REAL-TIME DEPTH MAP ESTIMATION FROM INFRARED STEREO IMAGES OF RGB-D CAMERAS

Jiageng Zhong [1], Ming Li [1,2,*], Xuan Liao[3], Jiangying Qin [1], Hanqi Zhang [1], Qi Guo [1]

[1]State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[2]Department of Physics, ETH Zurich, Zurich 8093, Switzerland - lisouming@whu.edu.cn
[3]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong 999077, China

KEY WORDS: Stereo Matching, Infrared Image, RGB-D Camera, Depth Map, Disparity

ABSTRACT:

RGB-D cameras are novel sensing systems that can rapidly provide accurate depth information for 3D perception, among which the type based on active stereo vision has been widely used. However, there are some problems exiting in use, such as the short measurement range and incomplete depth maps. This paper presents a robust and efficient matching algorithm based on semi-global matching to obtain more complete and accurate depth maps in real time. Considering characteristics of captured infrared speckle images, the Gaussian filter is performed firstly to restrain noise and enhance the relativity. It also adopts the idea of block matching for reliability, and a dynamic threshold selection of the block size is used to adapt to various situation. Moreover, several optimizations are applied to improve precision and reduce error. Through experiments on the Intel Realsense R200, the excellent capability of our proposed method is verified.

1. INTRODUCTION

Real-time and high-quality 3D space perception is a key technology for SLAM and AR (Endres and et al., 2013). In the current research, typical devices for 3D perception include RGB cameras, RGB-D cameras, and LiDAR. The RGB-D camera combines the advantages of LiDAR and RGB camera. It can quickly obtain high-quality geometric information and color information with low cost, thus has great research value and application potential, especially in indoor environments (Jiao and et al., 2017). There are several methods for RGB-D cameras to get depth maps. For example, Apple's Prime Sense sensor uses structured light (SL) to implement scene perception technology (Boehm, 2014). The Kinect v2 released by Microsoft uses the time-of-flight (ToF) principle to obtain depth maps with a higher frame rate but a lower resolution (Foix and et al., 2011). Intel's portable consumer-grade RGB-D cameras include the Intel R200 (2015), D415 and D435 (2018), which are based on active stereo vision (ASV) for data acquisition and processing (Kuan and et al., 2019). In particular, they are usually equipped with one NIR texture projector and a pair of NIR cameras and use stereo matching for depth estimation. They are widely used in robot navigation and positioning, indoor 3D Mapping and modelling because of their low cost and portability (Chen and et al., 2018). The binocular stereo matching module provided by Intel on the R200 is based on a local matching method (Keselman and et al., 2017). In this way, it can match infrared stereo images at a higher frame rate. However, there are many pixels that cannot be matched effectively, which causes many holes and a short valid detection distance, thereby limiting its application scenarios. In practice, its valid detection distance is no more than 4m.

In this paper, the infrared stereo image of R200 camera is used for experiment, and a stereo matching algorithm for infrared speckle image is proposed to improve the 3D space perception ability of RGB-D sensor based on active stereo vision technology, taking the deficiency of R200 into account.

2. RELATED WORK

Nowadays, the problem of getting high-quality depth maps through the stereo vision technique is one of the most actively studied problems in many applications. According to the different matching strategies, stereo matching algorithms can be simply divided into local methods and global methods. Local methods are based on correlation and can have high efficiency; therefore, they are suitable for real-time applications. Common local matching algorithms calculate disparity by comparing the information in the local window (Brown and et al., 2003.). They can perform pixelwise matching, so that they can obtain dense depth maps. However, there are often mismatches in the texture-less area, and it is difficult to retain depth continuity, so it is unlikely to obtain accurate matching results. Common global methods that can achieve higher accuracy include Dynamic Programming (Veksler, 2005), Belief Propagation (Sun and et al., 2003) and Graph Cut (Kolmogorov, 2001). They convert the matching problem into finding the global optimization of an energy function of the disparity image. However, they have much higher calculation cost and consume more memory during runtime, so that they are not suitable for real-time application.

Considering the characteristics of above two methods, researchers propose a semi-global strategy (Hirschmuller, 2005) which combines the advantages of both methods. The SGM algorithm has attracted considerable attention of many researchers. It performs 2D global optimization by constraining the 1D path in multiple directions, and maintains higher efficiency while obtaining higher quality disparity images. Since it does not include all pixels in the calculation, its complexity is lower than global methods, which allows it to run in real time. In view of its superiority, there are many researches based upon it. Its modified algorithms like tSGM in SURE (Rothermel and et al., 2012) and SGBM (Yang and et al., 2020) have been proposed according to the different characteristics of different scenes. Moreover, SGM-Nets (Seki and Pollefeys, 2017) uses SGM in

* Corresponding author

combination with a neural network, which can greatly enhance the performance in many situations.

While in practice, the problem on how to obtain high-quality disparity images with infrared speckle images in real time is not solved properly. Thereinto, the R200 is a representative RGB-D camera based on infrared speckle and stereo vision technology for the depth estimation of indoor scenes. The binocular stereo matching module provided by Intel on the R200 is based on a local matching method. In this way, it can match infrared stereo images at a higher frame rate.

However, there are many pixels that cannot be matched effectively, which causes many holes and a short valid detection distance, thereby limiting its application scenarios. For this, we propose an advanced infrared stereo matching algorithm. Inspired by the work of Semi-Global Matching (SGM), a semi-global strategy is adopted, in addition to improvement aimed at infrared the characteristics of speckle images. Experimental results are used to verify the validity and superiority of the method.

3. METHODOLOGY

3.1 R200′s Commercial Algorithm

Local matching method is used for stereo matching in the R200. The 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 (Hirschmuller and Scharstein, 2008). For a pixel in the match image, a Census transformation window with a size 7 × 7 is selected. Then a 0/1-bit string for the Census transformation can be obtained (Lu and et al., 2014). In the same way, the bit string for the search point of the target image is obtained. Then, a 64-disparity search is performed, and costs are aggregated with a 7 × 7 box filter. The best-fit candidate is selected. Finally, after a subpixel refinement and a set of filters, the disparity image is obtained.

3.2 Semi-Global Matching Algorithm

The semi-global matching algorithm has its variant. In this paper, SGM with BT (Birchfield and Tomasi) (Hirschmuller, 2005) is selected as the comparative method, whose key steps include cost calculation, cost aggregation and disparity computation.

In this algorithm, while pixelwise cost calculation is subject to interference from noise and other factors, an energy function that depends on the disparity image is defined to support smoothness by penalizing changes of neighbouring disparities, and the problem of matching is then transformed into finding the disparity image D that minimizes the energy function $E(D)$. Researches show that the effective method to achieve 2D global optimization is to accumulate 1D matching costs from multiple directions. In this way, the aggregated cost can be calculated. Then, by selecting the disparity d that minimizes cost for each pixel, the disparity image can be obtained. At last, a subpixel interpolation will be performed to improve accuracy. Furthermore, as there are some matching errors, it uses filters to eliminate them.

3.3 Our Proposed Algorithm

As stated earlier, existing methods are based on different requirements and applications, and there are still some problems to solve. In order to achieve better matching of infrared images for more accurate 3D perception with RGB-D cameras based on ASV, an improved method is proposed in this paper. Based on the SGM algorithm, the semi-global matching strategy is adopted in ours. And there are several improvements aimed at characteristics of infrared speckle images. A detailed flowchart of our algorithm is presented in Figure 1.

Because of the low power of the infrared projector of R200, the reflected infrared ray in many places in the scene is quite weak, which directly leads to texture-less regions of the infrared image (Zhu and Chang, 2019). Also, as the infrared light intensity can be affected by a variety of factors, for instance, the angle of incidence and distance, there are usually some noises in the image. To address these issues, Gaussian filtering is performed firstly after capturing two infrared stereo images. Gaussian filtering can not only reduce noises of the infrared images, but also can enhance the correlation of the stereo infrared images. As a result, abnormal value caused by noise is weakened and the correlation of texture-less regions is strengthened. Our experiments prove 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%.

The BT algorithm (Birchfield and Tomasi, 1999) is performed in cost calculation. The idea of block matching (Scharstein and Szeliski, 2002) is also adopted to merge the information of neighborhood pixels into the calculation, as the BT algorithm is a pixelwise method which is easily infected by noise and causes mismatches or errors. Through doing this, matching can be more robust. However, a fixed size block does not suit all circumstances. If the block size is too large, it will result in over-smoothness and more calculation. And if too small, it may have little effect. Therefore, before cost calculation, the dynamic threshold selection of the block size based on mutual information is applied. In other word, the block size is selected according to the mutual information between the two Gaussian filter images. Then our algorithm adopts BT for cost calculation. The cost calculated by our algorithm 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). Compared to the original BT, the second part of cost is to increase the similarity for better matching.

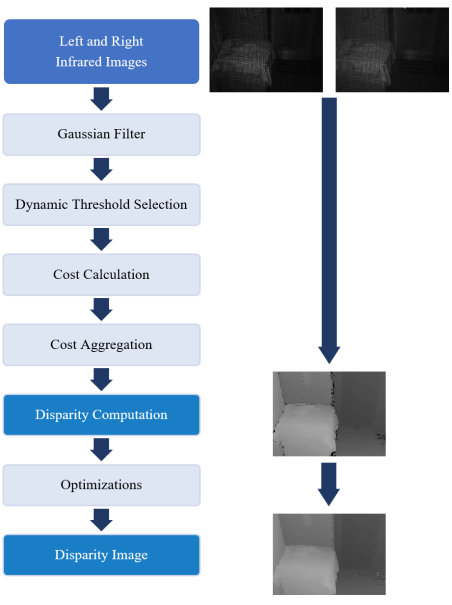

Figure 1. A flow chart of our algorithm.

In cost aggregation, based on the idea of SGM, a global smoothness constraint is approximated by combining many 1D constraints. So, the stereo matching problem is transformed into searching the optimal solution of the energy function. In this way, the algorithm can output high-quality disparity images at a high frame rate, which allows it to be applied to real-time application scenarios. Then there are several optimization steps to fix up some problems in the preliminary disparity image, including uniqueness test, sub-pixel interpolation, left-right consistency check and point cloud growth.

3.4 Stereo Depth

The output of the stereo matching algorithm is a disparity map, which is not a depth map that can be directly used for 3D perception. It needs to convert the disparity value to the depth value. The basic principle is shown in Figure 2. Here f is the focal length of the camera, and B is the baseline of the left and right infrared cameras. The point P is an object point, and P_L and P_R are respectively the image points on the left and right images. x_L and x_R are respectively the x-coordinates of P_L and P_R. The depth z is the distance of the point P from the camera.

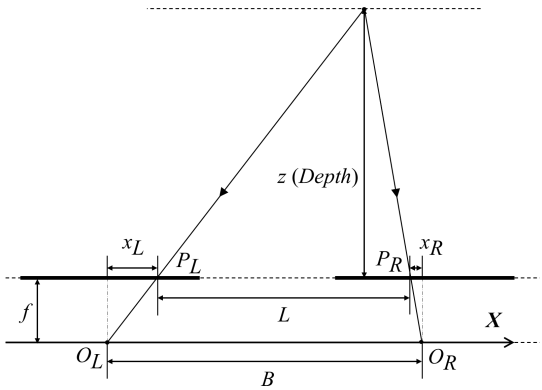

Figure 2. A schematic diagram of binocular stereo vision.

According to the principle of R200, the infrared speckle is first emitted by the infrared projector and irradiated on P, and then P is imaged in two infrared cameras on the left and right respectively. Ideally, the left and right camera focal lengths are equal, with only displacement on the x-axis and parallel main optical axis. Therefore, the imaging position of P in the two images is theoretically different only on the x-axis (corresponding to the x-axis in Figure 2), and the difference in its position is the disparity, denoted by d, and then L can be calculated.

$$d = x_L - x_R \tag{1}$$

$$L = B - d \tag{2}$$

According to the Similar Principle of Triangle, the Formula (3) can be obtained.

$$\frac{z}{L} = \frac{z+f}{B} \tag{3}$$

By combining Formula (1) - (3), the depth z is computed, as shown in Formula (4).

$$z = \frac{f \cdot B}{d} \tag{4}$$

The focal length f and baseline B in Formula (4) can be obtained by camera calibration, and the disparity d is calculated by stereo matching algorithm, thus the depth value can be computed.

4. EXPERIMENTS

4.1 Depth Maps of Different Matching Algorithms

To validate the efficiency of proposed algorithms, three indoor scenes have been selected for evaluation, and each scene corresponds to a row in Figure 3. The R200's commercial algorithm (RCA), SGM algorithm (SGM) and our algorithm are implemented for the comparative experiments. In Figure 3, the first column is the RGB images of three scenes. The second column is the infrared images acquired by the left infrared camera of the R200. And the practical effects of RCA, SGM and our algorithm are demonstrated in the third to fifth columns. Overall, from the visual effect of the depth maps of different methods, the depth maps obtained by our algorithm are the most complete compared to those of other methods, while those of RCA have the most holes and incomplete edges of objects. And the performance of the SGM is between RCA and our algorithm.

The RCA is a local method. As shown in Figure 3, the RCA can achieve good matching results in texture-rich areas, like the desk in Figure 3's scene (a), but does not work well in texture-less areas, like the floor in Figure 3's scene (c). The direct cause of the lack of texture is the weak infrared brightness which is easily affected by many factors, such as too far distance, too large angle of incidence, specular reflection on the surface. Concretely, the left part of the wall in Figure 3's scene (b) is far away from the camera and tends to reflect light easily, therefore, local reflection light is too weak and leads to indigent texture. Similar things exist in the floor in Figure 3's scene (c). The texture in these areas is so weak that it is hard for RCA to perform accurate and complete indoor 3D perception.

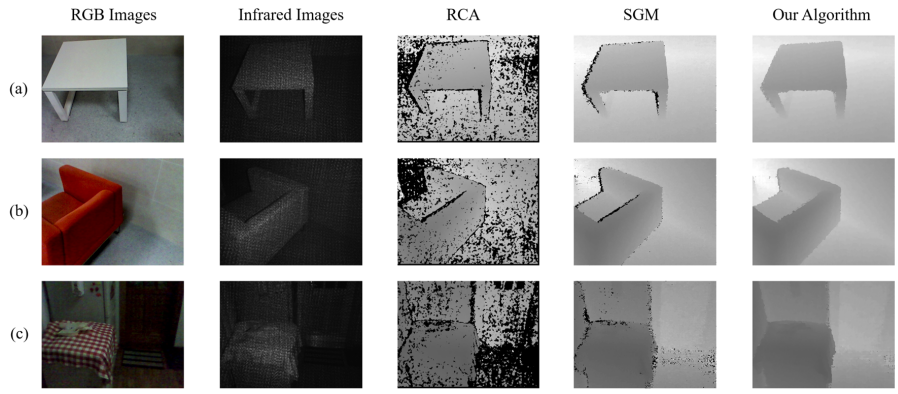

Figure 3. Comparison of results of different methods. (a–c) correspond to three scenes.

In contrast, the advantage of semi-global matching is obvious. SGM constructs a global energy function by means of a semi-global strategy for global optimization. That means that it takes into more pixels compared with RCA and can perform better than the RCA in texture-less areas, especially distant objects. So, SGM will no doubt have a longer detection range than RCA, which allows it to perform more comprehensive sensing. On the edges of objects with occlusion, the infrared speckle pattern may be incomplete or messy, so it is difficult to find right match and will lead to holes in the final depth map.

The last column of Figure 3 shows the result of our algorithm. From Figure 3, it can be concluded that the overall visual effect of our algorithm is better than SGM. In depth maps of ours, more complete edges and less abnormal value can be seen. Both algorithms adopt the semi-global matching strategy, but several improved methods are applied in our algorithm for the shortages of SGM. First, considering the effects of noise, our algorithm uses Gaussian Filter to suppress noise, while increasing the similarity. Next, block matching is used to integrate the information in an image block for robustness. Such as the floor in the lower right section of scene (c) of Figure 3, our algorithm can perform well whereas there are some abnormal values produced by SGM. It's worth mentioning that dynamic threshold selection of parameters ensures our algorithm's adaptability. Finally, due to last several optimization steps, the quality of depth maps is improved.

In order to evaluate these methods quantitatively, the error rate is used as the criterion. It should be noted that the error cannot be directly calculated due to the lack of standard datasets. Therefore, the error is captured by manual statistics. For the three scenarios in Figure 3, the error rates of depth information obtained by different algorithms are calculated, and the final result is shown in Figure 4. The error rate is normalized as the difference of error rates in different scenes can be an order of magnitude. The statistical data directly shows that among these three algorithms, SGM has the highest error rate, while our algorithm has the lowest. One of important reasons is that SGM lacks ways to eliminate wrong matching pixels. Therefore, errors occur especially at edges. As for our algorithm, the depth maps are smoother due to the usage of Gaussian Filter and block matching. Because there are more pixels used during matching, part of errors can be avoided. Moreover, the complexity of our algorithm and SGM is almost close. So, in general, our algorithm can provide completer depth data with a longer detection distance and higher accuracy in real time.

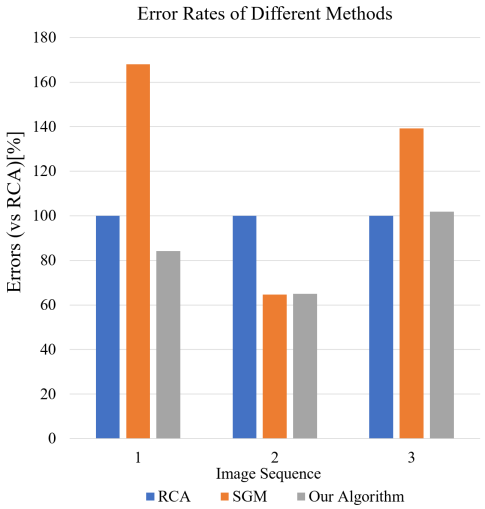

Figure 4. The statistical results of three stereo matching algorithms' error rates.

4.2 Error Analysis of Depth Value

The error of depth value is an important criterion to comprehensive evaluation of the effective detection distance and perception ability of stereo matching algorithms. Therefore, a test of the depth measurement errors is applied between RCA with the worst visual effect and our algorithm with the best visual effect. A white flat wall is used to test the precision of the two algorithms in the experiment. 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 addition, due to the influence of camera position, pixel physical size and other factors, the results have some systematic errors, which can be corrected by linear regression.

On the basis of Formula (4), the partial derivative of z with respect to d is calculated. Then, z is used to replace d in the formula to obtain the quantitative relationship between the error of depth value and the size of depth value. The mathematical expression is shown in Formula (5).

$$|\epsilon_z| = \frac{z^2}{f \cdot B} \cdot |\epsilon_d| \tag{5}$$

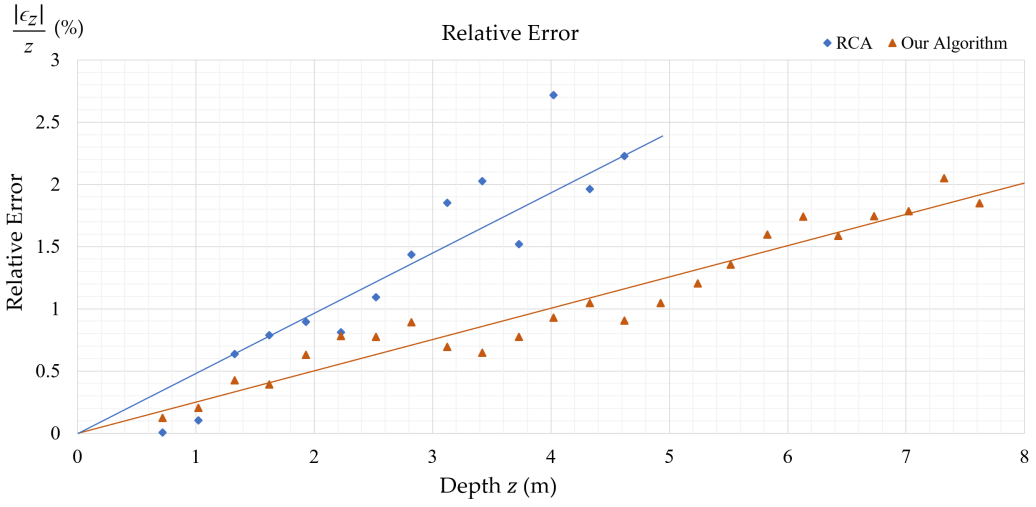

Figure 5. The relative error of RCA and our algorithm

Here, given that errors in disparity space are usually constant for a stereo system, $|\epsilon_d|$ can be treated as a constant. Besides, f and B are also constant and can be obtained by camera calibration. In addition, due to the influence of factors such as camera placement and pixel physical size, the experimental results have certain systematic errors, which can be corrected by methods such as unary linear regression. Considering that the relative error can usually better reflect the credibility of measurements, its mathematic expression is derived here as shown in Formula (6).

$$\frac{|\epsilon_z|}{z} = \frac{|\epsilon_d|}{f \cdot B} \cdot z \tag{6}$$

In Formula (6), $|\epsilon_d|$, f and B are constant. It is inferred that the relative error is linear in z, so it can be fitted with a linear model. The results of the experiment are shown in Figure 5.

As shown, both of the relative errors of two algorithms are no more than 1% within 2 m. When depth increases to 3 m, the relative error of RCA increases faster. At 5 m or more, RCA cannot get valid data. By contrast, our algorithm has better precision and its detection distance can reach to more than 7 m. As distance increases, light lessens and textures weaken. It is hard for RCA to make valid matches, and our algorithm can still match accurately by taking advantages of semi-global matching. Besides, the intensity of reflected infrared light frequently is variable, which will lead to measurement errors. Because the semi-global method uses more pixel values during calculation, our algorithm can reduce the influence of the instability more effectively and obtain more accurate measurements, especially in the distance.

5. CONCLUSION

In this paper, we presented a novel infrared stereo matching algorithm to improve the stereo vision of the Intel stereoscopic RGB-D sensors. Targeted at the characteristics of infrared speckle images, our algorithm uses Gaussian Filter to resist noise, and adopts the semi-global strategy and block strategy with a dynamic threshold selection to enhance the quality of matching. It has been shown that, compared with the existing methods, our algorithm can obtain depth maps with greater integrity, higher quality and a longer detection range in real time. As this kind of improvement can expand the using scene of the existing hardware with higher quality data, this work will be valuable.

ACKNOWLEDGEMENTS

This research was funded by the National Natural Science Foundation of China (NSFC), grant number 41901407, The College Students' Innovative Entrepreneurial Training Plan Program, grant number S2020634016.

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.

REFERENCES

Birchfield, S. and Tomasi, C. Depth discontinuities by pixel-to-pixel stereo. *Int. J. Comput. Vis*. 1999, 35, 269–293.

Boehm, J., 2014. Accuracy investigation for structured-light based consumer 3D sensors. *Photogrammetrie-Fernerkundung-Geoinformation*, 2014(2), 117-127.

Brown, M. Z., Burschka, D. and Hager, G. D., 2003. Advances in computational stereo. *IEEE transactions on pattern analysis and machine intelligence*, 25(8), 993-1008.

Chen, H., Wang, K. and Yang, K., 2018. 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*, pp. 142-146.

Endres, F., Hess, J., Sturm, J., Cremers, D. and Burgard, W., 2013. 3-D mapping with an RGB-D camera. *IEEE transactions on robotics*, 30(1), 177-187.

Foix, S., Alenya, G. and Torras, C., 2011. Lock-in time-of-flight (ToF) cameras: A survey. *IEEE Sensors Journal*, 11(9), 1917-1926.

Hirschmuller, H., 2005. 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*. Volume 2, pp. 807–814.

Hirschmuller, H, 2006. 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*, pp. 2386–2393.

Hirschmuller, H. and Scharstein, D., 2008. Evaluation of stereo matching costs on images with radiometric differences. *IEEE transactions on pattern analysis and machine intelligence*, 31(9), 1582-1599.

Jiao, J., Yuan, L., Tang, W., Deng, Z. and Wu, Q., 2017. A post-rectification approach of depth images of Kinect v2 for 3D reconstruction of indoor scenes. *ISPRS International Journal of Geo-Information*, 6(11), 349.

Julio, R. O., Soares, L. B., Costa, E. A. C. and Bampi, S., 2015. Energy-efficient Gaussian filter for image processing using approximate adder circuits. In: *2015 IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, pp. 450-453.

Keselman, L., Iselin, Woodfill, J., Grunnet-Jepsen, A. and Bhowmik, A, 2017. Intel RealSense Stereoscopic Depth Cameras. In: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-10.

Kolmogorov, V. and Zabih, R, 2001. Computing visual correspondence with occlusions using graph cuts. In: *Proceedings the Eighth IEEE International Conference on Computer Vision*, Volume 2, pp. 508–515.

Kuan, Y. W., Ee, N. O. and Wei, L. S., 2019. Comparative study of intel R200, Kinect v2, and primesense RGB-D sensors performance outdoors. *IEEE Sensors Journal*, 19(19), 8741-8750.

Lu, J., Zhang, X., Dong, D. and Fang, Y., 2014. A stereo matching algorithm based on census transformation and dynamic programming. In: *Proceedings of the 33rd Chinese Control Conference*, pp. 8271-8276).

Rothermel, M., Wenzel, K., Fritsch, D. and Haala, N, 2012. SURE: Photogrammetric Surface Reconstruction from Imagery. In: *Proceedings of the LC3D Workshop*, Vol. 8, No. 2.

Scharstein, D. and Szeliski, R, 2002. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International journal of computer vision*, 47, 7–42.

Seki, A. and Pollefeys, M., 2017. SGM-nets: Semi-global matching with neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 231-240.

Sun, J., Zheng, N.N. and Shum, H.Y., 2003. Stereo matching using belief propagation. *IEEE Transactions on pattern analysis and machine intelligence*, 25(7), 787-800.

Veksler, O, 2005. Stereo correspondence by dynamic programming on a tree. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 2, pp. 384–390.

Yang, W., Li, X., Yang, B. and Fu, Y., 2020. A novel stereo matching algorithm for digital surface model (DSM) generation in water areas. *Remote Sensing*, 12(5), 870.

Zabih, R. and Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. In: *European conference on computer vision*, pp. 151-158.

Zhu, C. and Chang, Y. Z., 2019. Stereo matching for infrared images using guided filtering weighted by exponential moving average. *IET Image Processing*, 14(5), 830-837.