Proceeding of the IEEE
International Conference on Information and Automation
Wuyi Mountain, China, August 2018

A Joint Calibration Method for the 3D Sensing System Composed with ToF and Stereo Camera

Shulan Huang[1,2], Feifei Gu[1], Zhiquan Cheng[1] and Zhan Song[1,3*]

[1]*Shenzhen Key Laboratory of Minimally Invasive Surgical Robotics and System, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China*

[2]*Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China*

[3]*The Chinese University of Hong Kong, Hong Kong, China*

{sl.huang1; ff.gu; zq.cheng; zhan.song}@ siat.ac.cn

***Abstract* – Time of Flight (ToF) and stereo vision are the two major technologies for rapid depth information acquisition. The ToF sensor is able to obtain depth map directly at a high rate of speed, but the resolution and accuracy is relatively low. The binocular stereo vision means can obtain high-accuracy disparity map based on triangulation principle, but its performance depends on the textures of the target surface. Therefore it is an interesting research topic to combine above two techniques. In order to carry out the fusion of the two 3D techniques, an efficient calibration method is investigated in this work. Based on a simple calibration rig, point correspondences between the depth image and color image are extracted and then used to calculate both intrinsic and extrinsic parameters of two devices. The experimental result shows that the proposed method could achieve better accuracy in comparison with classical calibration methods.**

Index Terms - ToF, stereo vision, camera calibration, 3D reconstruction.

I. INTRODUCTION

Vision-based 3D sensing techniques have gained wide applications in industry inspection, robotics, virtual reality and unmanned vehicles etc. Existing 3D sensing methods can be generally classified into stereo vision, Time of Flight (ToF) [1], structured light scanning and shape from motion etc. ToF-based 3D sensing techniques have the advantages of fast scanning speed and dense point cloud etc. For example, the famous Kinect V2 produced by Microsoft, consists of depth camera with a resolution of 512×424 pixels, a color camera with a resolution of 1920×1080 pixels, 4 microphone arrays and some electronic devices used for signal processing. The depth camera here can capture images at a highest frequency of 30 Hz with a measurable depth range of 0.5-4.5 m, a horizontal viewing angle of 70° and a vertical viewing angle of 60° [2]. The Kinect V2 emits lights with a time-varying intensity sine wave signal, then derives the object's depth from the phase difference between the emitted and received signal correspondingly, which is commonly known as continuous wave (Continuous-Wave, CW) modulation [3]. Compared with the binocular stereo vision technology, it could work well without strict requirement about rich texture of to-be-measured object. However, its measurement accuracy and depth resolution are lower than the stereo vision method, which is based on the triangulation principle and could achieve higher measurement accuracy.

How to combine the advantages of ToF technique and stereo vision to obtain real-time and high accuracy 3D reconstruction has become one important research topic in the domain of 3D vision. To realize this purpose, the first step is to calibrate the ToF sensor and stereo vision systems. There have been a lot of literatures to address the calibration of ToF sensor or the binocular stereo vision systems [4-7, 11].

There have been some works to address the calibration of ToF-stereo rig. Zhu et al. [8] established a look-up table from three-dimensional coordinates of ToF depth camera to two-dimensional coordinates of color camera by shooting a large number of checkerboard images, which avoided directly seeking external parameters, but the amount of data to be handled is very large. While Gudmundsson et al. [9] calibrated all cameras using Bouguet's Matlab calibration toolbox[1] based on the fact that the ToF-images (176×144) were cropped to 160×120 and each left and right image (1280×960) from the stereo rig was eight times down-scaled, so that they had the same size. In this way, the ToF camera was stereo-calibrated with the stereo rig, in which the left and right were then calibrated with each other in full size (1280×960). However, this method may bring down-sampling error. Dal Mutto et al. [12] estimate the intrinsic parameters with the standard calibration algorithms [7], and then minimized the sum of the Euclidean distance errors between all corresponding points from both ToF and stereo images. The limitation is that they apply a standard corner detector on the amplitude images acquired by ToF camera to obtain the coordinates of corners. Zhu et al. [10] used a checkerboard that is movable on a metric board with two guide rails, and generated per-pixel Look-Up-Tables (LUTs) to compensate the depth bias. They acquired raw parameters by calibrating color images from stereo and intensity images from ToF sensor, and then computed the per-pixel depth bias by comparing depth from the ToF sensor and that from stereo cameras in the world coordinate system. But the LUTs only stored a set of discrete value of depth bias.

This work was supported in part by the National Key R&D Program of China (2017YFB1103602), Shenzhen Science Plan (KQJSCX20170731165108047, JCYJ20170818160448602, JCYJ20170413152535587), and the National Natural Science Foundation of China (61773363, 51705513, U1613213, U1713213).

978-1-5386-8069-8/18//$31.00 ©2018 IEEE

Traditional checkerboard-based calibration schemes fail to work well for the depth camera, since its corner features cannot be reliably detected in the depth image. In [15, 16], some spheres are used to calibrate the ToF and stereo cameras. In [17], square or circular holes are drilled on the calibration plane, and these features are used for the system calibration. In some methods, the infrared images that obtained by the Kinect sensor are also utilized for the calibration of the camera pair as described in [18]. In [19], the authors summarized a large number of calibration methods of the RGB-D cameras, and compared the three most used calibration methods that have been applied to three different RGB-D sensors based on structured light and ToF. However, a good method of calibrating Kinect V2 is not given except using infrared (IR) images to calibrate the depth sensor.

In this paper we proposed a simple, efficient and reliable calibration method suitable for our heterogeneous system, which is composed by a Kinect V2 and two color cameras. Plenty of features and constrains in depth images are available easily with this new method, which improves the calibration accuracy. Besides, point correspondences between the depth image and color image from any standard camera are easy to get, making it possible for the depth sensor and color camera to be calibrated and optimized directly and jointly. Both the interior parameters and exterior parameters could be obtained at the same time, which greatly reduces the workload.

II. Joint Calibration of ToF and Binocular Cameras

A. Calibration Model of ToF Sensor - Kinect V2

The Kinect V2 is equipped with a 1080p full HD wide-angle color camera. Beyond that, there is an infrared camera adapting to low-light environment and three infrared enhanced launchers. The internal structure and important parts of Kinect V2 are shown in Fig. 1 with two enlarged views of the RGB camera and the depth/IR camera.

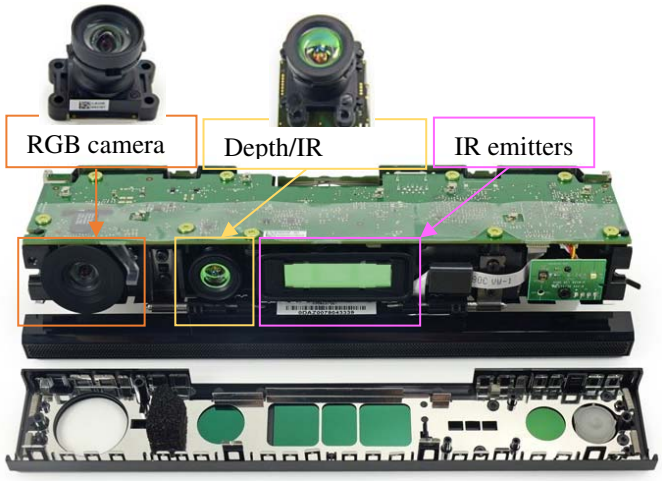

Fig. 1. Internal structure of ToF sensor - Kinect V2.

For the purpose of 3D reconstruction, it is necessary to calibrate the intrinsic and extrinsic parameters of the cameras. In general, the manufacturers only provide theoretical values, which will possibly be biased in practice in the process of production and transportation, even for the same type of cameras. Besides, to enlarge the field of view of the depth sensor, a CMOS focusing lens is used in the Kinect V2 as the depth camera [13] whose optical model is pinhole imaging. Radial distortion and tangential distortion of camera lens have a greater impact on the projection images compared to other distortion factors, whose mathematical models are summarized as follows.

$$\begin{cases} u_r = u\left(1+k_1r^2+k_2r^4+k_3r^6\right) \\ v_r = v\left(1+k_1r^2+k_2r^4+k_3r^6\right) \end{cases} \tag{1}$$

$$\begin{cases} u_t = u+\left[2p_1v+p_2\left(r^2+2u^2\right)\right] \\ v_t = v+\left[p_1\left(r^2+2v^2\right)+2p_2u\right] \end{cases} \tag{2}$$

Equation (1) and (2) represent the radial distortion and tangential distortion of lens respectivley.

For a 3D point denoted by capital X in the world coordinate system, its projection point is denoted by x on the image plane. Therefore, the relation varying from a 3D point to a 2D point is a linear transformation of the projection space. The complete transformation from point X to x in homogeneous coordinates can be expressed as:

$$\mathrm{x} = K\left[R \mid T\right]\mathrm{X} = P\mathrm{X} \tag{3}$$

where K represents the internal parameters of camera, R and T denote the external parameters between two coordinate system, with six degrees of freedom. P represents a 3×4 perspective projection matrix, which is also referred to as a camera matrix because it includes all internal and external parameters of the camera. Because checkerboard is employed for calibration, (3) can be denoted as (4).

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = K\begin{bmatrix} r_1 & r_2 & T \end{bmatrix}\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{4}$$

In order to calculate these arguments above, we define a new parameter H as (5), which is called a homography matrix and has 8 degrees of freedom. The n pairs of 2D-3D correspondence points provide $2n$ linear equations for H, which can be linearly resolved by Direct Linear Transformation (DLT) if $n \geq 4$. Then K can be solved from H.

$$H = K\begin{bmatrix} r_1 & r_2 & T \end{bmatrix} = \begin{bmatrix} h_1 & h_2 & h_3 \end{bmatrix} \tag{5}$$

As we know, the column vectors of R matrix are orthogonal each other, so we may deduce the following formula.

$$\begin{cases} h_1^T K^{-T} K^{-1} h_2 = 0 \\ h_1^T K^{-T} K^{-1} h_1 = h_2^T K^{-T} K^{-1} h_2 \end{cases} \tag{6}$$

For each image, two constraint equations of internal parameter can be obtained. For the parameter K (upper triangular matrix), $\omega = K^{-T}K^{-1}$ is a symmetric matrix. If the number of images is more than three, can be solved linearly by Direct Linear Transformation (DLT) algorithm and then K is obtained by orthogonal decomposition. Furthermore, r_1, r_2 and T can be derived from (5) consequently while r_3 is calculated from (7).

$$r_3 = r_1 \times r_2 \tag{7}$$

At this point, we come to the camera parameters K, R, T, which are used as the initial value called minimal solution in algebraic error. The final goal is to get the iterative solution of the following reprojection error for this minimization problem.

$$\min \sum_i \left\| P\mathrm{X}_i - \mathrm{x}_i \right\|^2 \tag{8}$$

In the end, the minimums of K, R, T are obtained under geometric error.

B. Calibration of the Heterogeneous System

The heterogeneous system consists of a Kinect V2 and two Canon EOS 700D cameras. In addition, due to the difference of storage mode and imaging, the images obtained by Kinect V2 are horizontally reversed with respect to SLR images. What is more, depth sensor of Kinect V2 provides depth maps representing depth information without textures, while SLR camera supplies color images describing texture information. These facts above inspire us to attach textures to depth maps.

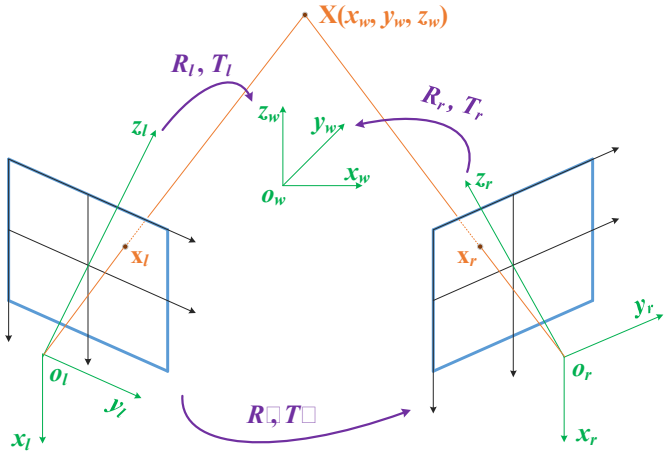

Fig. 2. Imaging model of the heterogeneous system.

In order to calibrate the heterogeneous system, we calibrate the depth sensor with one SLR camera at first, and then calibrate the two SLR cameras. The imaging model of two cameras is shown in Fig. 2. The exterior orientation arguments of the left and right camera are denoted as R_l, T_l, R_r, T_r. The relationship among them is expressed as (9).

$$\begin{cases} R' = R_r R_l^{-1} \\ T' = T_r - R_r R_l^{-1} T_l \end{cases} \tag{9}$$

From (6)-(9), both intrinsic and extrinsic parameters of any two cameras in the heterogeneous system are worked out. For three cameras, the rotation and translation from the first camera to the second camera are R_1, T_1, and the pose conversion from the second to the third camera is R_2, T_2 correspondingly. It can be deduced the transformation R_3, T_3 from the first to the third camera by (10).

$$\begin{cases} R_3 = R_2 \cdot R_1 \\ T_3 = R_2 \cdot T_1 + T_2 \end{cases} \tag{10}$$

The main calibration procedure of the heterogeneous system is shown in Fig. 3. The resolution of images provided by ToF sensor is quite low, and its imaging method is different from the standard camera or infrared camera. For these reasons, it is more difficult to calibrate than a common camera, especially to determine the pose transformation between the ToF sensor and the standard camera. The method of planar template is not able to obtain the desired result when it is used in primitive depth images.

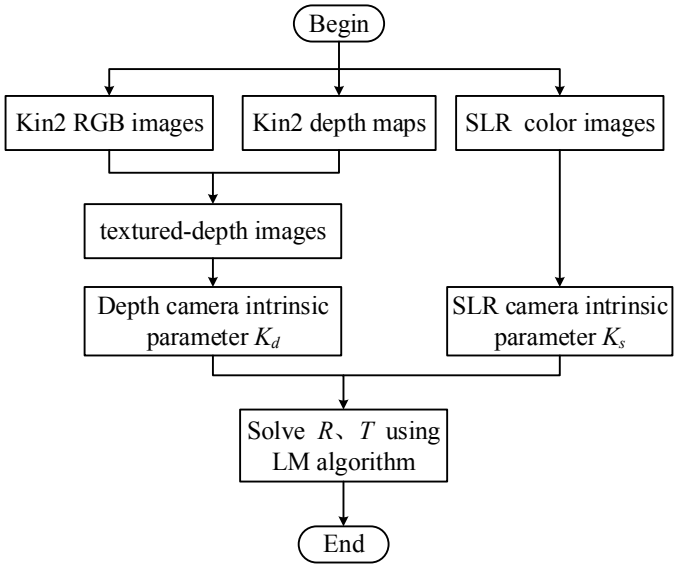

Fig. 3. Calibration between Kinect V2 and SLR camera.

The texture of the color image was mapped to the depth image and then a textured-depth image can be obtained. Since the resolution of images from RGB camera inside Kinect V2 is higher than that from depth sensor, the mapping texture is of rather good visual effect. After that, these images and color images from SLR are calibrated together.

TABLE I
INTERIOR PARAMETERS OF CAMERAS

cameras	data sources	f_x	f_y	u_0	v_0
depth/IR camera of Kinect V2	factory parameters	366.4652	366.4652	256.6184	203.7805
	error（+\-）	---	---	---	---
	Kim's method [18]	370.2049	370.7317	258.7926	199.6079
	error（+\-）	1.8688	1.9245	2.3251	2.1344
	our method	**365.4426**	**365.7826**	**260.8543**	**200.3256**
	error（+\-）	1.0367	1.0601	0.6883	0.6288
RGB Camera of Kinect V2	Terven's method [14]	1079.8578	1079.8578	976.8504	539.9836
	error（+\-）	---	---	---	---
	Bouguet's toolbox	1081.3867	1080.9698	978.9494	534.3203
	error（+\-）	2.2534	2.2782	4.2413	3.1812
SLR camera	Bouguet's toolbox	4776.9131	4780.3083	2686.8538	1662.6250
	error（+\-）	3.9464	4.0947	2.9283	1.9985

TABLE □
EXTERIOR PARAMETERS BETWEEN TWO CAMERAS

methods	Rotation from Kinect to SLR			Translation from Kinect to SLR
Kim's method [18]	0.99607734	0.0001145	-0.088487	147.6442081
	-0.00227324	0.9997022	-0.024295	-4.620772244
	0.08845768	0.0244012	0.995781	-23.60332734
Terven's method [14]	0.99618029	-0.0063368	-0.08703	143.5650863
	0.00393127	0.9996084	-0.027784	-4.548764284
	0.08717148	0.0273360	0.9958158	-22.12899615
our method	**0.99674377**	**0.0002527**	**-0.080634**	**147.231339**
	-0.00230984	**0.9996741**	**-0.02542**	**-4.567080897**
	0.08060095	**0.0255234**	**0.9964196**	**-20.35475035**

III. EXPERIMENTAL RESULTS

A. Experimental Setup

The key function of mapping textures of the RGB camera to the primal depth maps is *MapDepthFrameToColorSpace()*, which is encapsulated in Microsoft Software Development Kit (SDK). For convenience, a program of GUI interface was written before experiment to capture multiple images of multiple cameras at the same time.

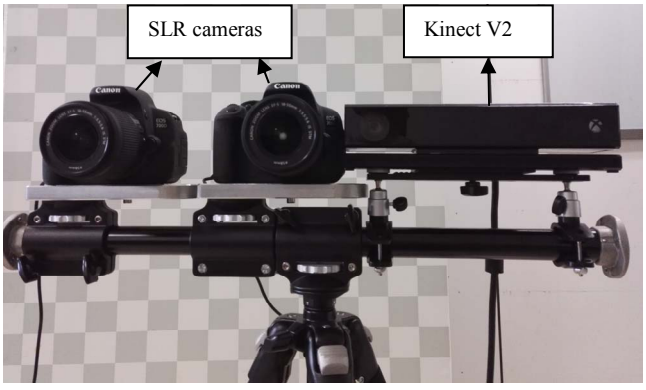

Fig. 4. The experimental setup consists of two DSLR and a ToF camera.

In our experiment, two SLR cameras are placed on the same side of the Kinect V2. The pattern of checkerboard consists of 12 lines and 15 columns, and the size of each square is 20×20 mm. The Kinect V2 and the SLR camera should be placed as close as possible so that the overlap of them is as large as possible. The experimental setup can be seen in Fig. 4.

B. Results of Calibration

The interior parameters of cameras involved in our experiment are shown in the TABLE I, from which we can see that the calibration accuracy of depth/IR camera with textured-depth maps is higher than that with IR images. The main reason is that textured-depth maps could increase the accuracy of corners detection. The calibration result of the depth camera shows that the mean reprojection error is only 0.152 pixel using the proposed method. In our contrast experiment, we also take advantage of the Kinect V2 toolbox [14] to get the interior parameters of depth camera and the exterior parameters between the two cameras inside the Kinect V2. Basing on this toolbox, we make use of RGB camera as a bridge to calibrate depth and SLR camera in an indirect way. The results of three methods to calibrate are displayed in TABLE II.

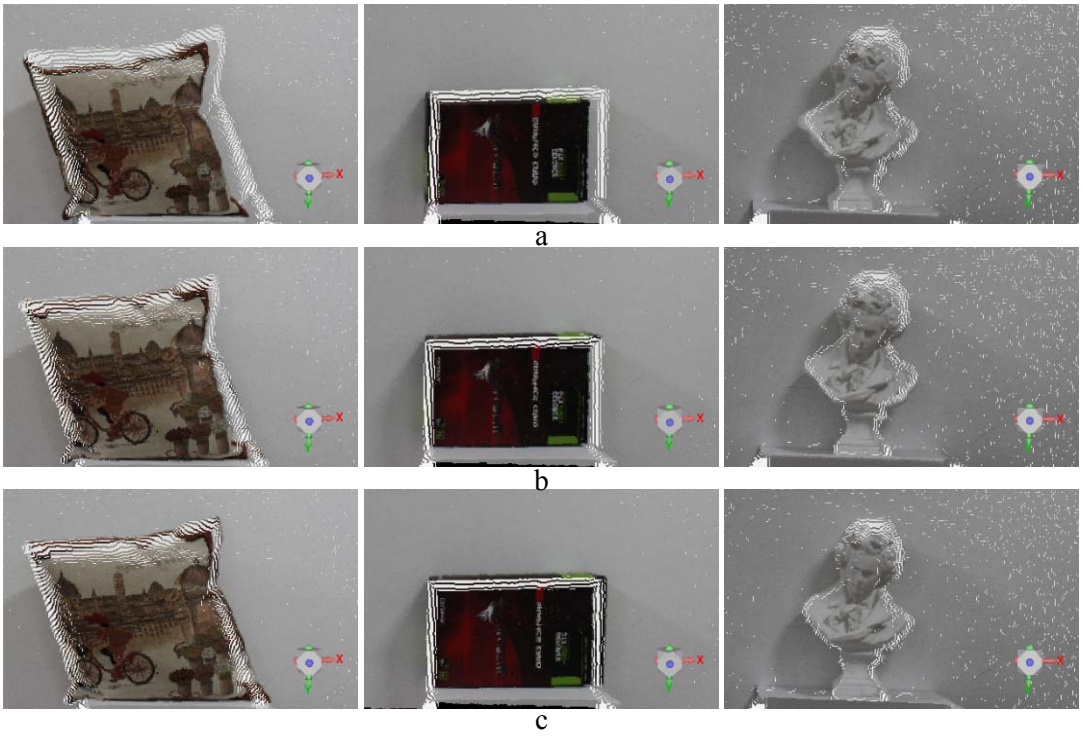

Fig. 5. Mapping of the high quality texture map from SLR camera to the ToF depth images. a) texture mapping results by the Kim's method [18]; b) texture mapping results by the Terven's method [14]; c) texture mapping results by our method.

On the basis of these parameters, textures derived from the SLR camera are aligned to the point clouds acquired by the ToF sensor and the results are exhibited in Fig. 5. From the results, we can see that our method gets the most accurate alignment in contrast to other two methods. To further validate our approach, we attach the textures of SLR camera to the depth maps employing the calibration results in TABLE II and then calculate the reprojection errors of checkerboard corners as shown in Fig. 6. From the result, we can see that the proposed method has minimum reprojection error. The worst result is obtained by Terven's method, the reason for which due to its dependence on Kinect 2 toolbox. It reads the internal parameters of the depth camera from the factory settings that deviate from actual values. Besides, the external parameters between the two cameras inside Kinect V2 are acquired by computing a less reasonable cost function. Compared with IR images, the textured-depth images are more stable and reliable for calibration.

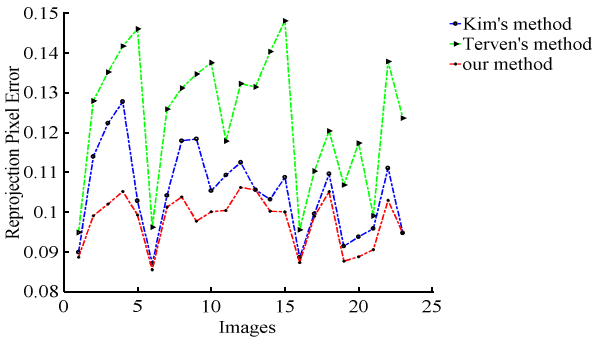

Fig. 6. Comparison of reprojection errors by different methods.

IV. Conclusion and Future Work

As one of the most typical representatives of ToF cameras, Kinect V2 handles depth accuracy and phase ambiguity well. To the best of our knowledge, there is few approach and research report that handles the problem of calibrating ToF sensor with cameras that are totally different from each other. We introduced to attach color textures to the original depth maps according to the characteristics of Kinect V2 itself, in this way corners could be detected for calibration. Experimental results show that out method is simple but effective and it could reduce computational complexity and shorten calibration process. In this paper, we presume the ToF sensor is free from systematic error. In future, we will establish an error-corrected model to compensate it and filter the collected data before being used, so as to increase the measurement reliability of ToF sensor. What is more, basing on the calibration method presented here, we will use the depth data acquired by Kinect V2 for the guidance of binocular stereo matching to realize rapid, high-quality and high-precision 3D reconstruction.

References

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

[2] Pagliari D, Pinto L. Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors. Sensors, 2015, 15(11): 27569-27589.

[3] Corti A, Giancola S, Mainetti G, et al. A metrological characterization of the Kinect V2 time-of-flight camera. Robotics and Autonomous Systems, 2016, 75: 584-594.

[4] Fuchs S, Hirzinger G. Extrinsic and depth calibration of ToF-cameras. IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1-6.
[5] Lindner M, Schiller I, Kolb A, et al. Time-of-flight sensor calibration for accurate range sensing. Computer Vision and Image Understanding, 2010, 114(12): 1318-1328.
[6] Horaud R, Csurka G, Demirdijian D. Stereo calibration from rigid motions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(12): 1446-1452.
[7] Zhang Z. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1330-1334.
[8] Zhu J, Wang L, Yang R, et al. Fusion of time-of-flight depth and stereo for high accuracy depth maps. IEEE Conference on Computer Vision and Pattern Recognition, 2008:1-8.
[9] Gudmundsson S A, Aanaes H, Larsen R. Fusion of stereo vision and time-of-flight imaging for improved 3D estimation. International Journal of Intelligent Systems Technologies and Applications, 2008, 5(3-4): 425-433.
[10] Zhu J, Wang L, Yang R, Davis J, and Pan Z. Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(7): 1400-1414.
[11] Hartley R and Zisserman A. Multiple View Geometry in Computer Vision. Cambridge University Press, 2004.
[12] Dal Mutto C, Zanuttigh P, Cortelazzo G M. A probabilistic approach to ToF and stereo data fusion. 3DPVT, Paris, France, 2010, 2: 69.
[13] Payne A, Daniel A, Mehta A, et al. A 512× 424 CMOS 3D Time-of-Flight image sensor with multi-frequency photo-demodulation up to 130Mhz and 2gs/s adc. IEEE International Solid-State Circuits Conference Digest of Technical Papers, 2014: 134-135.
[14] Terven J R, Córdova-Esparza D M. Kin2. A Kinect 2 toolbox for MATLAB. Science of Computer Programming, 2016, 130: 97-106.
[15] Staranowicz A N, Brown G R, Morbidi F, et al. Practical and accurate calibration of RGB-D cameras using spheres. Computer Vision and Image Understanding, 2015, 137: 102-114.
[16] Zhang C, Zhang Z. Calibration between depth and color sensors for commodity depth cameras. Computer vision and machine learning with RGB-D sensors. Springer, Cham, 2014: 47-64.
[17] Shibo L, Qing Z. A new approach to calibrate range image and color image from Kinect. IEEE International Conference on Intelligent Human-Machine Systems and Cybernetics, 2012, 2: 252-255.
[18] Kim C, Yun S, Jung S W, et al. Color and depth image correspondence for Kinect v2. Advanced Multimedia and Ubiquitous Engineering. Springer, Berlin, Heidelberg, 2015: 111-116.
[19] Villena-Martínez V, Fuster-Guilló A, Azorín-López J, et al. A quantitative comparison of calibration methods for RGB-D sensors using different technologies. Sensors, 2017, 17(2): 243.