

Article

A Real-Time Infrared Stereo Matching Algorithm for RGB-D Cameras' Indoor 3D Perception

Jiageng Zhong ¹, Ming Li ^{1,2,*} , Xuan Liao ^{2,3} and Jiangying Qin ¹

¹ School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; zhongjiageng@whu.edu.cn (J.Z.); jy_qin@whu.edu.cn (J.Q.)

² State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; liaoxuan@whu.edu.cn

³ Department of Land Surveying and Geo-Informatics, the Hong Kong Polytechnic University, Hong Kong

* Correspondence: lisouming@whu.edu.cn; Tel.: +86-138-7102-5925

Received: 12 June 2020; Accepted: 27 July 2020; Published: 28 July 2020



Abstract: Low-cost, commercial RGB-D cameras have become one of the main sensors for indoor scene 3D perception and robot navigation and localization. In these studies, the Intel RealSense R200 sensor (R200) is popular among many researchers, but its integrated commercial stereo matching algorithm has a small detection range, short measurement distance and low depth map resolution, which severely restrict its usage scenarios and service life. For these problems, on the basis of the existing research, a novel infrared stereo matching algorithm that combines the idea of the semi-global method and sliding window is proposed in this paper. First, the R200 is calibrated. Then, through Gaussian filtering, the mutual information and correlation between the left and right stereo infrared images are enhanced. According to mutual information, the dynamic threshold selection in matching is realized, so the adaptability to different scenes is improved. Meanwhile, the robustness of the algorithm is improved by the Sobel operators in the cost calculation of the energy function. In addition, the accuracy and quality of disparity values are improved through a uniqueness test and sub-pixel interpolation. Finally, the BundleFusion algorithm is used to reconstruct indoor 3D surface models in different scenarios, which proved the effectiveness and superiority of the stereo matching algorithm proposed in this paper.

Keywords: infrared image; stereo matching; RGB-D camera; depth map; 3D perception

1. Introduction

Indoor 3D environment perception technology is one of the key technologies for robot positioning and navigation, virtual reality, augmented reality and indoor mapping and localization [1–7]. With the rapid development of sensor technology, there are many devices that can be used for the point cloud acquisition and surface modeling of indoor scenes, such as LiDAR [8], RGB cameras [9], RGB-D cameras [5] and other commercial sensors, which are widely used in indoor 3D perception. The RGB-D camera combines the characteristics of two types of sensors, LiDAR and RGB cameras, to obtain point cloud data and RGB image data output in a time series, which is more conducive to real-time acquisition and the update of indoor 3D spatial structure and texture information. Moreover, it is inexpensive compared with devices integrating LiDAR, and covers extensive research and application prospects in close-range indoor 3D perception. One of the earliest consumer commercial RGB-D sensors is Apple's Prime Sense sensor, which uses structured light (SL) to implement scene perception technology. Similar devices include the Microsoft Kinect v1 and the Asus Xtion [10]. Microsoft then released the Kinect v2, a version of the RGB-D camera that uses the time-of-flight (ToF) principle for distance sensing, with a high frame rate, but a lower depth map resolution [11,12]. The emergence of RGB-D

cameras that include an infrared texture projector with a fixed pattern means RGB-D-style cameras have a higher depth map resolution, especially in close-range indoor 3D perception, where they can obtain complete and more accurate data. Intel's portable consumer-grade RGB-D cameras are the main sensors, including the Intel R200 (2015), D415 and D435 (2018), which are based on active stereo vision (ASV) for data acquisition and processing. Specifically, the technology typically uses one NIR texture projector paired with two NIR cameras for depth estimation [13]. With the advent of low-cost, portable RGB-D sensors, RGB-D cameras based on ASV are not sensitive to indoor textures, so there are a growing number of commercial companies and researchers interested in using such RGB-D cameras for 3D perception of indoor scenes [14]. Among them, the R200 is a representative RGB-D camera based on infrared speckle and stereo vision technology for the depth estimation of indoor scenes. Many researchers use it for robot indoor navigation and positioning, indoor 3D Mapping and other research [15]. The binocular stereo matching module provided by Intel on the R200 is based on a local matching method. Although it can match infrared stereo images with a higher frame rate, its depth map has the problem of many holes and a short valid detection distance. Specifically, the hole rate will often reach 40%, and the valid detection distance is less than 4 m. This is a limitation in much indoor 3D reconstruction and mapping work that requires dense 3D perception, which makes it unable to work well for many usage scenarios and requirements.

In view of the current performance and deficiency of the R200 commercial matching algorithm, a novel stereo matching algorithm, called Infrared Stereo Semi-Global Matching Algorithm (ISGSM), is proposed on the basis of the work of Semi-Global Matching (SGM) [16]. This method is based on the characteristics of an infrared speckle image. It adopts the strategy of a semi-global and sliding window, by which it can incorporate more data into the cost calculation. In this way, higher-quality stereo matching will be achieved, so that the obtained depth map has better integrity and higher accuracy, and significantly increases the detection range and accuracy of the R200. The validity and superiority of the method are verified by the experimental comparison and analysis of the R200's commercial algorithm [17] and the representative stereo matching algorithms [18]. The following sections of this paper are arranged as follows: Section 2 outlines the research progress of existing stereo vision technology; Section 3 introduces the existing typical algorithms, and describes the newly proposed ISGSM algorithm in detail; Section 4 explains the experimental methods and analyzes the experimental results; Section 5 is the discussion based on the experimental results; and Section 6 is the conclusions.

2. Related Work

At present, research on depth maps through the stereo vision technique is a research hotspot in the field of photogrammetric computer vision. It calculates the depth map of the scanned scene through matching of the images, and then can acquire dense unstructured point cloud data, which is the core technology for 3D scene perception and semantic segmentation. According to the characteristics and principles of these stereo matching algorithms, they can be simply divided into local methods and global methods [16]. The local methods mainly use the local information around the pixel of interest to calculate, involving less information and lower computational complexity. Common local matching algorithms include area-based and feature-based methods. The area-based matching algorithm is based on the principle of photometric invariance. The gray level of the neighborhood window is often used as the matching unit, and the correlation degree is used as the distinguish basis. In this way, a denser disparity image will be obtained. Among them, the more common one is the BM (Block Matching) algorithm [18]. Its prominent shortcoming is that the sharpness of the change of the correlation function is often insufficient in the texture-less area, and it is difficult to retain depth continuity, so it is unlikely to obtain accurate matching results. In view of these problems, Zabih et al. [19] have carried out some research improvements. They promoted rank transform to Census transform, so that they could avoid the correlation phase altogether and simply match pixels according to a set of semi-independent measures. The feature-based matching algorithm is based on

the principle of geometric invariance, which can overcome the shortcoming of the area-based matching algorithm's sensitivity to texture-less areas to a certain extent. Due to the statistical properties of the feature units and the regularity of the data structure, it is suitable for hardware design. However, there are problems where dense disparity images need to be attached to a more complex interpolation, and the performance of feature matching results relies heavily on the precision of the feature extraction. Prince [20] uses the local energy method to identify multi-directional subpixel features, and detects multiple types of features for matching, which improves the capability of feature-based local matching algorithms. However, most local matching algorithms are sensitive to noise, and the matching effect is not ideal in texture-less areas, occlusion areas or areas of discontinuous disparity. The global matching algorithms transform the matching problem of corresponding points into a global optimization problem of finding an energy function, the core of which lies in the energy function construction method and energy function optimization solution strategy. There are common global algorithms, such as Dynamic Programming [21], Graph Cut [22] and Belief Propagation [23]. Scharstein et al. [24] evaluated the performance of various optimization strategies and pointed out that Dynamic Programming can quickly search the optimal solution while satisfying the corresponding point sequence constraints. Its essence is finding the least matching cost paths between left and right images, providing global support for locally texture-less areas and, thereby, improving the matching accuracy, but it cannot effectively confluence the continuity constraints in horizontal and vertical directions. The matching accuracy of global algorithms is higher than that of the local algorithms, and the edges of the objects are also kept better. Unfortunately, the complexity of global algorithms is higher, and the processing time and hardware costs increase, which consume more memory during runtime. Therefore, it is difficult to achieve real-time processing and its application scenarios are relatively limited.

In view of the advantages and disadvantages of global methods and local methods in the stereo vision technique, the semi-global algorithm has attracted the interest of researchers and the attention of the industry. One of the most representative is the SGM algorithm [16], which combines the advantages of local methods and global methods, performs 2D global optimization by constraining the 1D path in multiple directions, and maintains higher efficiency while obtaining higher quality disparity images [16,25,26]. Meanwhile, the semi-global matching algorithm is less complex than global methods and can be processed in real time. In addition, the accuracy and detection distance of the semi-global matching algorithm, as well as the quality of disparity images, are significantly higher than local matching algorithms, which provides stunning visual effects and the ability for fine 3D perception. Therefore, the research on semi-global matching has become the research focus of the current stereo vision technique, especially in many indoor application scenarios that require both complete 3D perception accuracy and real-time processing, but it still has some shortcomings. Many researchers have made a lot of improvements based on SGM, among which the more prominent one is the tSGM algorithm [27]. The tSGM algorithm in SURE [28] provides a hierarchical coarse-to-fine solution for the SGM method to limit disparity search ranges and decreases the memory demand as well as the processing time. However, edges are not reconstructed as clearly as in the SGM algorithm [29], which will directly reduce the accuracy and integrity of 3D perception. Considering the use of stereo vision in structured environments, the CSGM (Consistent SGM) method [30] can handle structures well but increases the execution time by about 30%–50%. Based on the smallest spanning tree, the MST-SGM algorithm [31] is proposed, which has fewer matching black edges than the SGM method, but at the same time leads to more errors, which will reduce the accuracy of depth information. Combined with adaptive Census transformation, an improved SGM algorithm is proposed [32], which enables a color-aware filter to deal with light changes in outdoor scenes but does not keep the edges well. Plane fitting is performed on the basis of disparity images obtained by SGM [33], and it has achieved good results, but the computation cost also increases, and the real-time performance is not good enough. SGM-Nets [34] combines the SGM algorithm with a neural network, which can significantly improve the performance under the premise of sufficient prior knowledge. In addition, there is the SGBM algorithm [35] which can improve the accuracy of elevation estimation in the water area of

the optical satellite imagery through adaptive block matching, but it is only applicable to the poor texture area with an almost constant height, and its application scenarios are quite different from indoor scenarios. To sum up, researchers have proposed many improved algorithms based on the SGM algorithm in their respective research fields. However, there is still no perfect ready-made solution for such problems as the complete and accurate real-time indoor perception of a medical nursing robot, real-time precise 3D reconstruction and more accurate and subtle augmented reality experience.

3. Methodology

Figure 1 shows a complete solution of a real-time indoor 3D environment perception based on the RGB-D camera studied in this paper. First, calibrate the R200's cameras, and acquire RGB and infrared images. Then, use our stereo matching algorithm to acquire depth maps. Finally, the RGB images and point cloud obtained through depth maps are used to reconstruct the indoor 3D surface model of experimental scenes using the BundleFusion algorithm [36] (open source). The red box in Figure 1 is the main innovation and research content of this paper, which will be introduced and explained in detail later.

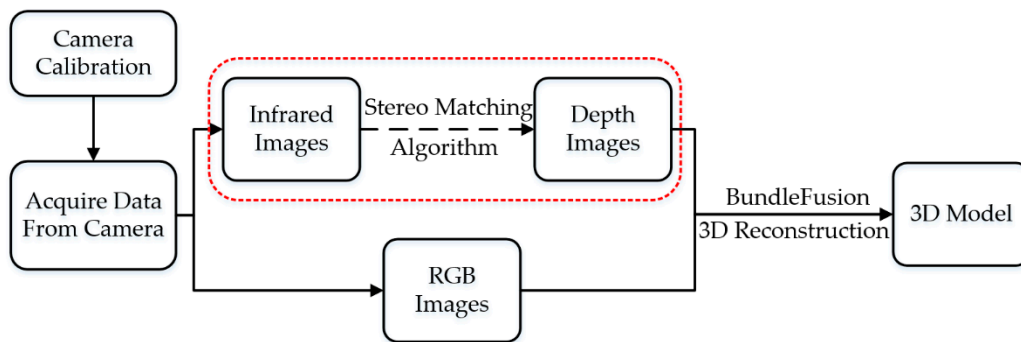


Figure 1. A complete solution for real-time 3D perception of indoor scenes.

In the first place, we used applications integrated in Matlab R2019a to calibrate the R200 [37–39]. After the calibration, we needed to cooperate with the R200 through a portable notebook computer to collect experimental data. In this paper, the software and hardware environment of the experimental laptop with R200 include Ubuntu16.04 LTS, Intel (R) Core (TM) i7 CPU, 8.00 GB RAM, NVIDIA GeForce MX150 GPU and the camera driver from librealsense-1.12.1. In the experiments, we could acquire 60 fps RGB images (640×480), infrared images (640×480) and depth maps (640×480) processed by the integrated module of the R200.

3.1. Stereo Matching Algorithm

3.1.1. Stereo Matching Algorithm of the R200

R200 uses a Census cost function to compare left and right images. Thorough comparisons of photometric correlation methods showed the Census descriptor to be among the most robust in handling noisy environments [17]. The main mathematical models of the algorithm are shown in Formulas (1) and (2). First, with a pixel $p(i, j)$ in the match image R as the center, select the Census transformation window with a size 7×7 . Then, compare the gray value of the center point and the pixel in the window successively. If it is larger, it is set to 1, and if it is smaller, it is set to 0. Finally, a 0/1-bit string can be obtained [40].

$$C(p, d) = \otimes_{p \in W} \Phi(R_p, R_{p'}) \quad (1)$$

$$\Phi(R_p, R_{p'}) = \begin{cases} 1, & R_p < R_{p'} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where W is the Census transformation window corresponding to the central pixel p , p' is the pixel in the window centered on p . R_p and $R_{p'}$ are the gray values of p , p' . Then, the bit string for the Census transformation of the window at point p can be obtained. Similarly, the bit string for the search point of the target image T is obtained. Finally, measured by Hamming distance, the level of similarity of the two-bit strings is quantified [41]. Then, a 64-disparity search is performed, and costs are aggregated with a 7×7 box filter. The best-fit candidate is selected, a subpixel refinement step is performed, and a set of filters are applied to filter out bad matches [17].

3.1.2. Block Matching Algorithm

The Block Matching Algorithm (BM) is a typical representative local stereo matching algorithm, which incorporates the idea of “block” [24]. BM has been proposed for a long time, and there are a variety of derived algorithms. There is a detailed introduction and comparison in [18]. In BM, the base image is divided into many small blocks, and each block is compared with the block collected from the matched image. It is achieved by moving and comparing the block. The process of moving is to simulate the movement of a small block from one position to another by creating a vector, and then looking horizontally for the most appropriate pixel block in another image, and finally calculating the disparity based on this. As for the matching method between blocks, SAD (sum of absolute differences) is used as the similarity measurement function in the contrast experiment of this paper [42]. The mathematical model can be expressed by Formula (3):

$$C = \sum_{i,j \in W} |I_L(x+i, y+i) - I_R(x+i+d, y+i)| \quad (3)$$

where d is the disparity value at this pixel and W is the support window. The best disparity at pixel (x_0, y_0) is the parameter d which minimizes the cost C . The principle of BM is simple, and its complexity is quite low, so it has good real-time performance. However, its depth value accuracy is poor, and there are many holes in the depth map.

3.1.3. Semi-Global Matching Algorithm

The SGM algorithm is one of the most representative semi-global matching algorithms, which is between local and global. It has three key steps: cost calculation, cost aggregation and disparity computation [16]. SGM has its variant. In this paper, SGM with BT [16] is selected as the comparative method.

Cost calculation. There are many methods for cost calculation, and SGM with BT [16] chooses the sampling insensitive measure of Birchfield and Tomasi [43] (hereinafter referred to as the BT algorithm), which is a pixelwise matching cost calculation method based on sampling. The cost of a match sequence is defined by a constant penalty for each occlusion, a constant reward for each match, and a sum of the dissimilarities between the matched pixels.

Cost aggregation. Pixelwise cost calculation is generally ambiguous and wrong matches can easily have a lower cost than correct ones, due to noise, etc. [16]. So, an additional constraint is added to support smoothness by penalizing changes of neighboring disparities. Then, the energy $E(D)$ that depends on the disparity image D is defined for this. $E(D)$ includes the pixelwise cost and the smoothness constraints, and its specific definition of the energy function $E(D)$ is shown in Formula (4):

$$E(D) = \sum_p C(p, D_p) + \sum_{q \in N_p} P_1 T[|D_p - D_q| = 1] + \sum_{q \in N_p} P_2 T[|D_p - D_q| > 1] \quad (4)$$

The first term is the sum of all pixel matching costs. The second term adds a constant penalty P_1 for all the pixels q in the neighborhood N_p of p , for which the disparity changes no more than one pixel. The third term adds a larger constant penalty P_2 for penalizing larger disparity changes. After constructing the energy function, the problem of matching is transformed into finding the disparity

image D that minimizes the energy function $E(D)$. Usually, the solution for this kind of problem depends on the dynamic programming method, which can efficiently perform the optimization problem. However, since the dynamic programming solution has difficulty relating the 1D optimizations of individual image rows to each other in the 2D image, it will easily suffer from streaking [16]. Therefore, a better idea is to consider accumulating 1D matching costs from multiple directions, not just one line. Summing the costs of all directions, the aggregated cost $S(p, d)$ can be better calculated.

Disparity computation. As in the local matching method, the disparity image D_b of the base image I_b is determined by selecting the disparity d that minimizes cost for each pixel p , i.e., $\min_d S(p, d)$.

Finally, the matching error should be eliminated. For a pixel p on the base image I_b , since disparity has been calculated, its corresponding pixel q on the match image I_m can be calculated. If the difference in disparity between two pixels is greater than 1, the disparity at p is regarded as an invalid value. This step can reduce the number of mismatches.

3.1.4. Our Infrared Semi-Global Stereo Matching Algorithm

There are noises in infrared images obtained by an infrared camera, and the intensity of infrared reflected light will be affected by factors such as the angle of incidence and distance. As mentioned above, in order to achieve the research purpose of this paper, the existing methods do not solve all these problems. Therefore, in order to achieve better matching of infrared images, we propose a new infrared stereo image matching algorithm—the ISGSM algorithm for higher quality depth maps. On the basis of SGM algorithm and other algorithms, it constructs a 2D global energy function for global optimization and is improved for cost calculations and disparity calculations. The detailed algorithm flow is shown in Figure 2. Subsequent experiments show that the ISGSM algorithm can perform infrared stereo matching better than existing algorithms. In this paper, it is verified with the R200 in indoor real-time 3D perception.

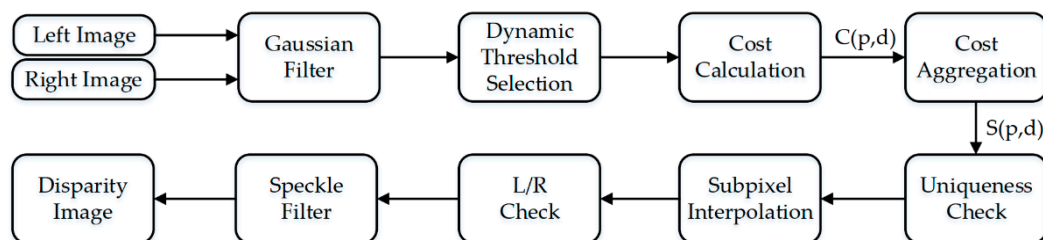


Figure 2. A flow chart of the ISGSM algorithm.

As the infrared projector of R200 emits infrared speckle with low power and the IR images usually lack texture [44], the reflected infrared ray in many places in the scene is very weak, or even not present, which directly leads to the lack of texture in many areas of the infrared image, not conducive to matching. As the matching cost directly depends on the similarity between the two primitives [35], Gaussian filtering operation is more conducive to reduce the matching cost. Therefore, after acquiring two infrared stereo images, we first perform Gaussian filtering on them (the default is a 3×3 window). On the one hand, Gaussian filtering reduces noises of the infrared images, and on the other hand, it can improve the correlation of the stereo infrared image. In this way, the region with a weak original signal can be strengthened and the influence of the abnormal value can be weakened. It is shown in this paper experiments that after Gaussian filtering, the correlation coefficient between the two images can be increased by about 9%, and the mutual information can be increased by about 13%.

In SGM with BT [16] (intensity-based matching), as the BT algorithm is pixelwise matching, which is easy to be interfered with by noise and has weak robustness, we integrate the idea of block into the SGM algorithm to integrate the information in an image block for robustness [17]. The idea is based on the BM algorithm, but the window size of “block” in the BM algorithm is a preset constant, which is not suitable for all kinds of scenes. Moreover, when calculating the cost of the

BT algorithm, the mutual information between left and right images is not fully utilized. Therefore, in this paper, before calculating the cost, the dynamic threshold selection of the window size based on mutual information is carried out, so that the algorithm can use different parameters in different scenarios, thereby enhancing its adaptability. The implementation: after Gaussian filtering, the mutual information between the two images is calculated, and then the size of the sliding window is selected according to the value of the mutual information. Formulas (5)–(7) are the mathematical expression of mutual information, and Formula (8) is the mathematical model of sliding window selection.

$$MI_{I_L, I_R} = H_{I_L} + H_{I_R} - H_{I_L, I_R} \quad (5)$$

$$H_I = - \int_0^1 P_I(i) \log P_I(i) di \quad (6)$$

$$H_{I_1, I_2} = - \int_0^1 \int_0^1 P_{I_L, I_R}(i_1, i_2) \log P_{I_L, I_R}(i_1, i_2) di_1 di_2 \quad (7)$$

$$L = L(H_{I_L, I_R}) \quad (8)$$

where MI_{I_1, I_2} is the mutual information of two images, H_I is the entropy of image I , H_{I_L, I_R} is the mutual entropy of I_L and I_R , L is the size of algorithm window and $L(H_{I_L, I_R})$ is a segmented function for dynamic threshold selection.

In calculating the cost, the BT algorithm is used. Different from the original BT algorithm, the costs of the BT algorithm used in this paper includes two parts: one is the costs calculated from the gray value of the left and right images, the other is the costs calculated from the result of the left and right images through the horizontal Sobel operator (SobelX). The above two parts of the costs are merged to obtain the final costs. In this way, the similarity can be improved. It should be noted that the horizontal gradient calculated here is not used directly but is processed in each segment. Each pixel on the image processed by the SobelX operator is mapped into a new pixel with a function. Here, P is the pixel value after filtering with the SobelX operator, and P_{new} is the new pixel value. Then their mapping function can be expressed by formula (9):

$$P_{new} = \begin{cases} 0, & P < -FParam \\ P + FParam, & -FParam \leq P \leq FParam \\ 2 * FParam, & P \geq FParam \end{cases} \quad (9)$$

where $FParam$ is a constant parameter used as a threshold for the subsection process. It can control the result within a certain range and optimize the performance of the algorithm.

When performing cost calculation, this paper adopts the idea of “block” and incorporates the information of neighborhood pixels into the calculation, which can make the result smoother. In cost aggregation, we draw on the idea of SGM, approximating a global, 2D smoothness constraint by combining many 1D constraints [16]. As transforming the problem of stereo matching into searching the optimal solution of the energy function, the final result can be comparable to that of global matching algorithms, while maintaining high efficiency.

After obtaining the preliminary disparity image, there are still some problems that need to be optimized. The optimization in this paper mainly includes the following steps:

- (1) Uniqueness test. The minimum computed cost function value should be smaller than the second-best value to a certain extent. Otherwise, the match will be considered invalid.
- (2) Sub-pixel interpolation. Since the image samples the real world, the disparity image cannot be exactly equal to the disparity of its corresponding object point. As there is a certain deviation, it is difficult to meet the needs of high-precision 3D perception and 3D reconstruction. Therefore, sub-pixel interpolation is needed to improve accuracy. The interpolation formulas are shown in

Formulas (10) and (11). Its essence is a parabolic interpolation: the disparity is the minimum value of the parabola.

$$\text{denom2} = \max(\text{Sp}(d-1) + \text{Sp}(d+1) - 2*\text{Sp}[d], 1) \quad (10)$$

$$d = d + \frac{\text{Sp}(d-1) - \text{Sp}(d+1) + \text{denom2}}{2*\text{denom2}} \quad (11)$$

where d is the original estimate disparity at this point, Sp is the aggregated costs.

- (3) Left-Right Consistency (LRC) check to eliminate errors.
- (4) Point cloud growth. The point cloud in object space can be restored from the disparity image. There is no depth data at the position in object space corresponding to the hole in the disparity image. The point cloud around can be used to fill it, and then it can be recovered to the disparity image, so as to repair the hole in the disparity image.

3.2. 3D Surface Model Reconstruction

3D reconstruction of real-world objects using imagery has been an active research field for decades in computer vision as well as in the photogrammetric community [27]. After Microsoft released the Kinect series of RGB-D cameras in 2010, the dense 3D reconstruction based on depth cameras has stirred up research booms. Early representative work was the KinectFusion [45] algorithm proposed by Microsoft's Newcombe in 2011. After that, there have emerged effective algorithms in succession such as BundleFusion [36], Kintinuous [46] and ElasticFusion [47]. Among them, the BundleFusion algorithm proposed by Stanford University in 2017 is one of the best methods for obtaining and reconstructing dense 3D point clouds based on RGB-D cameras. In this paper, the depth data obtained by different stereo matching methods is used to model the 3D surface of indoor scenes using the BundleFusion algorithm, and their performance in fine 3D surface reconstruction was verified by comparing their differences. The thought of the BundleFusion algorithm is shown in Figure 3:

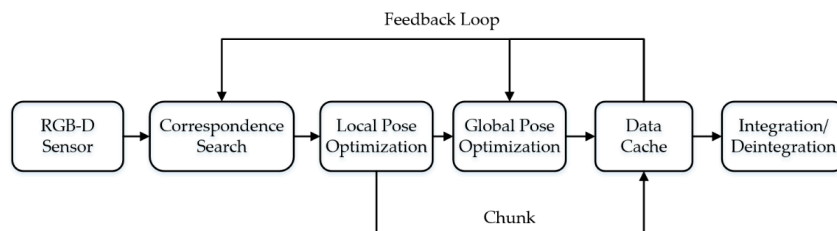


Figure 3. The flow diagram of BundleFusion algorithm.

4. Experiments Data and Results

The quality of infrared stereo matching will be greatly affected by infrared images which are created by infrared light falling on the infrared cameras. There are many factors that affect infrared light, including incident angle, material, distance, ambient light, etc. In order to verify the performance and effect of the ISGSM algorithm proposed in this paper, we collected data in scenes of different complexity. The performance of algorithms under different environmental conditions will be evaluated by changing the environmental factors, such as the scale and depth of the scenes and the intensity and incident angle of light. Figure 4 shows the real scenes of the five sets of data collected in the experiment.

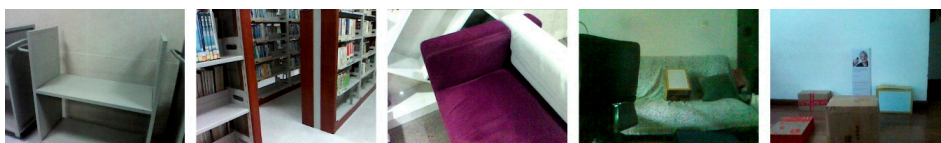


Figure 4. Five real scenes in the experiment.

It needs to be noted that the direct output of the stereo matching algorithm is disparity images, but in the application, such as 3D perception, the data used is actually depth maps. There is a process from disparity image to depth map, and its mathematical model is shown in Formula (12):

$$z = \frac{f \cdot B}{d} \quad (12)$$

where f is the focal length of the camera, B is the baseline length of the binocular camera, d is the disparity value corresponding to the pixel, and z is the depth value corresponding to the pixel.

4.1. Experimental Results Comparison of Different Stereo Matching Algorithms

In order to compare the experimental effects of several state-of-the-art stereo matching algorithms with our proposed algorithm on infrared stereo images of R200, the R200's commercial algorithm (RCA), BM algorithm (BM), SGM algorithm (SGM) [16] and ISGSM algorithm (ISGSM) were implemented with five different scenes in Figure 4. In Figure 5, each column corresponds to a scene. The first row is the RGB images of these scenes. The second row is the infrared images acquired by the left infrared camera. The third row shows depth maps output by R200's commercial algorithm. The fourth row is the experimental results of the BM algorithm. The fifth row of Figure 5 is the experimental results of the SGM algorithm. The experimental results of the ISGSM algorithm are in the sixth row. From the visual effect of the experimental results from the third row to the sixth row in Figure 5, it can be easily found that among the four algorithms, the depth map obtained by the ISGSM algorithm is the most complete with the least holes. Meanwhile, the R200's commercial algorithm has the most holes and the surfaces and edges of objects in these scenes are the most incomplete. However, in general, the ISGSM algorithm is better than the SGM algorithm, the SGM algorithm is better than the BM algorithm, and the BM algorithm is better than R200's commercial algorithm. Additionally, we also find that there are more holes in the occluded area of the object edge and the far away area of the scene in the depth map obtained by R200's commercial algorithm and the BM algorithm.

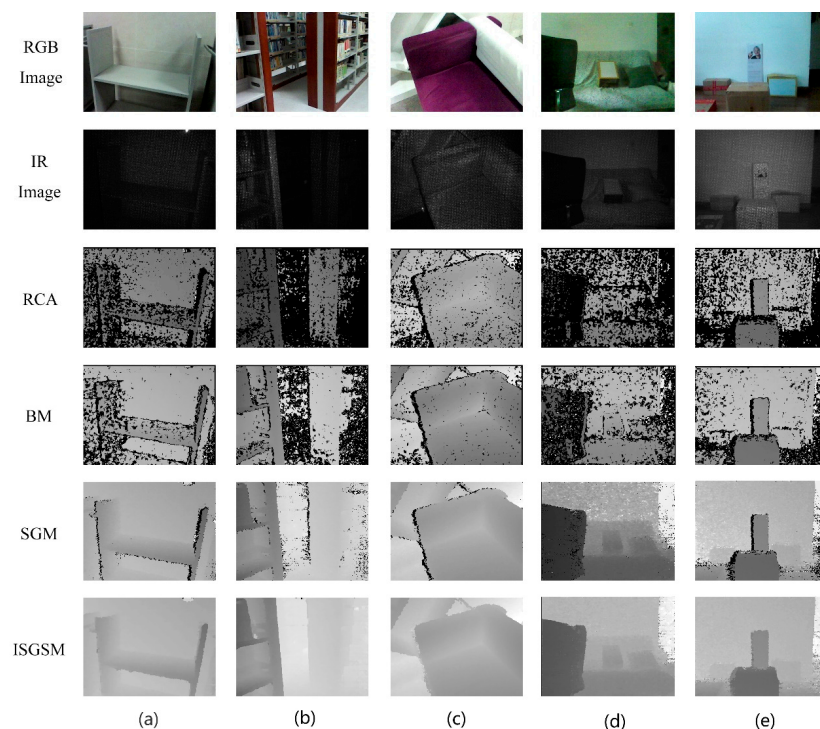


Figure 5. Comparison of results of different methods. (a–e) are, respectively, five scenes for the stereo matching experiment.

In order to verify the effective detection distance and perception ability of R200's commercial algorithm with the worst visual effect and the ISGSM algorithm with the best visual effect, the depth measurement errors of the two algorithms are tested in this paper.

In the experiment, we use a white flat wall to test the precision of the two algorithms. The distance between the R200 and the plane is changed by the caster. The step size is 300 mm, and the distance increases from about 700 mm, until the two algorithms cannot get effective depth data. In the experiment, on the one hand, there is a certain error in the position of R200, and on the other hand, there are errors in the camera's focal length, baseline and physical size of the pixels. These errors belong to systematic errors and can be eliminated by the method of linear regression. Figure 6 shows the RMSE (Root Mean Square Error) of the two algorithms for depth measurement. According to the experimental results, when the depth is within 2 m, RMSE of them is within 20 mm. When the distance increases to 3 m, the RMSE of the R200's commercial algorithm increases faster. Moreover, at 5 m or more, R200's commercial algorithm cannot get valid data. In contrast, the ISGSM algorithm has a higher accuracy within 6 m, and can obtain valid data within 8 m.

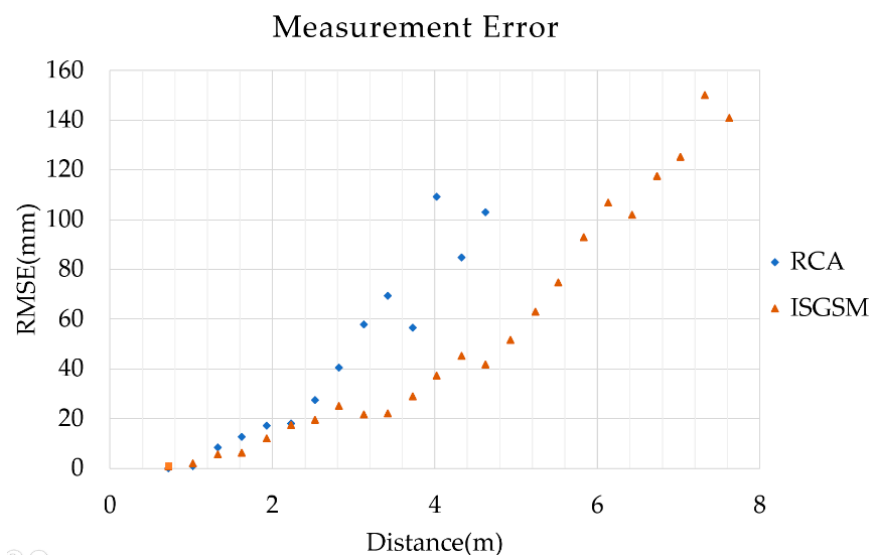


Figure 6. Measurement error of R200's commercial algorithm and the ISGSM algorithm.

Although the ISGSM algorithm obtains more complete, more detailed structure information and a longer distance depth map than the other three algorithms, whether more depth information will bring a higher depth error rate is also an evaluation index that must be considered in 3D perception. Therefore, for the five scenarios in Figure 4, we calculate the error rate of depth information obtained by different algorithms.

Due to the different scenes, there will be different error rates, the difference of which can be an order of magnitude. This is not conducive to better comparison and analysis. Therefore, we normalize the error rate, which means to divide the error rates of an algorithm by the error rate of R200's commercial algorithm as the indicator of error rates. The final result is shown in Figure 7. Compared with the R200's commercial algorithm, the BM algorithm and the SGM algorithm have higher error rates in the four scenes one, two, four and five. Furthermore, BM has the highest error rates overall, it even reaches 4.7 times the error rate of the R200's commercial algorithm, and the difference in different scenes is very large. The ISGSM algorithm's error rate is closer to that of the R200's commercial algorithm in scenes one, four, and five, and it is obviously lower in scenes two and three. Compared with BM and SGM, its fluctuation range is significantly smaller and the overall performance is more stable.

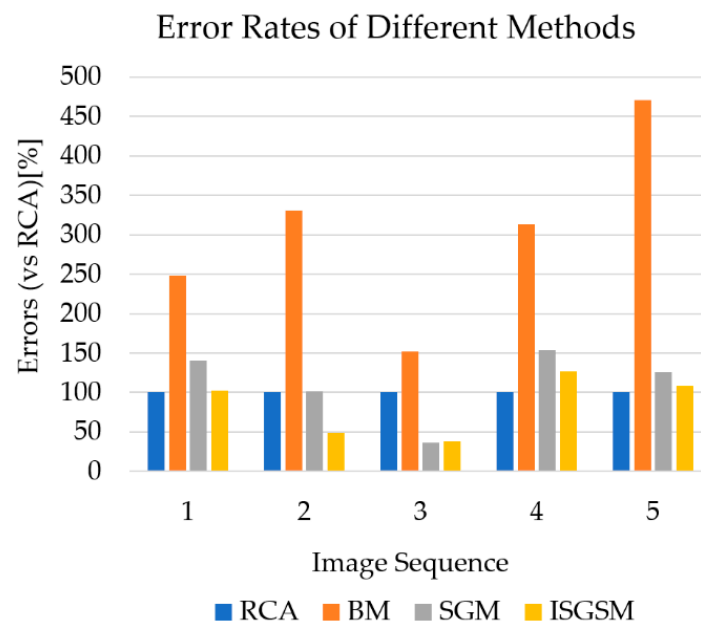


Figure 7. The statistical results of four stereo matching algorithms' error rates.

4.2. 3D Surface Modeling with Different Stereo Matching Algorithms

In order to evaluate the performance of depth maps obtained by different methods in real-time 3D perception, this paper uses the BundleFusion algorithm to reconstruct the indoor 3D surface model. The specific method is to collect the same amount of RGB image data and infrared binocular image data, and then use the R200's commercial algorithm, the SGM algorithm and the ISGSM algorithm to process the infrared binocular images to get depth maps. Next, RGB images and depth maps are used as the input data of the BundleFusion algorithm. Finally, we compare the obtained 3D surface models of different methods. It should be pointed out that the depth map obtained by BM not only has a relatively high hole rate, but also the highest error rate, and the accuracy is not stable in different environments, so it is not included in this experiment.

Figure 8 shows the results of real-time 3D surface reconstruction of an indoor scene. We collected 300 frames of images to reconstruct the surface model. By comparing the local details of models, it can be found that there are obvious differences in the 3D surface reconstructions. Different algorithms have significant differences in the integrity of surface reconstruction; the order of integrity of surface reconstruction is the ISGSM algorithm > the SGM algorithm > R200's commercial algorithm. For example, in the area identified by the red box area in Figure 8, as a result of more complete depth maps, the surface model reconstructed by ISGSM's depth maps is completer than the results of the R200's commercial algorithm and the SGM algorithm. According to the statistics of experimental data, the surface area of the reconstruction model corresponding to the R200's commercial algorithm is about 78.8% of the ISGSM algorithm, and the SGM algorithm is about 91.0% of the ISGSM algorithm. It proves that ISGSM has a distinct advantage in the integrity of 3D reconstruction. For the accuracy evaluation of the three stereo matching algorithms in 3D surface model modeling, we select the white flat desktop in Figure 8 as the study subject. The point cloud of the desktop is clipped from the models obtained by the three algorithms, and the point cloud data is used for plane fitting. Then, the RMSE of the fitting plane is calculated. Figure 9 shows the RMSE value of the fitting plane. The ISGSM algorithm has an RMSE of 1.53 mm with the highest accuracy, the R200's commercial algorithm with an accuracy of 1.64 mm, while the SGM algorithm has an RMSE of 2.94 mm with the lowest accuracy.

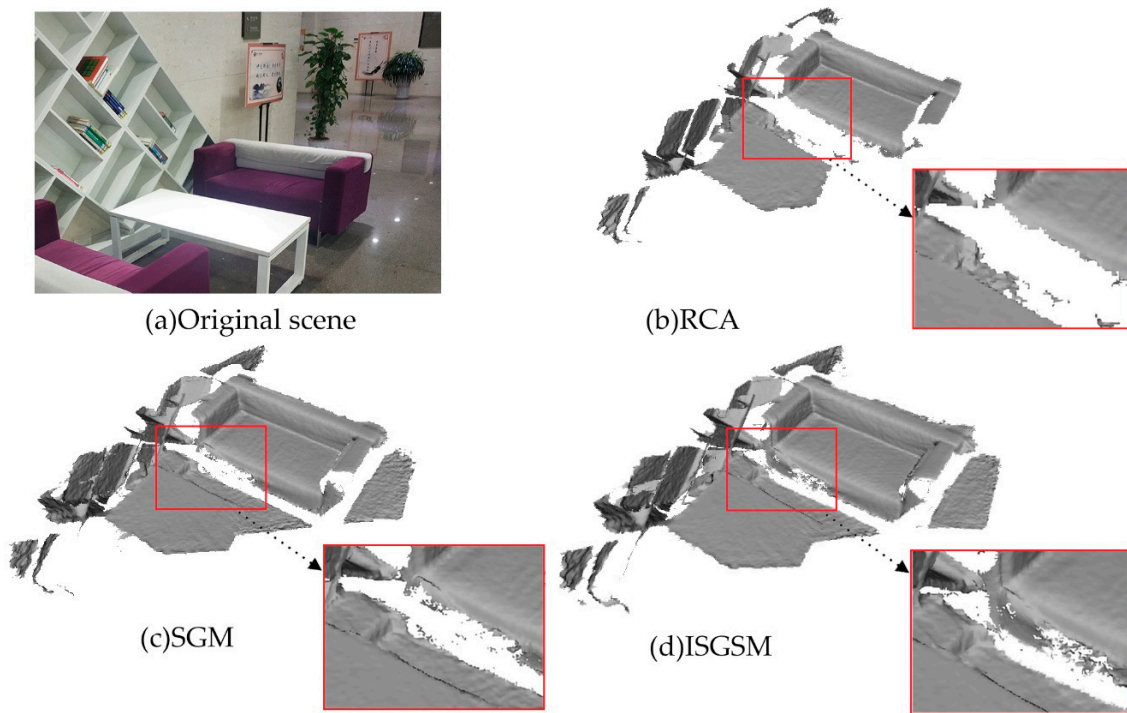


Figure 8. Comparison of 3D surface reconstruction results: (a) the real scene, (b) result of R200’s commercial algorithm, (c) result of the SGM algorithm, (d) result of the ISGSM algorithm.

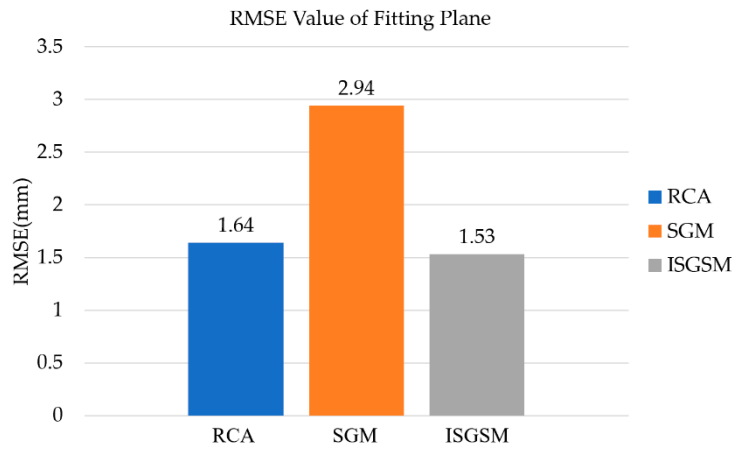


Figure 9. RMSE value of fitting plane using desktop point cloud.

5. Discussion

Through the above comparative experiments and their results analysis, it is proved that the novel ISGSM algorithm proposed in this paper can obtain a higher quality depth map in a larger detection range, which allows the R200 sensor to acquire denser depth information with higher accuracy and better perform more demanding and more complex 3D perception. The overall effect of the depth map calculated by ISGSM is much better than that of the RCA, BM and SGM, especially in areas where the brightness of infrared speckle is weak. The leading causes of weak infrared brightness include: the distance is too far, the angle of incidence is too large, the object surface reflection coefficient is low, the object has a specular reflection, etc. These reasons directly lead to the weak texture and the lack of matching information. This problem is more obvious in the RCA. In Figure 5, for areas with higher brightness, i.e., areas with higher infrared reflection intensity, due to their strong texture, the algorithm can better keep the edge characteristics of the indoor scene, and the continuity is better. However, for areas with weaker textures, i.e., parts with lower grayscale brightness. For instance, in the middle

of scene (b) of Figure 5 and the upper right of scenes (d) and (e) of Figure 5, the matching effect is poor because of the longer detection distance, with many holes in the corresponding depth map. As for the floor in Figure 5's scene (d) and (e), because of its smoothness and a large incident angle compared to the wall, the reflected infrared light is also too weak for RCA to match. As a result, it is hard to perform well for accurate and complete indoor 3D perception.

Whereas, BM enhances the correlation between left and right matching primitives by block matching strategy, and SGM constructs a global energy function by means of a semi-global strategy for global optimization. So, they perform better than the RCA in texture-less areas. However, because BM is very sensitive to noise, although the integrity of its depth map acquired in real time is better than the RCA, it also has problems such as many holes and so close a detection distance. Meanwhile, as there are a lack of efficient methods for reliability examination, there are usually a number of errors in its depth maps, especially in texture-less areas. We can find these deficiencies in Figures 5 and 7. SGM fully shows the advantages of the semi-global algorithm. Only in the edge area of the object, due to the occlusion, is the infrared speckle pattern incomplete or messy, which results in the lack of the available matching texture, so it is hard to match and will lead to holes in the depth map at the corresponding edges. Additionally, Figure 5 clearly shows that the semi-global algorithm has a superior advantage over BM and RCA in areas where infrared speckle is weak due to the long distance. In particular, in Figure 5's scene (d), there is a chair back to the R200 camera on the left side of the image. Although the chair is very close to the camera, the reflected light is still weak due to its leather material and the angle of the chair. Therefore, it is also difficult for local methods to match while the semi-global method can achieve good results.

It can be clearly seen from Figures 5, 7 and 9, the completeness and accuracy of ISGSM's depth maps are significantly better than those of SGM. The edges of the objects are also kept more completely than SGM in every scene. Primarily, this is because SGM lacks the means to suppress noise, while ISGSM uses a Gaussian filter to weaken the influence of noise on matching and enhance the mutual information of the left and right images. Secondly, dynamic threshold selection of parameters is taken for promoting the adaptability of ISGSM to different indoor scenes. For example, if the mutual information is detected below the set threshold, ISGSM will use a larger block for cost calculation in order to incorporate more information. Similar to SGM, ISGSM adopts the semi-global strategy to achieve 2D global optimization, which is also very important. Furthermore, the sub-pixel interpolation operation also makes the continuity of the depth map better, and the depth value will become more accurate. By applying the above-mentioned improvements comprehensively, ISGSM can obtain more dense and accurate depth maps. Besides, real-time 3D perception has high requirements on the efficiency of stereo matching. The complexity of ISGSM and SGM is almost close, so ISGSM can provide completer depth data with a longer detection distance and higher accuracy in real time.

6. Conclusions

In this paper, we propose a novel infrared stereo matching algorithm—ISGSM—to obtain high-quality depth maps for real time indoor 3D perception with the RGB-D sensor. In this method, the idea of semi-global matching and a sliding window is adopted, and the mutual information and correlation between binocular infrared images are enhanced by a Gaussian filter, which effectively suppresses image noise. The dynamic threshold selection of matching window size is also realized to improve the adaptability of the algorithm to different scenes. Meanwhile post-processing techniques, such as point cloud growth, reduce the holes in the depth map. These improvements make ISGSM able to achieve better matching and obtain more dense and precise depth maps. Through the specific experiment, it is shown that ISGSM can obtain depth maps with greater integrity, higher quality and a longer detection range in real time, especially at the edge of the object with finer details. Using the complete real-time indoor 3D perception solution which integrates the novel matching algorithm and BundleFusion, we demonstrate in the real indoor scene that our method is able to generate high-quality real-time reconstructions. The surface model it reconstructs has a higher accuracy and

better integrity. Therefore, we demonstrate that our approach outperforms state-of-the-art techniques. Additionally, the work presents an improved method for the stereo matching algorithm used on the popular RealSense RGB-D cameras. This work is quite valuable, as a software improvement like this could provide users all over the world with a vastly improved product at no additional cost and prolong the lifespan of these devices, as customers would not have to replace them for improved hardware.

Author Contributions: Ming Li and Jiageng Zhong proposed the methodology and wrote the paper; Jiageng Zhong and Ming Li conceived and designed the experiments; Jiageng Zhong, Ming Li, Xuan Liao and Jiangying Qin performed the experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China, grant numbers 2018YFB0505400, the National Natural Science Foundation of China (NSFC), grant number 41901407 and the LIESMARS Special Research Funding.

Acknowledgments: The authors would like to thank the LIESMARS of Wuhan university for the supporting computing environment. Meanwhile, we thank the editors and reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qin, J.; Li, M.; Liao, X.; Zhong, J. Accumulative Errors Optimization for Visual Odometry of ORB-SLAM2 Based on RGB-D Cameras. *Isprs Int. J. Geo-Inf.* **2019**, *8*, 581. [[CrossRef](#)]
2. Li, M.; Chen, R.; Liao, X.; Guo, B.; Zhang, W.; Guo, G. A Precise Indoor Visual Positioning Approach Using a Built Image Feature Database and Single User Image from Smartphone Cameras. *Remote Sens.* **2020**, *12*, 869. [[CrossRef](#)]
3. Stotko, P.; Weinmann, M.; Klein, R. Albedo estimation for real-time 3D reconstruction using RGB-D and IR data. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 213–225. [[CrossRef](#)]
4. Bäuml, B.; Schmidt, F.; Wimböck, T.; Birbach, O.; Dietrich, A.; Fuchs, M.; Friedl, W.; Frese, U.; Borst, C.; Grebenstein, M.; et al. Catching flying balls and preparing coffee: Humanoid rollin'justin performs dynamic and sensitive tasks. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 3443–3444.
5. Henry, P.; Krainin, M.; Herbst, E.; Ren, X.; Fox, D. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *Int. J. Robot. Res.* **2012**, *31*, 647–663. [[CrossRef](#)]
6. Endres, F.; Hess, J.; Sturm, J.; Cremers, D.; Burgard, W. 3-D mapping with an RGB-D camera. *IEEE Trans. Robot.* **2013**, *30*, 177–187. [[CrossRef](#)]
7. Remondino, F. Heritage recording and 3D modeling with photogrammetry and 3D scanning. *Remote Sens.* **2011**, *3*, 1104–1138. [[CrossRef](#)]
8. Zhang, J.; Singh, S. LOAM: Lidar Odometry and Mapping in Real-time. In Proceedings of the Robotics: Science and Systems Conference (RSS), Berkeley, CA, USA, 14–16 July 2014; pp. 109–111.
9. Kuhnert, K.D.; Stommel, M. Fusion of Stereo-Camera and PMD-Camera Data for Real-Time Suited Precise 3D Environment Reconstruction. In Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2006, Beijing, China, 9–15 October 2006; pp. 4780–4785.
10. Kuan, Y.W.; Ee, N.O.; Journal, L.S. Comparative Study of Intel R200, Kinect v2, and Primesense RGB-D Sensors Performance Outdoors. *IEEE Sens. J.* **2019**, *19*, 8741–8750. [[CrossRef](#)]
11. Foix, S.; Alenya, G.; Torras, C. Lock-in time-of-flight (ToF) cameras: A survey. *IEEE Sens. J.* **2011**, *11*, 1917–1926. [[CrossRef](#)]
12. Hisatomi, K.; Kano, M.; Ikeya, K.; Katayama, M.; Mishina, T.; Iwadate, Y.; Aizawa, K. Depth Estimation Using an Infrared Dot Projector and an Infrared Color Stereo Camera. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 2086–2097. [[CrossRef](#)]
13. Shengjun, T. RGB-D Indoor High-Precision 3D Mapping Method for Multi-View Image Enhancement. Ph.D. Thesis, Wuhan University, Wuhan, China, 2017.
14. Jiao, J.; Yuan, L.; Tang, W.; Deng, Z.; Wu, Q. A Post-Rectification Approach of Depth Images of Kinect v2 for 3D Reconstruction of Indoor Scenes. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 349.

15. Chen, H.; Wang, K.; Yang, K. Improving RealSense by Fusing Color Stereo Vision and Infrared Stereo Vision for the Visually Impaired. In Proceedings of the 2018 International Conference on Information Science and System, Wuhan, China, 20–22 April 2018; pp. 142–146.
16. Hirschmuller, H. Accurate and efficient stereo processing by semi-global matching and mutual information. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Washington, DC, USA, 20 June 2005; Volume 2, pp. 807–814.
17. Keselman, L.; Iselin, Woodfill, J.; Grunnet-Jepsen, A.; Bhowmik, A. Intel RealSense Stereoscopic Depth Cameras. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017.
18. Brown, M.Z.; Burschka, D.; Hager, G.D. Advances in computational stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 993–1008. [[CrossRef](#)]
19. Zabih, R.; Woodfill, J. Non-parametric local transforms for computing visual correspondence. In *European Conference on Computer Vision*; Springer: Berlin, Heidelberg, 1994; pp. 151–158.
20. Prince, S.J.; Eagle, R.A. Weighted directional energy model of human stereo correspondence. *Vis. Res.* **2000**, *40*, 1143–1155. [[CrossRef](#)]
21. Veksler, O. Stereo correspondence by dynamic programming on a tree. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20 June 2005; Volume 2, pp. 384–390.
22. Kolmogorov, V.; Zabih, R. Computing visual correspondence with occlusions using graph cuts. In Proceedings of the Eighth IEEE International Conference on Computer Vision, Vancouver, BC, Canada, 7–14 July 2001; Volume 2, pp. 508–515.
23. Sun, J.; Zheng, N.N.; Shum, H.Y. Stereo matching using belief propagation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 787–800.
24. Scharstein, D.; Szeliski, R. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [[CrossRef](#)]
25. Hirschmuller, H.; Scharstein, D. Evaluation of Stereo Matching Costs on Images with Radiometric Differences. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 1582–1599. [[CrossRef](#)]
26. Hirschmuller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [[CrossRef](#)] [[PubMed](#)]
27. Wenzel, K.; Rothmel, M.; Fritsch, D. SURE—The ifp Software for Dense Image Matching. In *Photogrammetric Week '13*; Fritsch, D., Ed.; Wichmann: Stuttgart, Germany, 2013; pp. 59–70.
28. Rothmel, M.; Wenzel, K.; Fritsch, D.; Haala, N. SURE: Photogrammetric Surface Reconstruction from Imagery. In Proceedings of the LC3D Workshop, Berlin, Germany, 4–5 December 2012.
29. Yan, L.; Fei, L.; Chen, C.; Ye, Z.; Zhu, R. A Multi-View Dense Image Matching Method for High-Resolution Aerial Imagery Based on a Graph Network. *Remote Sens.* **2016**, *8*, 799. [[CrossRef](#)]
30. Hirschmuller, H. Stereo vision in structured environments by consistent semi-global matching. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 2386–2393.
31. Chai, Y.; Yang, F. Semi-Global Stereo Matching Algorithm Based on Minimum Spanning Tree. In Proceedings of the 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Xi'an, China, 25–27 May 2018; pp. 2181–2185.
32. Lohman, M.; Kim, J. SGM-based dense disparity estimation using adaptive Census transform. In Proceedings of the International Conference on Connected Vehicles and Expo (ICCVE), Las Vegas, USA, 2–6 December 2013; pp. 592–597.
33. Humenberger, M.; Engelke, T.; Kubinger, W. A census-based stereo vision algorithm using modified semi-global matching and plane fitting to improve matching quality. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), San Francisco, CA, USA, 13–18 June 2010; pp. 77–84.
34. Seki, A.; Pollefeys, M. SGM-Nets: Semi-Global Matching with Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 1 July 2017; pp. 21–26.
35. Yang, W.; Li, X.; Yang, B.; Fu, Y. A Novel Stereo Matching Algorithm for Digital Surface Model (DSM) Generation in Water Areas. *Remote Sens* **2020**, *12*, 870. [[CrossRef](#)]

36. Dai, A.; Nießner, M.; Zollhöfer, M.; Izadi, S.; Theobalt, C. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph.* **2017**, *36*, 1. [[CrossRef](#)]
37. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [[CrossRef](#)]
38. Heikkila, J.; Silvén, O. A Four Step Camera Calibration Procedure with Implicit Image Correction. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 17–19 June 1997; Volume 22, pp. 1106–1112.
39. Camera Calibration Toolbox for Matlab. Available online: <http://www.vision.caltech.edu/bouguetj/calib-doc> (accessed on 11 December 2019).
40. Lu, J.; Zhang, X.; Dong, D.; Fang, Y. A stereo matching algorithm based on census transformation and dynamic programming. In Proceedings of the 33rd Chinese Control Conference, Nanjing, China, 28–30 July 2014; pp. 8271–8276.
41. Jin, S.; Cho, J.; Pham, X.D.; Lee, K.M.; Park, S.-K.; Kim, M.; Jeon, J. FPGA Design and Implementation of a Real-Time Stereo Vision System. *Ieee Trans. Circuits Syst. Video Technol.* **2010**, *20*, 15–26.
42. Fathi, M.; Sheikhaei, S.; Tavakoli, J. Low-cost and Real-time Hardware Implementation of Stereo Vision System on FPGA. In Proceedings of the 2019 27th Iranian Conference on Electrical Engineering (ICEE), Yazd, Iran, 30 April–2 May 2019; pp. 258–263.
43. Birchfield, S.; Tomasi, C. Depth discontinuities by pixel-to-pixel stereo. *Int. J. Comput. Vis.* **1999**, *35*, 269–293. [[CrossRef](#)]
44. Zhu, C.; Chang, Y.Z. Stereo matching for infrared images using guided filtering weighted by exponential moving average. *Iet Image Process.* **2020**, *14*, 830. [[CrossRef](#)]
45. Newcombe, R.A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Fitzgibbon, A.W. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, Switzerland, 26–29 October 2011; pp. 127–136.
46. Whelan, T.; Kaess, M.; Fallon, M.; Johannsson, H.; Leonard, J.J.; McDonald, J. Kintinuous: Spatially extended KinectFusion. In Proceedings of the 3rd RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras, Sydney, Australia, 9–10 July 2012.
47. Whelan, T.; Leutenegger, S.; Salas-Moreno, R.; Glocker, B.; Davison, A. ElasticFusion: Dense SLAM without a pose graph. *Proc. Robot. Sci. Syst.* **2015**, 1–9. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).