The final version is available at http://dx.doi.org/10.23919/MVA.2017.7986908

A Linear Method for Recovering the Depth of Ultra HD Cameras Using a Kinect V2 Sensor

Yuan Gao*† Matthias Ziegler† Frederik Zilly† Sandro Esquivel* Reinhard Koch*
*Institute of Computer Science, Christian-Albrechts-University of Kiel, 24118 Kiel, Germany
{yga, sae, rk}@informatik.uni-kiel.de
†Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany
{matthias.ziegler, frederik.zilly}@iis.fraunhofer.de

Abstract

Depth-Image-Based Rendering (DIBR) is a mature and important method for making free-viewpoint videos. As for the study of the DIBR approach, on the one hand, most of current research focuses on how to use it in systems with low resolution cameras, while a lot of Ultra HD rendering devices have been launched into markets. On the other hand, the quality and accuracy of the depth image directly affects the final rendering result. Therefore, in this paper we try to make some improvements on solving the problem of recovering the depth information for Ultra HD cameras with the help of a Kinect V2 sensor. To this end, a linear least squares method is proposed, which recovers the rigid transformation between a Kinect V2 and an Ultra HD camera, using the depth information from the Kinect V2 sensor. In addition, a non-linear coarse-to-fine method, which is based on Sparse Bundle Adjustment (SBA), is compared with this linear method. Experiments show that our proposed method performs better than the non-linear method for the Ultra HD depth image recovery both in computing time and precision.

1 Introduction

With so many Ultra HD resolution 3D TVs and high resolution Virtual Reality (VR) headsets having been launched into the market, the creation of the high-quality and high-resolution contents for these devices is becoming a research hotspot. Depth-Image-Based Rendering (DIBR) [1] is such a method which can be used for free-viewpoint video creation. The depth information in DIBR is very important because the accuracy of it is the key influence factor for the quality of free-viewpoint rendering. Therefore, in this paper, we focus on the depth information recovery for Ultra HD RGB cameras. The depth recovery for multiple RGB cameras has been well researched and can be classified into two categories. One is the light field-to-depth approach [2, 3], the other is stereo matching [4, 5].

Since most of the color-image-based methods above may not address the problem of recovering the depth information of regions without textures, it is also a good choice to solve this problem using the depth information from depth sensors. To achieve this, the calibration between the depth sensor and color cameras is a crucial step. Zhang *et al.* propose a maximum likelihood solution for this calibration problem using a Kinect V1 [6]. However, the distortion of the depth values is not addressed by their method. Herrera *et al.* propose a calibration algorithm for a Kinect V1 depth sensor and a color camera pair with distortion correction [7]. Hansard *et al.* find a 3D projective transformation for the ToF-stereo calibration of a time-of-flight (ToF) sensor and two RGB cameras [8]. Jung *et al.* design a special 2.5D pattern board for the calibration of a low resolution ToF sensor and a high resolution RGB camera [9].

The second version of the Microsoft Kinect (Kinect V2) is also based on the ToF technology and is one of the most high-speed and low-cost ToF sensors in the market. Besides, the difference between Kinect V2 and Kinect V1 is well studied in [10, 11], where it is stated that the Kinect V2 has higher accuracy than Kinect V1.

In this paper, we try to make full use of the ToF sensor in a Kinect V2 camera to map the depth information to an Ultra HD resolution camera. To this end, a linear least squares method is proposed. Specifically, a regular 2D checkerboard is employed to find corresponding points between the Kinect V2 sensor and the Ultra HD camera. Then, the rigid transformation between these two cameras is solved by the least squares method. Furthermore, a non-linear coarse-to-fine solution is also explored and compared with the linear one. The difference between the non-linear approach and the metric calibration method in [12] is that the corner points in the 3D space are recovered through the Kinect V2 sensor. Experiments are conducted on a camera rig with one Kinect V2 and one Sony DSLR camera. Experimental results show the superiority of the proposed linear method both in precision and computing time.

2 Methodology

In this section, the calibration process of a Kinect V2 camera and an Ultra HD camera is introduced in detail.

2.1 Preliminary

The internal parameters of pinhole cameras are important properties for camera calibration. To approximate these factors, substantial methods have been proposed [13]. For the Ultra HD camera in our system, the traditional checkerboard-based method is adopted [14]. For the Kinect V2 camera, the ToF sensor in it can also be modeled as a pinhole camera [15]. Its intrinsic parameters can be accessed through the Kinect for Windows SDK or computed in the same way as for a color camera. These intrinsic parameters are then used to compensate lens distortions of both cameras.

2.2 Linear Method

Suppose a pair of corresponding points in the Kinect V2 and camera image planes is measured by the checkerboard corner-based method. The point in the

3D coordinate system of the Kinect V2 is given as $\boldsymbol{x}_i = [x_i \quad y_i \quad z_i]^{\mathrm{T}}$.The corresponding point in the Ultra HD camera image plane is denoted as $\boldsymbol{u}_i = [u_i \quad v_i \quad 1]^{\mathrm{T}}$ in the homogeneous coordinates. The 3D point $\boldsymbol{x}_i$ is first transferred to the camera coordinate system of the Ultra HD camera using the rigid transformation defined by a rotation $\mathbf{R}$ and a translation $\boldsymbol{t}$, then projected to the image coordinate system of this Ultra HD camera. Therefore, the transformation and projection process can be described as:

$$\mathbf{K}(\mathbf{R}\boldsymbol{x}_i + \boldsymbol{t}) = \lambda \boldsymbol{u}_i \tag{1}$$

Here, $\mathbf{K}$ is the camera matrix of the Ultra HD camera and λ is a scaling factor. To simplify equation (1), we use

$$\boldsymbol{p}_i = \mathbf{K}^{-1}\boldsymbol{u}_i \tag{2}$$

on the right side, where $\boldsymbol{p}_i$ is a calibrated image point and $\boldsymbol{p}_i = [p_i \quad q_i \quad 1]^{\mathrm{T}}$. Therefore, equation (1) becomes:

$$\mathbf{R}\boldsymbol{x}_i + \boldsymbol{t} = \lambda \boldsymbol{p}_i \tag{3}$$

The scaling factor λ is calculated from equation (3) as:

$$\lambda = [r_{31} \quad r_{32} \quad r_{33}]\,\boldsymbol{x}_i + t_3 \tag{4}$$

Then, equation (3) can be written as:

$$\begin{bmatrix} \boldsymbol{x}_i^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} & -p_i\boldsymbol{x}_i^{\mathrm{T}} & 1 & 0 & -p_i \\ \mathbf{0}^{\mathrm{T}} & \boldsymbol{x}_i^{\mathrm{T}} & -q_i\boldsymbol{x}_i^{\mathrm{T}} & 0 & 1 & -q_i \end{bmatrix} \boldsymbol{h} = \mathbf{0} \tag{5}$$

where

$$\boldsymbol{h} = [r_{11} \quad r_{12} \quad \cdots \quad r_{33} \quad t_1 \quad t_2 \quad t_3]^{\mathrm{T}} \tag{6}$$

Here, $\boldsymbol{h}$ is a vector with twelve variables describing the rigid transformation between the ToF sensor of Kinect V2 and the Ultra HD camera. This problem can be solved by a Linear Least Squares (LLS) method like Singular Value Decomposition (SVD). At least six corresponding point pairs should be utilized to solve it. However, the drawback of this method is that the output $\mathbf{R}$ is not a standard rotation matrix, which is needed for refinement by other parameter estimation methods, *e.g.*, Bundle Adjustment (BA).

2.3 Non-Linear Method

A non-linear method is exploited here in order to make a comparison with the linear one. A coarse-to-fine strategy is adopted to make the method more robust.

2.3.1 Coarse Estimation

Suppose there are a Kinect V2 and an Ultra HD camera. A corresponding point pair in the Kinect V2 3D coordinate and the Utra HD camera image plane are denoted as $\boldsymbol{x}_i$ and $\boldsymbol{u}_i$ as in Section 2.2. The coarse estimation step is designed to give a coarse estimation of the rigid transformation from the Kinect V2 depth sensor coordinate system to the Ultra HD camera coordinate system, which is defined as ($\mathbf{R}$, $\boldsymbol{t}$). The rigid transformation estimation is obtained by minimizing the following formula:

$$\min_{\mathbf{R},\boldsymbol{t}} \sum_{i=1}^{n} \|\boldsymbol{u}_i - \hat{\boldsymbol{u}}(\boldsymbol{x}_i, \mathbf{K}, \mathbf{R}, \boldsymbol{t})\|^2 \tag{7}$$

Here, n stands for the number of corresponding point pairs. Note that the 3D coordinate system of the ToF sensor in the Kinect V2 camera is used as the reference coordinate system here, which is different from [14], where the 3D coordinates in a model plane are treated as the reference coordinate system. Equation (7) describes a non-linear least squares problem, which can be solved by the Levenberg-Marquardt optimization approach.

2.3.2 Estimation Refinement

After the coarse estimation process, the $\mathbf{R}$ and $\boldsymbol{t}$ for transferring the 3D points from a Kinect V2 ToF sensor to an Ultra HD camera have been computed. To make this problem more general, suppose there is one Kinect V2 camera with multiple Ultra HD cameras, the number of which is denoted as m. The rigid transformation between the Kinect V2 and the Ultra HD camera j is expressed as $\mathbf{R}_j$ and $\boldsymbol{t}_j$. The estimation refinement step can be formulated as:

$$\min_{\mathbf{K}_j,\mathbf{R}_j,\boldsymbol{t}_j} \sum_{j=1}^{m+1} \sum_{i=1}^{n} v_{ij} \|\boldsymbol{u}_{ij} - \hat{\boldsymbol{u}}(\boldsymbol{x}_i, \mathbf{K}_j, \mathbf{R}_j, \boldsymbol{t}_j)\|^2 \tag{8}$$

Here, $\mathbf{K}_j, \mathbf{R}_j, \boldsymbol{t}_j$ with $j = 1, \ldots, m$ relate to the Ultra HD cameras and $\mathbf{K}_{m+1}, \mathbf{R}_{m+1}, \boldsymbol{t}_{m+1}$ to the Kinect V2 sensor. Furthermore, $\boldsymbol{u}_{ij}$ is the ground truth point in the j-th image plane corresponding to a 3D point $\boldsymbol{x}_i$ in the world coordinate system, and $v_{ij} \in \{0, 1\}$ denotes the visibility between these two points. To solve this problem, the 3D coordinate system of the ToF sensor in the Kinect V2 camera is set as the world coordinate system. A generic Sparse Bundle Adjustment (SBA) method is employed to solve this non-linear least squares problem efficiently [16].

(a) Frontal view (b) Vertical view

Figure 1. Cameras in our system.

3 Experiment

3.1 Experimental Settings

System: The system is built on a rig on a tripod, using a Sony α7R II DSLR camera mounted with a Canon lens and a Kinect V2 sensor. The positions and orientations of these two cameras are illustrated in Fig. 1. The displacement between the centers of the two cameras is around 19 centimeters. Note that the original resolution of the Sony camera is $7,952 \times 5,304$ pixels, which is downsampled to the Ultra HD resolution of $3,840 \times 2,561$ pixels for the experimental evaluations described here. The depth sensor in a Kinect V2 has a resolution of 512×424 pixels.

Field of view: The Field of View (FOV) of the depth sensor in the Kinect V2 camera is $\approx$ 70 degrees [17]. Since the focal length of the Sony camera can be adjusted, we set the FOV of the Ultra HD camera to a similar FOV as the Kinect V2 sensor. An example capture of both cameras for the same scene is shown in (a) and (b) in Fig. 2.

Table I. The RMSE and computing time of different methods.

Method	RMSE (pixel)	Time (ms)
Linear Method	0.781	1.2
Non-linear Method (Coarse)	2.045	1.6
Non-linear Method (Refine)	0.993	16.0

Intrinsic parameter: The intrinsic parameters of both the Ultra HD camera and the ToF sensor in the Kinect V2 camera are estimated by a standard checkerboard-based calibration process. The checkerboard used in our experiments has 266 (19×14) corners and the size of each black or white square field is 15×15 mm. The internal parameters are then utilized to undistort all the output views of both cameras.

Corresponding point pair: To evaluate the effects of both methods, corresponding point pairs need to be found in advance. Here, a bigger checkerboard with 54 (9×6) corners is placed in the jointly visible areas of both cameras twice. The size of each square of this checkerboard is 52×52 mm. Therefore, in total 108 corner points are detected automatically in each camera. It should be noted that in the view of the Kinect V2 camera, the infrared view is actually used for detecting the corner points and the depth values of them are estimated by using the same specific filter as described in [18] on the corresponding depth image.

Ultra HD depth recovery: Because there is a significant difference in resolutions of these two cameras, it is prone to get a recovered depth image with most of the information missing by directly performing the rigid transformation from the low-resolution Kinect V2 ToF sensor to the Ultra HD camera image. To solve this problem, an oversampling strategy in DIBR is employed here [19]. Rigid transformation is done after oversampling the depth image in the Kinect V2 with a factor of 10 using the Nearest-Neighbor method.

Evaluation standard: Here, the Root-Mean-Square Error (RMSE) is adopted to evaluate the effects of our proposed method. It estimates the precision in pixels only in the Ultra HD image plane using the same corresponding point pairs as above.

All experiments are conducted on an Intel Core $i3-4030U$ laptop with 16 GB memory and no GPU acceleration.

3.2 Results and Analysis

The quantitative error report of our proposed linear method and the non-linear method is shown in Table I. The linear method outperforms the coarse-to-fine non-linear method. The reason for this may be that the depth information of the points in the ToF sensor of the Kinect V2 is not accurate enough, while the SBA algorithm heavily relies on the accurate structure of these points [20]. The computation time of both algorithms is also exhibited in Table I. The linear method is around 14 times faster than the non-linear one.

The visualization of the final recovered depth image for the Ultra HD camera is illustrated in Fig. 2. Note that, in (a), each pixel corresponds to a depth pixel which is not shown here. It is called registered color image corresponding to a registered depth image, which is plotted with the help of the depth-to-color map in the Kinect for Windows SDK for a better understanding. The recovered color (c) and recovered depth (d) are the recovered results for the Ultra HD camera using the registered color image and registered depth image respectively with our proposed linear method. Both of them have the same resolution as the Ultra HD camera view in (b). It can be found that the depth image (d) is well recovered except for some occlusion regions which are caused by the displacement between cameras.

Both the linear and the non-linear method can be extended to the case of multiple Ultra HD cameras. For lack of space, the case of only two cameras is evaluated here. In addition, the proposed linear calibration method can also be used for recovering the color information for the depth map using a color camera.

4 Conclusion

In this paper, the problem of recovering the depth information of a high resolution Ultra HD camera using a low resolution Kinect V2 sensor is tried to be solved. A linear solution method is proposed for this problem, which is also compared with a coarse-to-fine non-linear method. Experimental results demonstrate the effectiveness and efficiency of this linear method, which performs better than the other. Moreover, the recovered Ultra HD depth image still has room for quality improvement, which will be our next research goal.

Acknowledgement

This project has received funding from the European Union's Framework Programme for Research and Innovation Horizon 2020 (2014-2020) under the Marie Skłodowska-Curie Actions Grant Agreement No. 676401, the Fraunhofer and Max Planck cooperation program within the German pact for research and innovation (PFI), Intel-VCI-CAU and the German Research Foundation (DFG) No. K02044/8-1.

References

[1] Christoph Fehn, René De La Barré, and Siegmund Pastoor, "Interactive 3-dtv-concepts and key technologies," *Proceedings of the IEEE*, vol. 94, no. 3, pp. 524–538, 2006.

[2] Ting-Chun Wang, Alexei A. Efros, and Ravi Ramamoorthi, "Depth estimation with occlusion modeling using light-field cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 38, no. 11, pp. 2170–2181, 2016.

[3] Łukasz Dąbała, Matthias Ziegler, Piotr Didyk, Frederik Zilly, Joachim Keinert, Karol Myszkowski, H-P Seidel, Przemysław Rokita, and Tobias Ritschel, "Efficient multi-image correspondences for on-line light field video processing," *Computer Graphics Forum*, vol. 35, no. 7, pp. 401–410, 2016.

[4] Beau Tippetts, Dah Jye Lee, Kirt Lillywhite, and James Archibald, "Review of stereo vision algorithms and their suitability for resource-limited systems," *Journal of Real-Time Image Processing*, vol. 11, no. 1, pp. 5–25, 2016.

[5] Xiaoyan Hu and Philippos Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 11, pp. 2121–2133, 2012.

(a) Kinect V2 view (registered color image)

(b) Ultra HD camera view

(c) Recovered color

(d) Recovered depth

Figure 2. Experimental results. (a) and (b) are the captured views for the same scene in the Kinect V2 and the Ultra HD cameras. (c) and (d) are the recovered color and depth views in the Ultra HD resolution.

[6] Cha Zhang and Zhengyou Zhang, "Calibration between depth and color sensors for commodity depth cameras," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–6.

[7] Daniel Herrera, Juho Kannala, and Janne Heikkilä, "Joint depth and color camera calibration with distortion correction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 10, pp. 2058–2064, 2012.

[8] Miles Hansard, Georgios Evangelidis, Quentin Pelorson, and Radu Horaud, "Cross-calibration of time-of-flight and colour cameras," *Computer Vision and Image Understanding (CVIU)*, vol. 134, pp. 105–115, 2015.

[9] Jiyoung Jung, Joon-Young Lee, Yekeun Jeong, and In So Kweon, "Time-of-flight sensor calibration for a color and depth camera pair," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 37, no. 7, pp. 1501–1513, 2015.

[10] Andrea Corti, Silvio Giancola, Giacomo Mainetti, and Remo Sala, "A metrological characterization of the kinect v2 time-of-flight camera," *Robotics and Autonomous Systems (RAS)*, vol. 75, Part B, pp. 584–594, 2016.

[11] S Zennaro, M Munaro, S Milani, P Zanuttigh, A Bernardi, S Ghidoni, and E Menegatti, "Performance evaluation of the 1st and 2nd generation kinect for multimedia applications," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2015, pp. 1–6.

[12] Vaibhav Vaish, Bennett Wilburn, Neel Joshi, and Marc Levoy, "Using plane + parallax for calibrating dense camera arrays," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004, vol. 1, pp. 2–9.

[13] Richard Hartley and Andrew Zisserman, *Multiple view geometry in computer vision*, Cambridge university press, 2003.

[14] Zhengyou Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 22, no. 11, pp. 1330–1334, 2000.

[15] Radu Horaud, Miles Hansard, Georgios Evangelidis, and Menier Clément, "An overview of depth cameras and range scanners based on time-of-flight technologies," *Machine Vision and Applications (MVA)*, vol. 27, no. 7, pp. 1005–1020, 2016.

[16] Manolis IA Lourakis and Antonis A Argyros, "Sba: A software package for generic sparse bundle adjustment," *ACM Transactions on Mathematical Software (TOMS)*, vol. 36, no. 1, pp. 2, 2009.

[17] Ziyun Cai, Jungong Han, Li Liu, and Ling Shao, "Rgb-d datasets using microsoft kinect or similar sensors: a survey," *Multimedia Tools and Applications (MTA)*, pp. 1–43, 2016.

[18] Valeria Garro, Carlo Dal Mutto, Pietro Zanuttigh, and Guido M Cortelazzo, "A novel interpolation scheme for range data with side information," in *Conference for Visual Media Production (CVMP)*. IEEE, 2009, pp. 52–60.

[19] Sveta Zinger, Luat Do, and PHN de With, "Free-viewpoint depth image based rendering," *Journal of Visual Communication and Image Representation*, vol. 21, no. 5, pp. 533–541, 2010.

[20] Marvin Lindner, Ingo Schiller, Andreas Kolb, and Reinhard Koch, "Time-of-flight sensor calibration for accurate range sensing," *Computer Vision and Image Understanding (CVIU)*, vol. 114, no. 12, pp. 1318–1328, 2010.