisition", In: Proc. of International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, Plzen-Bory, Czech Republic, pp.49-56, January, 2007.

[10] Daniel Scharstein and Richard Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms". International Journal of Computer Vision, Vol. 47, No. 1, pp. 7-42, May, 2002.

[11] Gelfand Natasha, Adams Andrew, Park Sung Hee and Pulli Kari, "Multi-exposure Imaging on Mobile Devices", MM '10 Proceedings of the International Conference on Multimedia, Firenze, Italy, pp.823-826, October, 2010.

[12] M. Brown and D. G. Lowe, "Recognising panoramas," Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, Nice, France, Vol. 2, pp. 1218-1225, October, 2003.

[13] M. Brown and D. G. Lowe, "Automatic Panoramic Image Stitching using Invariant Features", International Journal of Computer Vision, Vol. 74, No. 1, pp. 59-73, 2007.

[14] Herbert Bay, Andreas Ess, Tinne Tuytelaars and Luc Van Gool, "Speeded-Up Robust Features (SURF)", Computer Vision and Image Understanding, Vol. 110, No. 3, pp. 346-359, 2008.

[15] M.A. Fischler and R.C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography", Communications of the ACM, 24(6) pp. 381-395, 1981.

[16] Ramin Zabih and John Woodfill, "Non-parametric Local Transforms for Computing Visual Correspondence", In Proceedings of European Conference on Computer Vision, Stockholm, Sweden, pp.151-158, May, 1994.

[17] Hirschmuller Heiko, "Stereo Processing by Semiglobal Matching and Mutual Information", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, No. 2, pp.328-341, February, 2008.

[18] Istvan Haller and Sergiu Nedevschi, "Design of interpolation functions for subpixel-accuracy stereo-vision systems", IEEE Trans. on Image Process., Vol. 21, No. 2, pp.889-898, February, 2012.