Hindawi
Journal of Electrical and Computer Engineering
Volume 2021, Article ID 9152035, 9 pages
https://doi.org/10.1155/2021/9152035

Research Article

Method of Using RealSense Camera to Estimate the Depth Map of Any Monocular Camera

Li-fen Tu and Qi Peng

School of Physics and Electronic Information Engineering, Hubei Engineering University, Xiaogan 432000, China

Correspondence should be addressed to Qi Peng; petersky0316@163.com

Received 10 April 2021; Revised 29 April 2021; Accepted 4 May 2021; Published 18 May 2021

Academic Editor: Yang Li

Robot detection, recognition, positioning, and other applications require not only real-time video image information but also the distance from the target to the camera, that is, depth information. This paper proposes a method to automatically generate any monocular camera depth map based on RealSense camera data. By using this method, any current single-camera detection system can be upgraded online. Without changing the original system, the depth information of the original monocular camera can be obtained simply, and the transition from 2D detection to 3D detection can be realized. In order to verify the effectiveness of the proposed method, a hardware system was constructed using the Micro-vision RS-A14K-GC8 industrial camera and the Intel RealSense D415 depth camera, and the depth map fitting algorithm proposed in this paper was used to test the system. The results show that, except for a few depth-missing areas, the results of other areas with depth are still good, which can basically describe the distance difference between the target and the camera. In addition, in order to verify the scalability of the method, a new hardware system was constructed with different cameras, and images were collected in a complex farmland environment. The generated depth map was good, which could basically describe the distance difference between the target and the camera.

1. Introduction

With the rapid development of technology in the robotics industry, the application fields are becoming more and more extensive, such as vegetable picking [1], industrial testing [2], medical assistance [3], automatic driving [4], etc., and epidemic prevention robot [5], which has been a hot topic in recent years. Visual information is the main channel for most robots to interact with the outside world. It usually relies on image sensors to collect video images and then analyzes the images through various algorithms to obtain information of interest, such as the location, shape, color, and category of the target. The image sensor usually acquires a two-dimensional image, which lacks the distance from the target to the camera, that is, depth information. Compared with RGB information, depth information introduces the distance from the target to the camera and adds a spatial dimension, which can better understand the scene and significantly improve the accuracy of robot detection, recognition, and positioning applications [6].

In general, the depth map acquisition methods are divided into two categories: active and passive. The most obvious feature of the active method is that the equipment itself needs to transmit energy to complete the collection of depth information. This ensures that the depth image is obtained independently of the color image. In recent years, active depth map acquisition methods mainly include TOF (Time of Flight) [7], structured light and Kinect [8, 9], lidar [10, 11], and so on. The principle of the TOF camera [7] to obtain a depth image is as follows: by transmitting continuous near-infrared pulses to the target scene, the light pulses reflected back from the object are received by the sensor. By comparing the phase difference between the emitted light pulse and the light pulse reflected by the object, the transmission delay between the light pulse can be calculated and the distance of the object relative to the emitter

can be obtained, resulting in a depth image. It has a larger infrared sensor size, a wider field of view angle, and a higher quality depth map. However, the resolution of the depth image is much lower than the resolution of the color image. The depth value is disturbed by significant noise, especially at the edges of the object. In addition, TOF cameras are usually expensive. The principle of depth image acquisition based on structured light [8] is that the structured light is projected to the scene, and the corresponding pattern with structured light is captured by the image sensor. Since the pattern of structured light will change due to the shape of the object, the depth information of each point in the scene can be obtained by calculating the position of the pattern image in the captured image and the degree of deformation by using the triangulation principle. This method can obtain the three-dimensional information of the target with high accuracy and speed. However, because the structured light method is easily affected by the strong natural light outdoors, it can not be used in an outdoor environment. Moreover, depth loss will occur when the object is black or the surface of the object is smooth. Kinect [9] adopts a technology called Light Coding. Different from traditional structured Light, Kinect's light-encoded infrared transmitter emits a "stereo code" with three-dimensional depth. This method can also obtain the three-dimensional information of the target with high accuracy and speed, but the effective range is small, the depth value is missing, and the edge of the depth image does not correspond to the edge of the color image and has some noise. The depth information acquisition principle of lidar [10, 11] is that laser is fired into space at a certain time interval, and the signal of each scanning point is recorded from the lidar to the objects in the measured scene, as well as the interval time between the signal reflected to the lidar after the object, so as to calculate the distance between the surface of the object and the lidar. Because of its wide ranging range and high measurement accuracy, lidar is widely used in artificial intelligence systems of outdoor three-dimensional space perception, such as obstacle avoidance navigation of autonomous vehicles, three-dimensional scene reconstruction, and other applications. However, its price is relatively high, and the texture information of the target is lacking.

Passive depth acquisition methods mainly include binocular or multiocular stereo matching [12–14] and monocular depth estimation [15]. The binocular or multiocular stereo matching method uses multiple cameras separated by a certain distance to obtain multiple images of the same scene at the same time. A stereo matching algorithm is used to find the corresponding pixels in multiple images, and then the disparity information is calculated according to the principle of triangulation. The disparity information can be transformed to represent the depth information of objects in the scene. This method has low hardware requirements and can be applied to both indoor and outdoor scenes. However, it has high computational complexity to carry out pixel-by-pixel stereo matching, and it is difficult to match in scenes lacking texture and scenes with severe lighting changes. Moreover, this method requires a complex calibration of the camera. Once the calibration is completed, the relative position of the camera cannot be moved and it is inflexible to use. Monocular image depth estimation [15] is a method that only relies on a single-view image or video data for depth estimation. Because the camera is projecting the three-dimensional space onto the image plane, it will inevitably cause the loss of depth information. Therefore, it has long been regarded as a pathological problem to recover depth information only through a single image, and it is difficult to achieve. However, in recent years, deep learning has developed rapidly. Convolutional Neural Network (CNN) [16] constantly refreshed records in various fields of computer vision with its efficient ability of image feature extraction and expression, which provided a new idea for the estimation of depth information of monocular images [17–20]. This method has the advantages of low hardware cost, flexibility in use, and high-precision in-depth map generation. However, learning and modeling need to be carried out first, so a large number of data sets and complex operation processes are required, and the universality is not strong, so it is not suitable for popularization.

In this paper, the characteristics of active and passive acquisition depth maps are integrated, and the idea of combining hardware and software is adopted to obtain the depth map of a monocular camera, which can improve the existing monocular camera. A RealSense camera is added to the installation location of the camera. The RGB image obtained by RealSense is matched with the original high-precision RGB image obtained by the single camera to acquire the spatial position correspondence of the points. Then, the depth map obtained by RealSense is mapped according to the position correspondence to fit the depth map of the original monocular camera. This method retains the performance of the original camera. The image resolution and field of view range remain unchanged, which overcomes the defect of obtaining the depth map only by hardware. In addition, the acquisition of depth map does not calculate 3D coordinates through multicamera image coordinates, so there is no need to calibrate the hardware, nor to learn and model the scene, and does not require a large amount of prior knowledge which is suitable for popularization and application.

2. Depth Map Generation Method for Monocular Camera

2.1. Hardware Structure. The hardware structure of the system is relatively simple. Any monocular camera and RealSense camera are fixed together. Generally, the two cameras are required to be closely connected horizontally or up and down. The purpose of this installation is based on the assumption that the same scene captured by the two cameras has the same depth, and the depth map of the RealSense camera is used to fit the depth map of a conventional monocular camera. For ordinary monocular cameras, there will be very little error in the generated absolute depth because the image planes of the two cameras do not exactly coincide, but the relative depth of different objects in the scene will not be affected.

According to this structural requirement, the designed hardware system is shown in Figure 1, where Figure 1(a) is

the system structure diagram, and Figure 1(b) is the system physical diagram.

In Figure 1(a): Camera ① is an arbitrary monocular camera. In this paper, a Micro-vision RS-A14K-GC8 industrial camera is adopted. Camera ② uses an Intel RealSense D415 depth camera. In addition, camera ② can also use the Intel RealSense D435 depth camera. The monocular camera ① is mounted on the quick mounting plate of the first pan-tilt ③. The base of the first pan-tilt ③ is fixed on the pan-tilt fixing plate ⑤. ⑥ is a transfer frame made of ductile material for cushioning against earthquakes. Camera ② is mounted on the quick mounting plate of the second pan-tilt ④. The base of the second pan-tilt ④ is fixed on the pan-tilt fixing plate ⑤. In addition, the monocular camera ① is equipped with heat sinks ⑦ in the top, bottom, left, and right directions. This is because industrial cameras have high power consumption and easily generate heat when in use, and they must be dissipated. In terms of structural features, the monocular ① and the RealSense camera ② need to be closely matched in the horizontal or vertical direction. In this specific embodiment, the monocular camera ① and the RealSense camera ② are closely matched in the horizontal direction. A tripod ⑧ is installed under the pan-tilt fixing plate ⑤. By adjusting the posture of tripod ⑧, the optical axis of monocular camera ① and the optical axis of RealSense camera ② can be kept horizontal.

The hardware system design can realize the system upgrade by adding a RealSense camera without replacing the existing single-camera system. The image information acquired by the original system is completely retained, and the depth map corresponding to the original image can be fitted.

2.2. Depth Map Fitting Algorithm. Suppose the RGB image collected by the original monocular camera is I_{RGB}. The RealSense camera captures two kinds of images. One is an RGB image, denoted as R_{RGB}, and the other is a depth map corresponding to the RGB image, denoted as R_D. The purpose of this algorithm is to obtain the corresponding relationship between the feature point positions by matching I_{RGB} and R_{RGB}, and then fit the depth map corresponding to the RGB images collected by the monocular camera with R_D, which is represented by I_D. A schematic diagram of the algorithm steps is shown in Figure 2.

Step 1. Sampling the RGB image captured by the original monocular camera and converting it to the same resolution as the image R_{RGB} captured by the RealSense camera. Due to the low resolution of RealSense cameras, for example, the two models used in this paper are D415 and D435, respectively, the maximum depth map resolution that can be output is 1280×720, while common cameras have high resolution. Since the resolution of RealSense cameras is relatively small, the two models used in this paper are D415 and D435, respectively, and the maximum resolution of the depth map that can be output is 1280×720 while the resolution of commonly used ordinary cameras is higher. Therefore, first of all, I_{RGB} should be downsampled to transform into a new image i_{RGB} with the same resolution as R_{RGB}. Assuming that the resolution of R_{RGB} is $x \times y$, then i_{RGB} can be obtained by equation (1), where cv2.resize is the image sampling function. For detailed usage, please refer to the OpenCV document [22]:

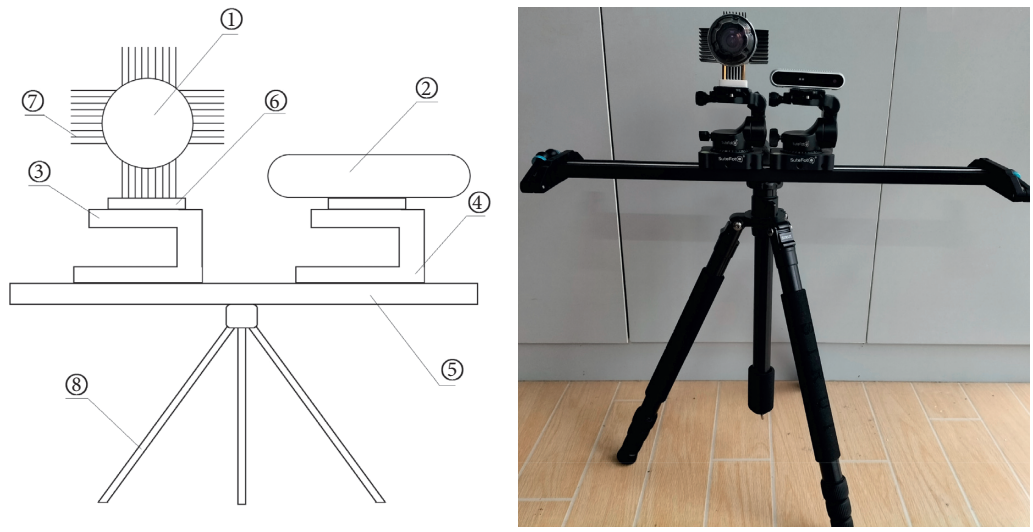
$$i_{\mathrm{RGB}} = \mathrm{cv2.resize}\left(I_{\mathrm{RGB}}, (x, y)\right). \tag{1}$$

At the same time, it is also necessary to perform superpixel segmentation [22] on i_{RGB}, and the scene to be analyzed should be divided into regions. The image after segmentation is represented by S_{RGB}.

Step 2. Match the feature points between i_{RGB} and R_{RGB}, eliminate the points with large errors, and retain the good matching points. Due to the big difference between the two camera models, the field of view angle will be different, and the image captured by the camera with the larger field angle will have more noncoincidence areas on both sides. However, since the cameras are closely connected, the image similarity of the overlapping part in the middle of the field of view is very high, which can reduce the difficulty of matching. Therefore, more matching point pairs are usually generated in this area

Suppose that the image coordinate of a pair of matching points in i_{RGB} is (m, n), and the coordinate in R_{RGB} is (m', n'). Due to the different models of the two cameras, these two coordinates are generally different. However, the two cameras are closely connected from left to right or from top to bottom, and the front and back positions are consistent. Therefore, for the same scene, the absolute depth is similar, while the relative depth is the same. Since the RGB image of the RealSense camera has a one-to-one correspondence with the points in the depth map, we use the depth value of the point at coordinate (m', n') in image R_D as the depth value at coordinate (m, n) in image i_{RGB}. According to this correspondence, the depth values of all matching point positions can be generated to form a new image i_{DP}. Use $i_{\mathrm{RGB}}(m, n)$ to represent the gray value of the image i_{RGB} at coordinate (m, n), $R_{\mathrm{RGB}}(m', n')$ to represent the gray value of image R_{RGB} at coordinate (m', n'), $i_{\mathrm{DP}}(m, n)$ to represent the gray value of image i_{DP} at coordinate (m, n), and $R_D(m', n')$ to represent the gray value of image R_D at coordinate (m', n'). Then the gray value of each point of image i_{DP} can be obtained by the following equation:

$$i_{\mathrm{DP}}(m, n) = \begin{cases} R_D(m', n'), & \text{if point } (m, n) \text{ in } i_{\mathrm{RGB}} \text{ matches point } (m', n') \text{ in } R_{\mathrm{RGB}}, \\ 0, & \text{else.} \end{cases} \tag{2}$$



(a) (b)

Figure 1: The hardware system. (a) The structure diagram. (b) The physical diagram.

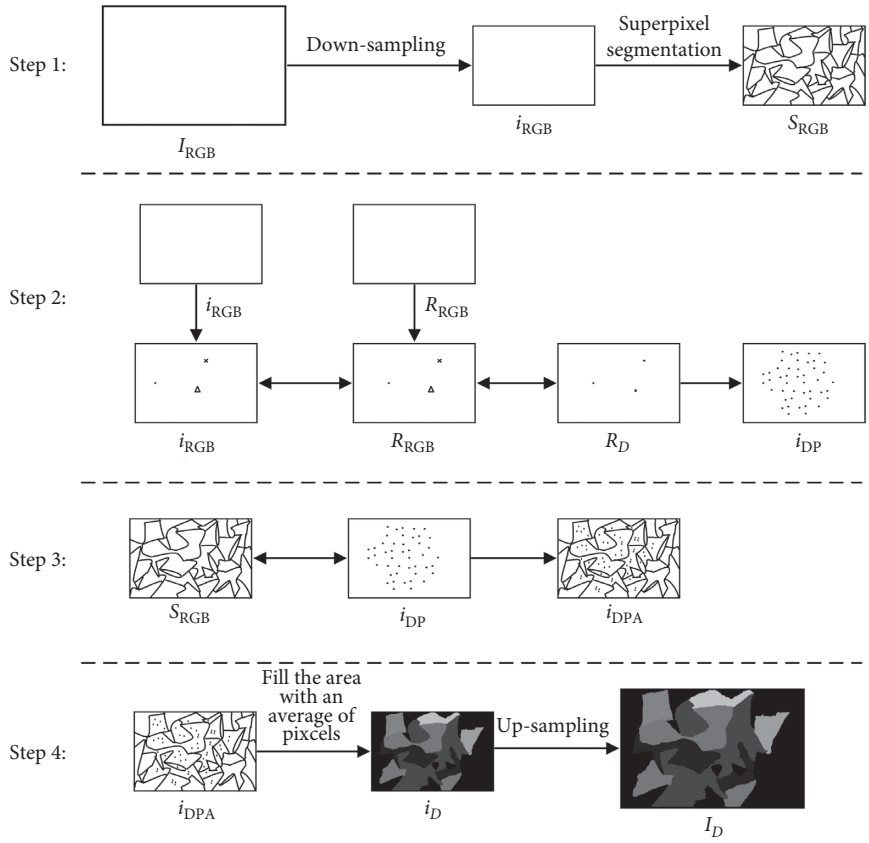


Figure 2: The algorithm flow chart.

In the schematic diagram of Figure 2, the background of i_{DP} is painted white in order to more clearly express the features of image i_{DP}. The actual algorithm is to set the gray value of the background area that has not been successfully matched to 0.

Step 3. Segment i_{DP} into regions, and S_{RGB}, the result of superpixel segmentation in the first step is needed. Superpixel segmentation composes adjacent pixels with similar texture, color, brightness, and other characteristics to form irregular pixel blocks with certain visual significance. These

(a) (b)

Figure 3: The camera system used in this experiment. (a) The front. (b) The side.

small blocks are taken as a unit and filled with the same depth value. Therefore, this step is to partition the depth values at the feature points of the hash distribution in i_{DP} according to the result of S_{RGB} and generate i_{DPA}. Assuming that there are N partitions in S_{RGB}, then i_{DPA} is to divide the i_{DP} image into the same N areas. Use $S_{\mathrm{RGB}}(s)$ to represent the s-th partition of S_{RGB}, use $i_{\mathrm{DPA}}(m,n)$ to represent the gray value of the image i_{DPA} at the coordinate (m,n), and use $i_{\mathrm{DPA}}(s)$ to represent the s-th partition of the image i_{DPA}, and then the image i_{DPA} can be described by the following equations:

$$i_{\mathrm{DPA}}(m,n) = i_{\mathrm{DP}}(m,n), \tag{3}$$

$$i_{\mathrm{DPA}}(s) = S_{\mathrm{RGB}}(s). \tag{4}$$

The schematic diagram i_{DPA} in Figure 2 is the partition result. The area dividing line is added to express the meaning of the algorithm more clearly. The actual algorithm does not have a dividing line.

Step 4. For each region in i_{DPA}, the average depth of all feature points is counted as the depth value of this region and the filling operation is carried out. The filling result is represented by i_D. Suppose there are M feature points in a certain region of i_{DPA}, and $i_{\mathrm{DPA}}(j)$ is used to represent the gray value of the j-th feature point. In this area, use $i_D(m,n)$ to represent the gray value of the point of the image i_D at the coordinate (m,n), then use equation (5) to calculate the gray value of all points in this area at the corresponding coordinates.

$$i_D(m,n) = \frac{1}{M}\sum_{j=1}^{M} i_{\mathrm{DPA}}(j). \tag{5}$$

Equation (5) is used in each region of i_{DPA} to calculate the average depth of each region and complete the region filling. The effect is shown as i_D in Figure 2. In actual operation, it is found that for scenes with few feature points, there will be some areas without feature points, so the background gray value 0 is used to fill them.

Finally, i_D is upsampled to fit the depth map I_D with the same resolution as the original image I_{RGB}. Assuming that the resolution of I_{RGB} is $x' \times y'$, then I_D is obtained by equation (6), where cv2.resize is the image sampling function. For detailed usage, please refer to the OpenCV document [21].

$$I_D = \mathrm{cv2.resize}\left(i_D, \ (x' \times y')\right). \tag{6}$$

Except for the areas where feature points are not detected and the areas outside RealSense cameras that may not be able to capture images in I_D, the depth values of other points correspond to I_{RGB} in a one-to-one correspondence.

3. Experiment and Result Analysis

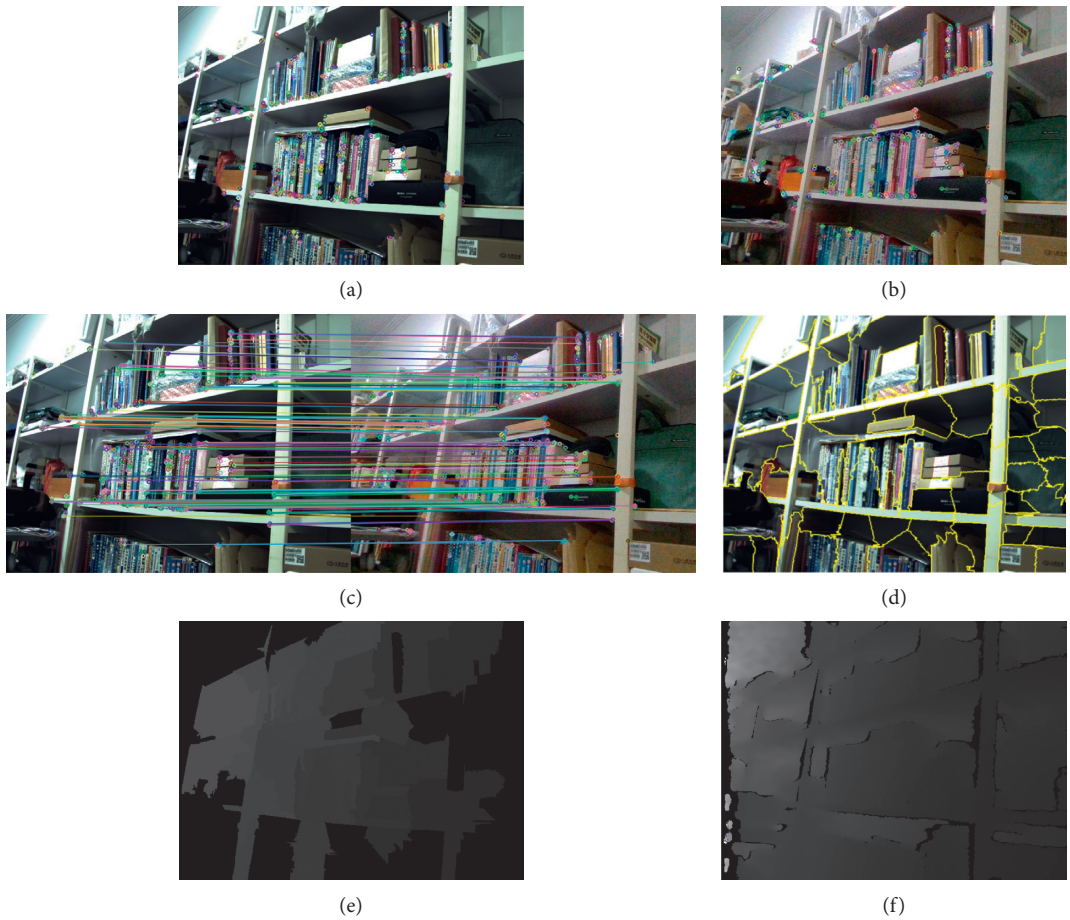
In order to verify the effectiveness of this method, a hardware system was constructed using the Micro-vision RS-A14K-GC8 industrial camera and the Intel RealSense D415 depth camera, as shown in Figure 3. Among them, the industrial camera consumes a lot of power and easily generates heat when in use, so heat sinks are installed in the four directions of up, down, left, and right. The size of the screw hole on the tripod pan-tilt quick release plate is 1/4 inch, but there are two M3 aperture screw holes under the industrial camera, which are not matched. Therefore, a 3D printer was used to make a conversion frame. The block printed by the 3D printer has good flexibility, good cushioning, and antivibration ability.

Here, the resolution of the industrial camera is 4384×3288. The resolution of the depth camera is varied. Due to the requirements of the algorithm, a resolution with an aspect ratio consistent with that of the industrial camera is selected, which is 640×480 here. An example of the image collected by this camera group is shown in Figure 4, where (a) is the large image I_{RGB} collected by the industrial camera, (b) is the RGB image R_{RGB} collected by the depth camera. Since the two are installed in close positions, the image scenes are very close. (c) is the depth map R_D collected by the depth camera, which coincides with the RGB map point-to-point.

Industrial applications have high requirements for image clarity and contrast, so the image quality of R_{RGB} cannot be used, only I_{RGB} can be used, but I_{RGB} lacks the corresponding depth map. Therefore, the purpose of this paper is

(a) (b) (c)

Figure 4: Examples of images captured by the camera group. (a) Industrial camera RGB image. (b) Depth camera RGB image. (c) Depth map of the depth camera.



(a) (b)

(c) (d)

(e) (f)

Figure 5: Algorithm test results. (a) The feature point detection result for I_{RGB}. (b) The feature point detection result for R_{RGB}. (c) The feature point matching result. (d) Superpixel segmentation results. (e) The depth map generated by the algorithm in this paper. (f) Depth map taken by depth camera.

to use the algorithm described in Figure 2 to generate a point-to-point depth map I_D with respect to I_{RGB}. The three images shown in Figure 4 are used to verify the algorithm, and some intermediate results and final results are shown in Figure 5, where (a) is the feature point detection result for I_{RGB}, (b) is the feature point detection result for R_{RGB}, (c) is the feature point matching result. Although R_{RGB} is affected by the resolution and the image quality is poor, it has little influence on the feature points. The detected feature points are basically the same as the high-resolution industrial camera, and the matching results are also good. (d) is the image after superpixel segmentation, which is S_{RGB}, (e) is the depth map I_D corresponding to the large image I_{RGB} of the industrial camera finally fitted by the algorithm, (f) is the depth map R_D corresponding to the small image I_{RGB} automatically generated by the depth camera. Since the two scenes are the same, they can be used for comparison with I_D.

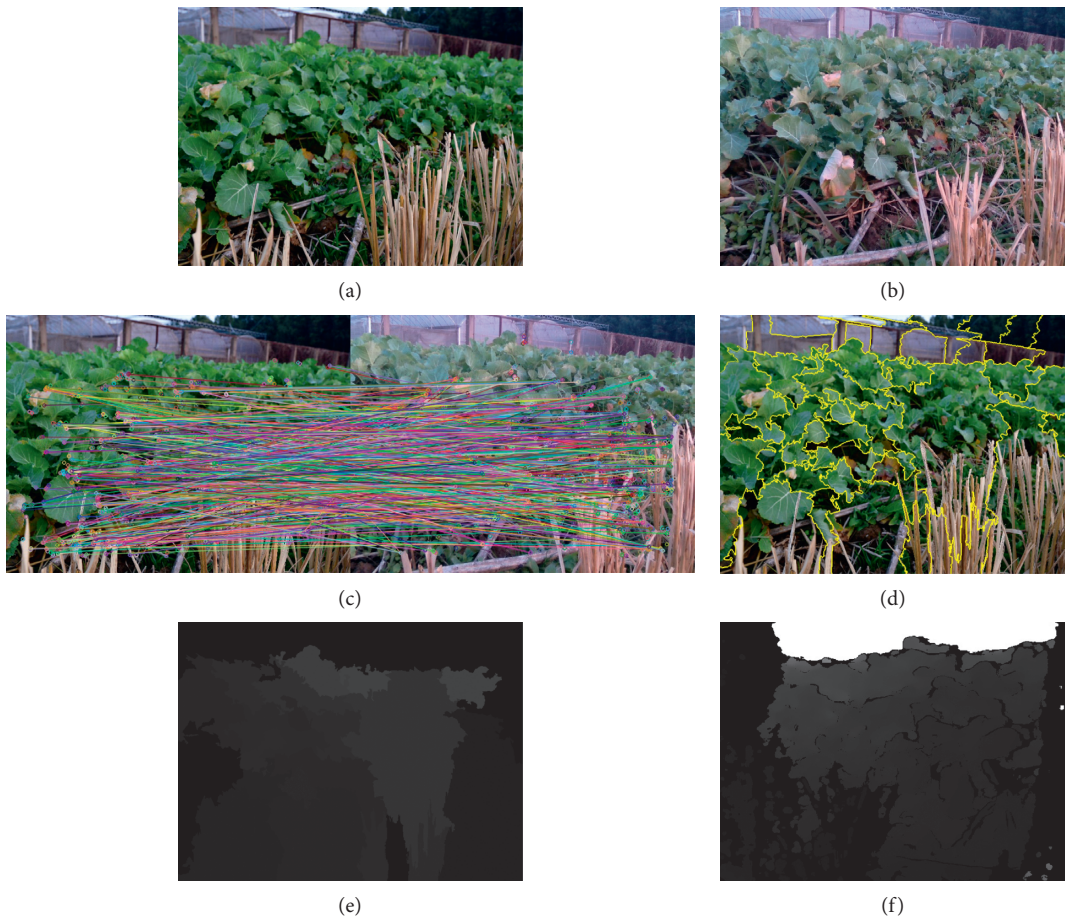


Figure 6: Algorithm test results. (a) The RGB image captured by Nikon D7500. (b) The RGB image captured by RealSense D415. (c) The feature point matching result. (d) Superpixel segmentation results. (e) The depth map generated by the algorithm in this paper. (f) Depth map taken by depth camera.

The black areas in Figures 5(e) and 5(f) are both depth-missing areas. By comparing the two figures, it can be found that the depth map (f) automatically generated by the depth camera is relatively complete, with only a small area of depth missing on both sides of the image, while the algorithm in this paper has more missing depth areas, which are mainly distributed around the image. There are two main reasons for the lack of depth in this algorithm: One is that although the industrial camera and the depth camera are closely connected to each other, they are still not completely consistent. The shooting picture is not completely overlapped, which will result in a lack of depth around them. The second is that some areas are relatively smooth and lack feature points, which may occur in the middle and around the image. By analyzing the results in Figure 5(e), except for a few depth-missing areas, the results of other depth areas are relatively good, which can basically describe the distance difference between the target and the camera.

The main innovation of the method of obtaining the depth map of the monocular camera proposed in this paper is that a RealSense camera is directly installed close to the original monocular camera without replacing the original hardware system, and the depth map of the monocular camera can be fitted according to the position mapping relationship between the two cameras, while the high resolution of the original camera can be retained. In order to verify the scalability of this method, the following two additional groups of experiments were conducted for verification.

The first group of experiments: the industrial camera used in this system was replaced with Nikon D7500, a common hand-held SLR camera, and the RealSense camera still used D415. The scene is replaced by an outdoor farmland environment, which has problems such as texture similarity and occlusion. It is difficult for traditional binocular stereo vision to generate a dense depth map through algorithm. Using the method designed in this paper, the results are shown in Figure 6:

The second group of experiments: the industrial camera used in this system was replaced with an ordinary mobile phone, and the RealSense camera was replaced with model D435. The scene is also a complex outdoor farmland environment. The method designed in this paper is adopted, and the result is shown in Figure 7:

The results in Figures 6 and 7 show that the method used to obtain the depth map in this paper has high scalability.

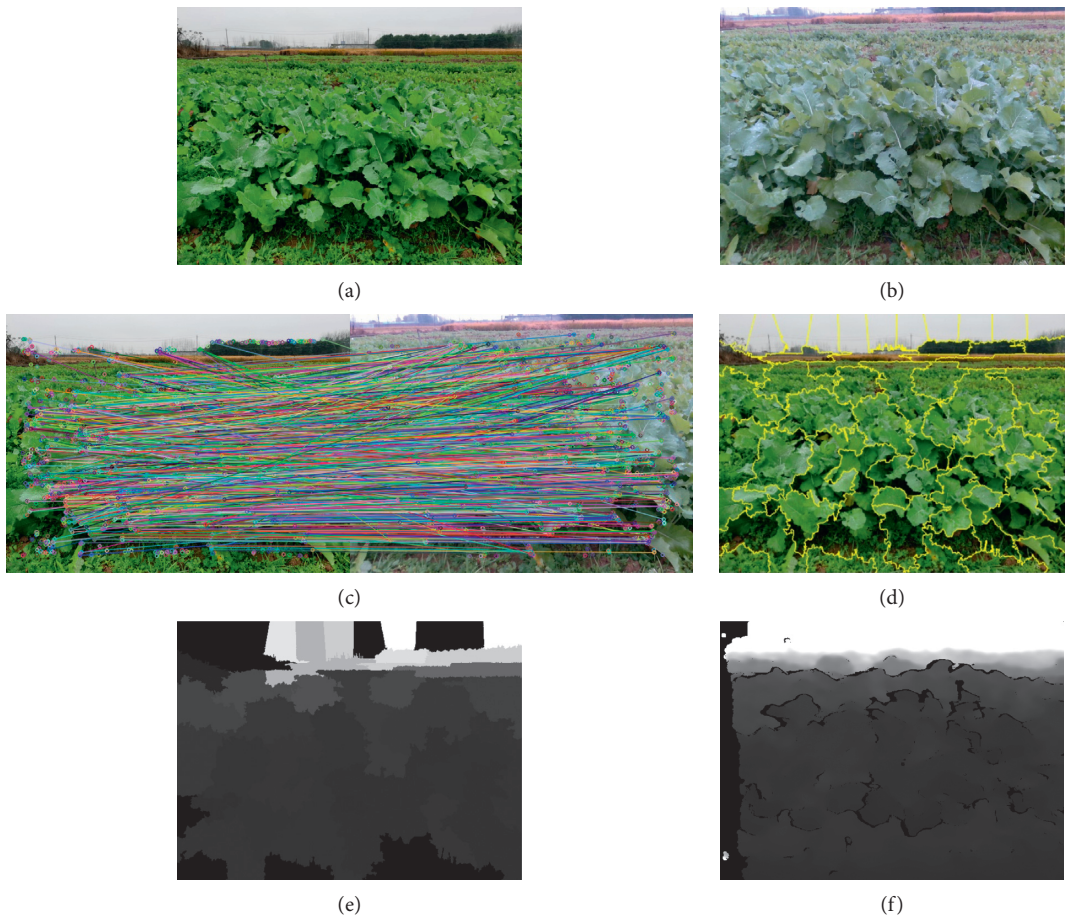


Figure 7: Algorithm test results. (a) The RGB image captured by mobile phone. (b) The RGB image captured by RealSense D435. (c) The feature point matching result. (d) Superpixel segmentation results. (e) The depth map generated by the algorithm in this paper. (f) Depth map taken by depth camera.

Industrial cameras, SLR cameras, mobile phones, and other photographic devices can be used to obtain high-resolution RGB images. Both the RealSense D415 and D435 models can be used to assist in generating the depth maps corresponding to the RGB images of the above monocular cameras. It can be used in various indoor or outdoor scenes. In fact, the method proposed in this paper aims to directly transform the existing monocular camera system with the help of simple hardware equipment at a lower cost so that it has the function of depth information acquisition, and adds a spatial dimension to the traditional monocular image detection system, so as to better understand the scene.

4. Conclusion

This paper provides a depth map fitting method for the existing monocular image or video detection system by combining the ideas of active and passive depth map acquisition methods. By using this method, any single-camera detection system can be upgraded online. Under the premise of not changing the original system, the depth information of the original single camera can be obtained by adding low-cost hardware and combining it with a simpler algorithm so as to realize the effective utilization of resources. Different hardware is used to test the method in different environments separately, and the experiment proves that the method has good validity and scalability. Compared with the existing depth information acquisition methods, the characteristics and advantages of this method are as follows:

(1) Compared with the traditional active method, although this method also requires additional hardware equipment, the cost is lower. More importantly, the system retains the performance of the original monocular camera, so the RGB images obtained by this system have a higher resolution.

(2) Compared with the traditional passive method, this method has lower requirements on the location of the hardware, as long as the two cameras are closely connected to each other and no complicated camera calibration process is needed. In addition, due to the assistance of the depth map of the depth camera, this paper only needs a relatively simple algorithm to restore the point-to-point depth map corresponding to the original monocular camera.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the Key Project of Science and Technology Research Plan of Hubei Provincial Department of Education (D20192701) and the Hubei Provincial Education Research Project (T201716).

References

[1] W. K. Jia, Y. Zhang, J. Lian, Y. J. Zheng, D. A. Zhao, and C. J. Li, "Apple harvesting robot under information technology: a review," *International Journal of Advanced Robotic Systems*, vol. 17, no. 3, 2020.

[2] Y. F. Hong and J. P. Tan, "Trajectory planning of a planar cable-driven robot for industrial detection," *Journal of Physics Conference Series*, vol. 1570, no. 1, 2020.

[3] T. Ginoya, Y. Maddahi, and K. Zareinia, "A historical review of medical robotic platforms," *Journal of Robotics*, vol. 2021, Article ID 6640031, 13 pages, 2021.

[4] Y. F. Cai, K. S. Tai, H. Wang, Y. C. Li, and L. Chen, "Research on behavior recognition algorithm of surrounding vehicles for driverless car," *Qiche Gongcheng/Automotive Engineering*, vol. 42, no. 11, pp. 1464–1472, 2020.

[5] C. Guo, M. Y. Wang, K. F. Gao, J. N. Liu, and W. W. Zuo, "Location-based service technologies for major public health events: illustrated by the cases of COVID-19 epidemic," *Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University*, vol. 46, no. 2, pp. 150–158, 2021.

[6] F. Hsiao and P. Lee, "Autonomous indoor passageway finding using 3D scene Reconstruction with stereo vision," *Journal of Aeronautics, Astronautics and Aviation*, vol. 52, no. 4, pp. 361–370, 2020.

[7] H. J. Issaq, T. D. Veenstra, T. P. Conrads, and D. Felschow, "The seldi-tof ms approach to proteomics: protein profiling and biomarker identification," *Biochemical and Biophysical Research Communications*, vol. 292, no. 3, pp. 587–592, 2002.

[8] H. Nguyen, Y. Z. Wang, and Z. Y. Wang, "Single-shot 3d shape reconstruction using structured light and deep convolutional neural networks," *Sensors (Switzerland)*, vol. 20, no. 13, pp. 1–13, 2020.

[9] M. Tolgyessy, M. Dekan, L. Chovanec, and P. Hubinsky, "Evaluation of the azure Kinect and its comparison to Kinect V1 and Kinect V2," *Sensors*, vol. 21, no. 2, pp. 1–25, 2021.

[10] L. Chen, Y. He, J. Chen, Q. Li, and Q. Zou, "Transforming a 3-D LiDAR point cloud into a 2-D dense depth map through a parameter self-adaptive framework," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, pp. 165–176, 2017.

[11] G. Q. Chen, Z. Z. Mao, H. L. Yi et al., "Pedestrian detection based on panoramic depth map transformed from 3d-lidar data," *Periodica Polytechnica Electrical Engineering and Computer Science*, vol. 64, no. 3, pp. 274–285, 2020.

[12] W.-P. Ma, W.-X. Li, and P.-X. Cao, "Binocular vision object positioning method for robots based on coarse-fine stereo matching," *International Journal of Automation and Computing*, vol. 17, no. 4, pp. 562–571, 2020.

[13] M. Yao, W. B. Ouyang, and B. G. Xu, "Hybrid cost aggregation for dense stereo matching," *Multimedia Tools and Applications*, vol. 79, no. 31-32, pp. 23189–23202, 2020.

[14] P. Rogister, R. Benosman, S. H. Ieng, P. Lichtsteiner, and T. Delbruck, "Asynchronous event-based binocular stereo matching," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 2, pp. 347–353, 2012.

[15] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: a review," *Neurocomputing*, vol. 438, pp. 14–33, 2021.

[16] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015)*, pp. 2650–2658, Santiago, Chile, February 2015.

[17] A. Gordon, H. H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: unsupervised monocular depth learning from unknown cameras," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2019)*, pp. 8976–8985, Seoul, South Korea, October 2019.

[18] H. Fu, M. M. Gong, C. H. Wang, K. Batmanghelich, and D. C. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pp. 2002–2011, Salt Lake City, USA, June 2018.

[19] Y. R. Chen, H. T. Zhao, Z. W. Hu, and J. C. Peng, "Attention-based context aggregation network for monocular depth estimation," *International Journal of Machine Learning and Cybernetics*, vol. 11, pp. 1–14, 2021.

[20] R. D. Mendes, E. G. Ribeiro, N. D. Rosa, and V. Grassi, "On deep learning techniques to boost monocular depth estimation for autonomous navigation," *Robotics and Autonomous Systems*, vol. 136, Article ID 103701, 2021.

[21] G. R. Bradski and A. Kaehler A, *Learning OpenCV*, Oreilly Media, Newton, MA, USA, 2018.

[22] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.