The final version is available at http://dx.doi.org/10.1109/3DV.2017.00013

A Novel Self-Calibration Method for a Stereo-ToF System Using a Kinect V2 and Two 4K GoPro Cameras

Yuan Gao, Sandro Esquivel, Reinhard Koch
Christian-Albrechts-University of Kiel
24118 Kiel, Germany
{yga, sae, rk}@informatik.uni-kiel.de

Joachim Keinert
Fraunhofer Institute for Integrated Circuits IIS
91058 Erlangen, Germany
joachim.keinert@iis.fraunhofer.de

Abstract

A new light-field movie capture device using a Kinect V2 sensor and two 4K GoPro cameras is presented in this paper. Due to the uncontrollable tilt of the Kinect V2 camera, it is hard to obtain a constant rigid transformation between the stereo- and ToF-camera systems. To this end, a novel self-calibration method is proposed, which takes advantage of the geometric constraints from the scene and the cameras. Specifically, a camera orientation approximation approach is utilized to estimate the rigid transformation of the stereo-ToF system based on reliable point pairs filtered by the geometric constraints. Besides, a depth correction step is exploited to improve the depth accuracy of the Kinect V2 sensor. Moreover, a depth fusion strategy for the stereo- and ToF-depth data is proposed to provide more accurate depth images in 4K resolution. Experimental results demonstrate the effectiveness of the proposed depth correction step, stereo-ToF calibration method and depth fusion strategy.

1. Introduction

With more and more 4K Ultra High Definition (UHD) 3D TVs and high-fidelity Virtual Reality (VR) Head-Mounted Displays (HMDs) having been launched into the consumer market, how to create high-resolution and high-quality contents for these devices is becoming a research hotspot. The Depth-Image-Based Rendering (DIBR) approach is such kind of method which is capable of producing free-viewpoint videos for the above devices [33, 42, 9]. Since the accuracy and the resolution of the depth images in DIBR are the critical influence factors for the rendering of free-viewpoint videos, in this paper, how to recover the depth information for a 4K resolution RGB camera is investigated. Currently, the recovery of the depth information using multiple RGB cameras has been well researched, which can be classified into two categories. One is the light field-to-depth approach [35, 4], the other is the stereo matching method [34, 19]. Apart from the RGB cameras-based algorithms, the depth information can also be acquired from external sensors. For example, the second version of the Microsoft Kinect (Kinect V2) is one of the most low-cost and high-speed Time-of-Flight (ToF) sensors in the market [3]. The comparison between the Kinect V2 and the first generation of Microsoft Kinect (Kinect V1) is well studied in [36, 21, 31, 38], where the Kinect V2 exhibits a better performance in a lot of multimedia applications than the Kinect V1.

(a) Frontal view (b) Vertical view

Figure 1. A multi-camera rig for capturing light-field movies. The Kinect V2 sensor is in the middle of the two 4K GoPro cameras.

Motivation: The multi-camera rig in Fig. 1 is a prototypical movie capture system for making light-field movies. Two GoPro Hero3+ cameras with a Kinect V2 sensor are mounted on this rig. In order to increase angular resolution and minimize lens distortion, the original lenses of both GoPro cameras have been replaced by two customized lenses with the same Field of View (FOV) of about 70 degrees [4]. The camera resolutions of these two GoPro cameras are 4K $(4,000 \times 3,000)$. There are two challenging problems of this multi-camera rig that need to be solved. First, the GoPro cameras are well fixed on the camera rig, while the Kinect V2 is impossible to be fixed because of the changeable tilt of it, which easily leads to the inconsistency of the camera extrinsic parameters after the multi-camera rig being moved. How to automatically calibrate this stereo-ToF system using the scene and camera constraints is challenging. The other challenging problem is that it is difficult to recover a 4K-resolution depth image for one of the GoPro

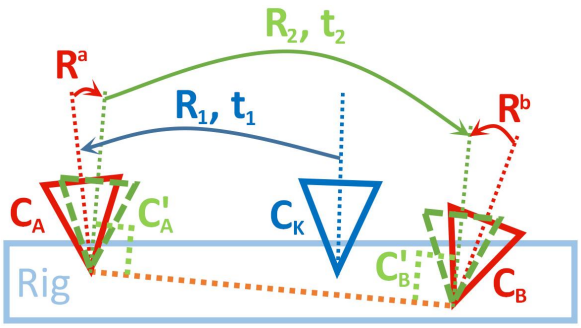


Figure 2. A virtual vertical view of the multi-camera rig for describing the camera calibration problems.

cameras, considering that the distance between the two GoPro cameras is quite large (around 39 cm), and little literature focuses on the stereo- and ToF-depth fusion problem in 4K resolution.

To solve the two challenging problems of the multi-camera rig, a depth correction step, a stereo-ToF calibration method and a depth fusion strategy are proposed in this paper. In particular, the depth correction step increases the depth accuracy of the Kinect V2 camera. The stereo-ToF calibration method is based on the reliable point pairs, which are detected by an off-the-shelf feature point detector and filtered using geometric constraints in the cameras and the scene [12]. Besides, the camera rotation matrix can be linearly approximated because the Kinect V2 and the GoPro cameras have similar orientations. The depth fusion strategy exploits the rigid transformation result of the stereo-ToF calibration method to fuse the depth information from the stereo matching method and the ToF sensor at the pixel level. Experimental results show that the depth correction step contributes to the stereo-ToF calibration. The stereo-ToF calibration method using the camera orientation approximation achieves the best performance compared with baseline approaches. Moreover, the depth fusion strategy is capable of creating depth images of better quality than using the stereo matching method or the ToF sensor alone.

The paper is organized as follows. Section 2 focuses on the introduction of related works. Section 3 outlines a depth correction step, a coarse-to-fine camera calibration framework using the reliable point pairs from the filtering of the scene and camera constraints, and a pixel-level depth fusion strategy. Section 4 is devoted to the experiments and analysis of the camera calibration and the depth fusion. Finally, section 5 concludes and summarizes this paper.

2. Related Work

The Perspective-n-Point (PnP) problem is first described in [10], which stands for the problem of how to estimate the camera pose of a calibrated camera using n known 3D reference points in the world coordinate frame and their corresponding 2D points on the camera image plane of this calibrated camera. The solutions to the PnP problem can be classified into two categories: iterative and non-iterative methods. As for the iterative-based methods, Lu *et al*. minimize an object-space collinearity error for computing orthogonal rotation matrices, which is proven to be globally convergent [28]. Zhang proposes a closed-form solution for estimating the camera intrinsic and extrinsic parameters, and then refines them with the Levenberg-Marquardt algorithm [39]. With regard to the non-iterative-based methods, Lepetit *et al*. express the non-iterative solution to the PnP problem as a vector standing for a weighted sum of the null eigenvectors and their method achieves the computational complexity in n [22]. Li *et al*. also present an O(n) solution by estimating the coordinates of two special end points [23].

As for the fusion of stereo- and ToF-depth data in non-4K resolution, several methods have been proposed [30]. Zhu *et al*. use the MAP-MRF Bayesian framework to solve the stereo and ToF data fusion problem and a belief propagation-based method is applied to fulfill the depth inference [41, 40]. Gandhi *et al*. utilize a Bayesian fusion method within an efficient seed-growing algorithm to solve the same problem [11]. More recently, Dal Mutto *et al*. also use the MAP-MRF framework to fuse the stereo- and ToF-depth data with considering the mixed pixel effect [6]. Evangelidis *et al*. address the stereo-ToF fusion problem by solving a set of local energy optimization problems hierarchically [8]. Marin *et al*. extend the Local Consistency (LC) fusion framework of [7] with taking into account of the depth data confidence [29]. With regard to the calibration between the stereoscopic camera pair and the ToF sensor in the publicly-available datasets [8, 6], checkerboard-based methods are taken advantage of [16, 5]. However, both of these two datasets do not contain color image contents in 4K resolution.

3. Methodology

In this section, a coarse-to-fine framework for the calibration process of the stereo-ToF system is presented after a brief introduction to the self-calibration problem and the reliable point pair detection step. Besides, a depth fusion strategy for different depth sources is described in the end of this section.

3.1. Preliminary

Since there are two GoPro cameras and one Kinect V2 sensor on the multi-camera rig, for the sake of describing convenience, the left and right GoPro cameras are denoted as $\mathbb{C}_{\mathrm{A}}$ and $\mathbb{C}_{\mathrm{B}}$ respectively, and the Kinect V2 camera is denoted as $\mathbb{C}_{\mathrm{K}}$. Each camera has two basic spaces: camera 3D space and camera image space. The intrinsic matrices of the GoPro and Kinect V2 cameras are defined as $\mathbf{K}_a$, $\mathbf{K}_b$, $\mathbf{K}_k$ respectively, and the lens distortions of them are assumed to

have been corrected. It should be noted that the coordinate system of the Kinect V2 camera $\mathbb{C}_\mathrm{K}$ coincides with that of the RGB sensor in it, by which color and depth images in Full High Definition (FHD) resolution are captured. More details concerning this are explained in section 4.1.

The self-calibration of this stereo-ToF system is defined as to estimate the rigid transformation ($\mathbf{R}_1$, $\boldsymbol{t}_1$) from the Kinect V2 sensor $\mathbb{C}_\mathrm{K}$ to the left GoPro camera $\mathbb{C}_\mathrm{A}$. The calibration of the stereoscopic GoPro camera pair is expressed as to measure the rotation rectification matrices $\mathbf{R}^a$ and $\mathbf{R}^b$, which turn $\mathbb{C}_\mathrm{A}$ into a virtual camera $\mathbb{C}'_\mathrm{A}$ and $\mathbb{C}_\mathrm{B}$ into another virtual camera $\mathbb{C}'_\mathrm{B}$. Typically, the intrinsic camera matrices of the virtual cameras $\mathbb{C}'_\mathrm{A}$ and $\mathbb{C}'_\mathrm{B}$ are the same, *i.e.* $\mathbf{K}'_a = \mathbf{K}'_b$. The straight line going through the optical centers of both GoPro cameras are parallel to the coplanar camera image planes of $\mathbb{C}'_\mathrm{A}$ and $\mathbb{C}'_\mathrm{B}$. The rigid transformation from $\mathbb{C}'_\mathrm{A}$ to $\mathbb{C}'_\mathrm{B}$ is then defined as ($\mathbf{R}_2$, $\boldsymbol{t}_2$), where $\mathbf{R}_2$ should be an identity matrix. The above calibration descriptions are illustrated in Fig. 2 as well. It can be found that $\mathbb{C}_\mathrm{K}$ is not in the middle of the stereo GoPro pair. The reason is that the RGB sensor in $\mathbb{C}_\mathrm{K}$ is physically closer to $\mathbb{C}_\mathrm{B}$ than to $\mathbb{C}_\mathrm{A}$ as shown in Fig. 1 (a).

Depth Correction: The depth accuracy of a Kinect V2 camera has a constant offset of -18 mm, which is well evaluated in [36]. It is important to correct the depth images from the Kinect V2 sensor by compensating this accuracy offset before fusing the stereo- and ToF-depth data. Otherwise the misalignment of the point clouds derived from the ToF sensor and the stereo matching method would happen, which is analyzed in section 4.2.

3.2. Reliable Point Pair Detection

Interest point detection has been well studied in the computer vision field [15]. The traditional SIFT keypoint detector and descriptor are used here for their robustness [27]. Another reason for only evaluating SIFT is that, since $\mathbb{C}_\mathrm{K}$ is able to capture FHD resolution color images and $\mathbb{C}_\mathrm{A}$ can capture 4K resolution images, any classical interest point detection algorithm is capable of detecting adequate reliable feature points for the following processes. Besides, the influence of choosing another type of keypoint detector and descriptor on the calibration result is negligible for the experiments.

The detected feature points in the camera image spaces of $\mathbb{C}_\mathrm{A}$ and $\mathbb{C}_\mathrm{K}$ are then exploited to compose reliable matched point pairs. Here, the k-Nearest-Neighbors (KNN) [1] and ratio test [27] methods are utilized to fulfill this task. Afterwards, the resulting corresponding pairs still contain some outliers, which are filtered by using epipolar constraints with the RANdom SAmple Consensus (RANSAC) algorithm [10, 17]. The remaining corresponding point pairs are assumed to be accurate for the subsequent processes.

3.3. Coarse-to-Fine Framework

A coarse-to-fine framework [13] is generally composed of a coarse estimation step based on the solution to the PnP problem, and an estimation refinement step based on the bundle adjustment algorithm [25], which are introduced as follows.

3.3.1 Coarse Estimation

Suppose the number of the corresponding point pairs is n. A corresponding point pair is denoted as ($\boldsymbol{u}_i^a$, $\boldsymbol{u}_i^k$), where $\boldsymbol{u}_i^a = \begin{bmatrix} u_i^a & v_i^a & 1 \end{bmatrix}^\mathrm{T}$ is a 2D point in the camera image space of $\mathbb{C}_\mathrm{A}$, and $\boldsymbol{u}_i^k$ is a 2D point in the camera image space of $\mathbb{C}_\mathrm{K}$, having the same format as $\boldsymbol{u}_i^a$. For $\boldsymbol{u}_i^k$, the corresponding 3D point $\boldsymbol{x}_i^k = \begin{bmatrix} x_i^k & y_i^k & z_i^k & 1 \end{bmatrix}^\mathrm{T}$ in the camera 3D space of $\mathbb{C}_\mathrm{K}$ is calculated by using the intrinsic camera matrix $\mathbf{K}_k$ and the depth information of $\mathbb{C}_\mathrm{K}$. The coarse estimation step is essentially to estimate the rigid transformation from the camera coordinates of $\mathbb{C}_\mathrm{K}$ to the camera coordinates of $\mathbb{C}_\mathrm{A}$ by calculating the below formula:

$$\min_{\mathbf{R}_1^2, \boldsymbol{t}_1^2} \sum_{i=1}^{n} \| \boldsymbol{u}_i^a - \hat{\boldsymbol{u}}(\boldsymbol{x}_i^k, \mathbf{K}_a, \mathbf{R}_1^2, \boldsymbol{t}_1^2) \|^2, \tag{1}$$

where:

$$\hat{\boldsymbol{u}}(\boldsymbol{x}, \mathbf{K}, \mathbf{R}, \boldsymbol{t}) = \mathrm{proj}(\mathbf{K} \begin{bmatrix} \mathbf{R} & \boldsymbol{t} \end{bmatrix} \boldsymbol{x}). \tag{2}$$

The results of the coarse estimation stage are denoted as ($\mathbf{R}_1^2$, $\boldsymbol{t}_1^2$). Several methods have been proposed for solving the above PnP problem, *e.g.*, LHM [28], EPnP [22], and RPnP [23]. Based on the specific camera structure of our stereo-ToF system, a camera orientation approximation-based PnP solution is presented as below.

Camera Orientation Approximation: When observing the camera configuration in Fig. 1, it can be found that $\mathbb{C}_\mathrm{K}$ and $\mathbb{C}_\mathrm{A}$ have a minor orientation difference, which indicates that the rotation matrix $\mathbf{R}_1^2$ can be approximated by a linear method in [26]. Suppose the camera orientation difference between $\mathbb{C}_\mathrm{K}$ and $\mathbb{C}_\mathrm{A}$ is expressed as $\boldsymbol{r} = \begin{bmatrix} \alpha & \beta & \gamma \end{bmatrix}^\mathrm{T}$. The approximated rotation matrix $\mathbf{R}_1^2$ is denoted as:

$$\mathbf{R}_1^2 = \begin{bmatrix} 1 & -\gamma & \beta \\ \gamma & 1 & -\alpha \\ -\beta & \alpha & 1 \end{bmatrix}. \tag{3}$$

After transforming the 2D image point $\boldsymbol{u}_i^a$ into a 2D point $\boldsymbol{p}_i^a = \begin{bmatrix} p_i^a & q_i^a & 1 \end{bmatrix}^\mathrm{T}$ on the normalized image plane of $\mathbb{C}_\mathrm{A}$ with using $\mathbf{K}_a$, the rigid transformation progress can be written as:

$$\begin{bmatrix} \mathbf{R}_1^2 & \boldsymbol{t}_1^2 \end{bmatrix} \boldsymbol{x}_i^k = \lambda \boldsymbol{p}_i^a. \tag{4}$$

The scaling factor λ can be derived from the combination of equation (3) and (4), expressed as follows:

$$\lambda = \begin{bmatrix} -\beta & \alpha & 1 & \boldsymbol{t}_1^2(3) \end{bmatrix} \boldsymbol{x}_i^k. \tag{5}$$

Afterwards, equation (4) is written as:

$$\mathbf{A}\begin{bmatrix}\boldsymbol{r}\\ \boldsymbol{t}_1^2\end{bmatrix}=\begin{bmatrix}p_i^a z_i^k - x_i^k\\ q_i^a z_i^k - y_i^k\end{bmatrix}, \tag{6}$$

where:

$$\mathbf{A}=\begin{bmatrix}-p_i^a y_i^k & (p_i^a x_i^k + z_i^k) & -y_i^k & 1 & 0 & -p_i^a\\ -(q_i^a y_i^k + z_i^k) & q_i^a x_i^k & x_i^k & 0 & 1 & -q_i^a\end{bmatrix}. \tag{7}$$

The linear least-squares problem presented in equation (6) and (7) can be solved by using the SVD algorithm to compute a pseudo-inverse or using normal equations, requiring at least three corresponding point pairs, *i.e.* $n \geqslant 3$. The approximated rotation matrix $\mathbf{R}_1^2$ is then converted to a standard rotation matrix by normalization.

3.3.2 Estimation Refinement

Due to the depth precision and flying pixel problems of any Kinect V2 device [36, 31, 24], the 3D point $\boldsymbol{x}_i^k$ generated from $\boldsymbol{u}_i^k$ is not equal to the ground truth 3D point in the camera coordinate system of $\mathbb{C}_\mathrm{K}$, which is further refined by using the formula as below:

$$\min_{\mathbf{R}_1^j, \boldsymbol{t}_1^j, \boldsymbol{x}_i^k} \sum_{j=1}^{2}\sum_{i=1}^{n} \|\boldsymbol{u}_i^j - \hat{\boldsymbol{u}}(\boldsymbol{x}_i^k, \mathbf{K}_j, \mathbf{R}_1^j, \boldsymbol{t}_1^j)\|^2. \tag{8}$$

Here, the input parameter $\mathbf{R}_1^1$ is a 3×3 identity matrix and $\boldsymbol{t}_1^1$ is a zero vector. Besides, the input parameters $\mathbf{R}_1^2$ and $\boldsymbol{t}_1^2$ are the results of the previous coarse estimation step. When $j=1$, it indicates that j-relevant parameters are also related to camera $\mathbb{C}_\mathrm{K}$. Therefore, $\mathbf{K}_1$ is as same as $\mathbf{K}_k$, and $\boldsymbol{u}_i^1$ is equal to $\boldsymbol{u}_i^k$. As for $j=2$, it turns into the case of camera $\mathbb{C}_\mathrm{A}$ where $\mathbf{K}_2 = \mathbf{K}_a$ and $\boldsymbol{u}_i^2 = \boldsymbol{u}_i^a$. The nonlinear least-squares optimization problem defined in (8) can be solved by a robust bundle adjustment approach efficiently [37].

The final rigid transformation $(\mathbf{R}_1, \boldsymbol{t}_1)$ from the camera coordinates of $\mathbb{C}_\mathrm{K}$ to the camera coordinates of $\mathbb{C}_\mathrm{A}$ is expressed as follows:

$$\begin{aligned}\mathbf{R}_1 &= \mathbf{R}_1^2(\mathbf{R}_1^1)^\mathrm{T},\\ \boldsymbol{t}_1 &= \boldsymbol{t}_1^2 - \mathbf{R}_1\boldsymbol{t}_1^1.\end{aligned} \tag{9}$$

Note that, equation (9) can also stand for the stereo-ToF calibration result of only using the coarse estimation method by replacing $\mathbf{R}_1^1$ with an identity matrix, and $\boldsymbol{t}_1^1$ with a zero vector.

3.4. Depth Fusion

The depth data of the Kinect V2 sensor and the stereo matching method are fused together in the camera image space of $\mathbb{C}'_\mathrm{A}$. The rigid transformation $(\mathbf{R}_3, \boldsymbol{t}_3)$ from the camera coordinates of $\mathbb{C}_\mathrm{K}$ to the camera coordinates of $\mathbb{C}'_\mathrm{A}$ is denoted as:

$$\begin{aligned}\mathbf{R}_3 &= \mathbf{R}^a\mathbf{R}_1,\\ \boldsymbol{t}_3 &= \mathbf{R}^a\boldsymbol{t}_1.\end{aligned} \tag{10}$$

The 3D points in the camera 3D space of $\mathbb{C}_\mathrm{K}$ are projected onto the camera image plane of $\mathbb{C}'_\mathrm{A}$ using $(\mathbf{R}_3, \boldsymbol{t}_3)$ and $\mathbf{K}'_a$. Since there exists an image resolution difference between $\mathbb{C}_\mathrm{K}$ and $\mathbb{C}'_\mathrm{A}$, the above 3D-point-projection solution may cause information loss, *i.e.* sparse points on the destination image plane. To solve this problem, a universal oversampling strategy in DIBR is employed here [42]. The oversampling rate $s\,(=4)$ is applied to adjust the resolution of all the captured images of $\mathbb{C}_\mathrm{K}$ and the intrinsic camera matrix $\mathbf{K}_k$. Afterwards, in the camera image space of $\mathbb{C}'_\mathrm{A}$, there are two depth images. One is projected from the Kinect V2 sensor, which is denoted as $\mathbf{D}'_k$. The other is the depth result of using the stereo matching method with the GoPro camera pair, which is expressed as $\mathbf{D}'_s$. The image resolutions of $\mathbf{D}'_k$ and $\mathbf{D}'_s$ are the same as that of the 4K GoPro camera $\mathbb{C}_\mathrm{A}$ or $\mathbb{C}_\mathrm{B}$. Let $(i,\,j)$ be the coordinates of a 2D point on the camera image plane of $\mathbb{C}'_\mathrm{A}$, the fusion strategy for creating the final fused depth image $\mathbf{D}'_f$ is described as below:

$$\mathbf{D}'_f(i,\,j) = \begin{cases}\mathbf{D}'_k(i,\,j), & \text{if } \mathbf{D}'_k(i,\,j) > 0\,;\\ \mathbf{D}'_s(i,\,j), & \text{else.}\end{cases} \tag{11}$$

The above depth fusion strategy is designed by considering the observation that the valid depth points in $\mathbf{D}'_k$ are normally denser than those in $\mathbf{D}'_s$. Besides, for this pixel-level depth fusion strategy, an accurate stereo-ToF calibration is the key to avoiding fusion artifacts in $\mathbf{D}'_f$.

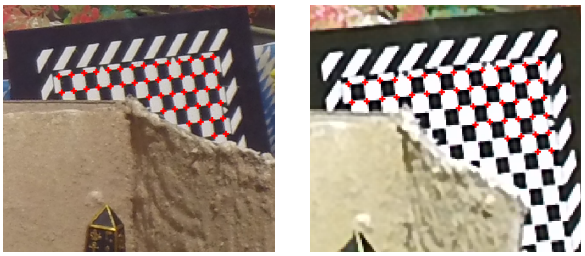
4. Experiments

Experimental data are captured by the multi-camera rig device as illustrated in Fig. 1. An example image of the captured scene is presented in Fig. 5 (a). The details concerning the experimental parameter configuration and analysis are introduced as follows.

4.1. Experimental Settings

Image capture: A control computer is in charge of synchronizing the capture progress of the Kinect V2 sensor $\mathbb{C}_\mathrm{K}$ and two 4K GoPro cameras $\mathbb{C}_\mathrm{A}, \mathbb{C}_\mathrm{B}$ using the signal and time stamp technologies. The Kinect for Windows SDK is utilized to get the captured data from the Kinect V2 camera. Note that, this SDK is able to output color images and their corresponding registered depth images from the RGB sensor in $\mathbb{C}_\mathrm{K}$[1]. More specifically, both of these two

[1] Refer to the 'MapColorFrameToDepthSpace' method in the Kinect for Windows SDK.



(a) (b)

Figure 3. Detected corners of the $\mathbb{C}_{\mathrm{A}}$ view in (a) and of the $\mathbb{C}_{\mathrm{K}}$ view in (b).

Table I. The RMSE results of the stereo-ToF calibration using different coarse estimation approaches.

Coarse-to-Fine Framework	LHM	EPnP	RPnP	Ours
	Before depth correction			
Coarse estimation	1.82	3.14	1.45	0.87
Estimation Refinement	-	2.19	-	-
	After depth correction			
Coarse estimation	1.78	2.83	1.54	**0.87**
Estimation refinement	-	2.60	-	-

types of images used in experiments are in FHD resolution $(1,920 \times 1,080)$.

Camera Calibration: The intrinsic parameters and the radial and decentering distortion coefficients [2] of all the cameras on the multi-camera rig are measured by using a conventional checkerboard-based method [39]. The same calibration method is then exploited to estimate the extrinsic parameters of the GoPro camera pair in order to compute the rotation rectification matrices ($\mathbf{R}^a$, $\mathbf{R}^b$), the intrinsic camera matrices ($\mathbf{K}'_a$, $\mathbf{K}'_b$), and the perspective transformation matrix $\mathbf{Q}$ used for projecting the disparity map into the camera 3D space of $\mathbb{C}'_{\mathrm{A}}$. Note that there is no need to repeat this step every time because the two GoPro cameras have been fixed on the multi-camera rig.

Stereo-GoPro Depth: The Semi-Global Matching (SGM) algorithm is one of the most effective and efficient stereo matching methods [18], which is used here to estimate the disparity map of $\mathbb{C}'_{\mathrm{A}}$ with the rectified stereo images. The disparity map of $\mathbb{C}'_{\mathrm{A}}$ is projected into the camera 3D space of $\mathbb{C}'_{\mathrm{A}}$ with $\mathbf{Q}$ and then projected back onto the camera image plane of $\mathbb{C}'_{\mathrm{A}}$ to calculate $\mathbf{D}'_s$. Regarding the SGM method, the minimum disparity value is set to 256 and the maximum disparity value is set to $1,280$. The matched block size is equal to 9.

Point Pair Detection: The SIFT detector and descriptor algorithms are implemented by referring to their default implementations in OpenCV. The parameter r for ratio test is set to 0.8. The threshold ε of epipolar constraints in the RANSAC framework is set to 0.5 pixels.

Evaluation Metric: For the evaluation of the stereo-ToF calibration, a checkerboard appearing in both views of $\mathbb{C}_{\mathrm{A}}$ and $\mathbb{C}_{\mathrm{K}}$ is manually labeled at the locations of the common visible corners. Afterwards, a corner refinement approach with sub-pixel accuracy is applied to refine the positions of these corners [32]. As illustrated in Fig. 3, there are m $(= 47)$ common-corner point pairs in both views of $\mathbb{C}_{\mathrm{A}}$ and $\mathbb{C}_{\mathrm{K}}$, each of which is expressed as $(\boldsymbol{u}_i^a, \boldsymbol{u}_i^k)$ as the description in section 3.3.1. When transforming the 2D point $\boldsymbol{u}_i^k$ to a 3D point $\boldsymbol{x}_i^k$ in the camera 3D space of $\mathbb{C}_{\mathrm{K}}$, the intensity-related distance error of the checkerboard in the Kinect V2 device is required to be considered [31, 20, 24]. To compensate the depth error of the corner point $\boldsymbol{u}_i^k$, a specific filter is adopted from [14]. Finally, the Root-Mean-Square Error (RMSE) metric is utilized to evaluate the error of the stereo-ToF calibration:

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m} \|\boldsymbol{u}_i^a - \hat{\boldsymbol{u}}(\boldsymbol{x}_i^k, \mathbf{K}_a, \mathbf{R}_1, \boldsymbol{t}_1)\|^2}. \qquad (12)$$

For the evaluation of the depth fusion strategy, the Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM) and Non-Black Region Proportion (NBRP) approaches are tried. In particular, the DIBR approach is exploited to render new views for $\mathbb{C}'_{\mathrm{B}}$ using the depth images $\mathbf{D}'_s, \mathbf{D}'_k, \mathbf{D}'_f$ respectively. The virtually-rendered images are then compared with the ground-truth images captured by $\mathbb{C}'_{\mathrm{B}}$ using PSNR, SSIM and NBRP metrics.

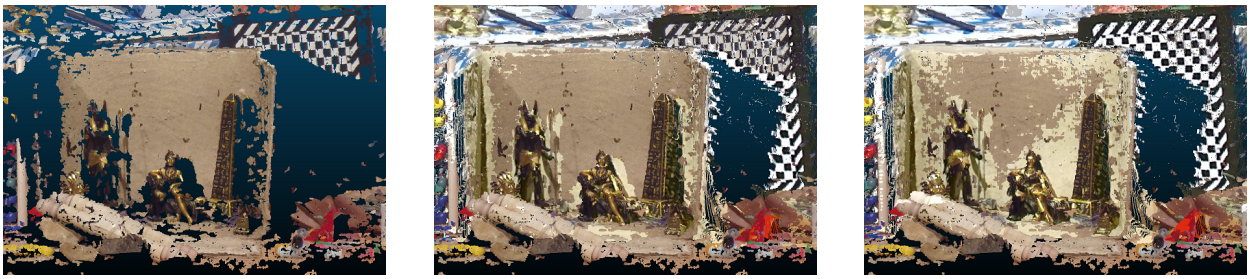
All experiments are conducted on an Intel Core i3-4030U laptop with 16 GB RAM and no GPU acceleration, using the captured datasets from the control computer. Both source code and datasets are going to be released on our website[2].

4.2. Results and Analysis

The proposed depth correction step, stereo-ToF calibration method and depth fusion strategy are analyzed quantitatively and qualitatively.

Quantitative Evaluation: The RMSE results of evaluating the precision of the stereo-ToF calibration methods are illustrated in Table I. Here, the symbol '-' indicates that the result of the estimation refinement step is worse than that of the coarse estimation step. In other words, the reliable point pairs detected by the approaches in section 3.2 make the calibration refinement algorithm get stuck in a local minimum. The calibration refinement algorithm works only for the case of the EPnP-based coarse estimation, while its performance is the worst among these four methods. The depth correction step slightly helps improving the RMSE results of LHM and EPnP for the coarse estimation stage. However, it has no influence on the calibration result of the proposed stereo-ToF calibration method using the camera orientation approximation. In summary, the proposed stereo-

[2] https://ygaokiel.github.io



(a) (b) (c)

Figure 4. The point clouds of a part of the scene in the camera 3D space of $\mathbb{C}'_{\mathrm{A}}$. The point cloud rendered with using the depth image $\mathbf{D}'_s$ is presented in (a). The point clouds from $\mathbf{D}'_s$ and $\mathbf{D}'_k$ without using the depth correction step are shown in (b). With using the depth correction step, the point clouds from $\mathbf{D}'_s$ and $\mathbf{D}'_k$ are exhibited in (c), in which the misalignment disappears compared with (b).

Table II. The performances of different depth sources for the rendering of virtual views using different evaluation metrics.

Depth Source	Before depth correction			After depth correction		
	PSNR	SSIM	NBRP (%)	PSNR	SSIM	NBRP (%)
$\mathbf{D}'_s$	9.29	0.156	61.29	-	-	-
$\mathbf{D}'_k$	11.29	0.235	74.50	11.35	0.245	74.28
$\mathbf{D}'_f$	11.98	0.247	**81.35**	**12.04**	**0.259**	81.15

ToF calibration method achieves the best performance compared with the other three baseline approaches.

The evaluation results of the depth fusion strategy are shown in Table II. The depth correction step improves the PSNR and SSIM values when using $\mathbf{D}'_k$ and $\mathbf{D}'_f$ for the virtual rendering, which indicates that the depth correction is a necessary step for depth accuracy of a Kinect V2 sensor. In addition, using $\mathbf{D}'_f$ from the depth fusion strategy achieves the best virtual-rendering performance in all of these three evaluation metrics, which exhibits the effectiveness of the proposed depth fusion strategy.

Qualitative Evaluation: To evaluate the effectiveness of the depth correction step, the point clouds with using the depth images $\mathbf{D}'_s, \mathbf{D}'_k$ are visualized in the camera 3D space of $\mathbb{C}'_{\mathrm{A}}$ as shown in Fig. 4. The rendering results without and with using the depth correction step are presented in Fig. 4 (b)(c). It can be found that, with using the depth correction step, the misalignment in the places of the golden tower and the checkerboard is eliminated, which indicates that the depth correction step contributes to the self-calibration of this stereo-ToF system.

As for the evaluation of the depth fusion strategy, the projected virtual images on the camera image plane of $\mathbb{C}'_{\mathrm{B}}$ using the depth images $\mathbf{D}'_s, \mathbf{D}'_k, \mathbf{D}'_f$ are shown in Fig. 5. The large black region near the bottom boarder of Fig. 5 (c) is mainly because the vertical FOV of the RGB sensor in the Kinect V2 camera is smaller than that of the GoPro camera $\mathbb{C}_{\mathrm{A}}$. The missing areas near the right boarders of Fig. 5 (b)(c) are caused by the camera displacement of $\mathbb{C}'_{\mathrm{A}}$ and $\mathbb{C}'_{\mathrm{B}}$. The image quality of the virtually projected image in Fig. 5 (b) is worse than that in Fig. 5 (c), which indicates that the Kinect V2 camera is more reliable than the GoPro camera pair for this specific scene. Moreover, the image quality of Fig. 5 (d) is better than those of Fig. 5 (b)(c), which shows that the proposed depth fusion strategy is an effective way of taking full advantage of both $\mathbf{D}'_s$ and $\mathbf{D}'_k$.

5. Conclusion

In this paper, a novel self-calibration method is proposed for a light-field movie capture device composed of a Kinect V2 and two 4K GoPro cameras. The proposed self-calibration method utilizes the geometric constraints in the scene and the cameras to overcome the disadvantage of the changeable tilt of the Kinect V2 camera. In addition, the camera orientation approximation step is used by our self-calibration method, which outperforms other baseline approaches. Moreover, a depth correction step is proven to be beneficial to the self-calibration of this stereo-ToF system. Furthermore, a depth fusion strategy is presented in this paper as well, which relies on the depth correction step and the rigid transformation result of the stereo-ToF calibration method, and is shown to be effective in rendering depth images of higher quality in 4K resolution.

6. Acknowledgments

The work in this paper was funded from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676401, European Training Network on Full Parallax Imaging, Intel-VCI-CAU, the German Research Foundation (DFG) No. K02044/8-1 and the Fraunhofer and Max Planck cooperation program within the German pact for research and innovation (PFI).

(a) Ground truth (Color image of $\mathbb{C}'_{\mathrm{B}}$)

(b) Projected virtual color image using $\mathbf{D}'_s$

(c) Projected virtual color image using $\mathbf{D}'_k$

(d) Projected virtual color image using $\mathbf{D}'_f$

Figure 5. Projected virtual color views on the camera image plane of $\mathbb{C}'_{\mathrm{B}}$ using different depth sources.

References

[1] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992. 3

[2] D. C. Brown. Close-range camera calibration. *Photogrammetric Engineering*, 37(8):855–866, 1971. 5

[3] A. Corti, S. Giancola, G. Mainetti, and R. Sala. A metrological characterization of the Kinect V2 time-of-flight camera. *Robotics and Autonomous Systems (RAS)*, 75:584–594, 2016. 1

[4] Ł. Dąbała, M. Ziegler, P. Didyk, F. Zilly, J. Keinert, K. Myszkowski, H.-P. Seidel, P. Rokita, and T. Ritschel. Efficient multi-image correspondences for on-line light field video processing. *Computer Graphics Forum*, 35(7):401–410, 2016. 1

[5] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. A probabilistic approach to ToF and stereo data fusion. In *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, pages 1–8, 2010. 2

[6] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. Probabilistic ToF and stereo data fusion based on mixed pixels measurement models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(11):2260–2272, 2015. 2

[7] C. Dal Mutto, P. Zanuttigh, S. Mattoccia, and G. Cortelazzo. Locally consistent ToF and stereo data fusion. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 598–607, 2012. 2

[8] G. D. Evangelidis, M. Hansard, and R. Horaud. Fusion of range and stereo data for high-resolution scene-modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(11):2178–2192, 2015. 2

[9] C. Fehn, R. De La Barré, and S. Pastoor. Interactive 3-DTV-concepts and key technologies. *Proceedings of the IEEE*, 94(3):524–538, 2006. 1

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2, 3

[11] V. Gandhi, J. Čech, and R. Horaud. High-resolution depth maps based on ToF-stereo fusion. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4742–4749, 2012. 2

[12] Y. Gao, S. Esquivel, R. Koch, M. Ziegler, F. Zilly, and J. Keinert. A novel Kinect V2 registration method for large-displacement environments using camera and scene constraints. In *IEEE International Conference on Image Processing (ICIP)*, pages 997–1001, 2017. 2

[13] Y. Gao, M. Ziegler, F. Zilly, S. Esquivel, and R. Koch. A linear method for recovering the depth of Ultra HD cameras using a Kinect V2 sensor. In *IAPR International Conference on Machine Vision Applications (MVA)*, pages 494–497, 2017. 3

[14] V. Garro, C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. A novel interpolation scheme for range data with side information. In *Conference for Visual Media Production (CVMP)*, pages 52–60, 2009. 5

[15] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *International Journal of Computer Vision (IJCV)*, 94(3):335–360, 2011. 3

[16] M. Hansard, G. Evangelidis, Q. Pelorson, and R. Horaud. Cross-calibration of time-of-flight and colour cameras. *Computer Vision and Image Understanding (CVIU)*, 134:105–115, 2015. 2

[17] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[18] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(2):328–341, 2008. 5

[19] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(11):2121–2133, 2012. 1

[20] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight cameras in computer graphics. *Computer Graphics Forum*, 29(1):141–159, 2010. 5

[21] M. Kraft, M. Nowicki, A. Schmidt, M. Fularz, and P. Skrzypczyński. Toward evaluation of visual navigation algorithms on RGB-D data from the first-and second-generation Kinect. *Machine Vision and Applications (MVA)*, pages 1–14, 2016. 1

[22] V. Lepetit, F. Moreno-Noguer, and P. Fua. EPnP: An accurate O(n) solution to the PnP problem. *International Journal of Computer Vision (IJCV)*, 81(2):155–166, 2009. 2, 3

[23] S. Li, C. Xu, and M. Xie. A robust O(n) solution to the perspective-n-point problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(7):1444–1450, 2012. 2, 3

[24] M. Lindner, I. Schiller, A. Kolb, and R. Koch. Time-of-flight sensor calibration for accurate range sensing. *Computer Vision and Image Understanding (CVIU)*, 114(12):1318–1328, 2010. 4, 5

[25] M. I. A. Lourakis and A. A. Argyros. SBA: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software (TOMS)*, 36(1):2:1–2:30, 2009. 3

[26] K.-L. Low. Linear least-squares optimization for point-to-plane ICP surface registration. *Technical Report, Department of Computer Science, University of North Carolina at Chapel Hill*, TR04-004, 2004. 3

[27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004. 3

[28] C.-P. Lu, G. D. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(6):610–622, 2000. 2, 3

[29] G. Marin, P. Zanuttigh, and S. Mattoccia. Reliable fusion of ToF and stereo depth driven by confidence measures. In *European Conference on Computer Vision (ECCV)*, pages 386–401, 2016. 2

[30] R. Nair, K. Ruhl, F. Lenzen, S. Meister, H. Schäfer, C. S. Garbe, M. Eisemann, M. Magnor, and D. Kondermann. A survey on time-of-flight stereo fusion. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pages 105–127. 2013. 2

[31] H. Sarbolandi, D. Lefloch, and A. Kolb. Kinect range sensing: Structured-light versus time-of-flight Kinect. *Computer Vision and Image Understanding (CVIU)*, 139:1–20, 2015. 1, 4, 5

[32] D. Scaramuzza, A. Martinelli, and R. Siegwart. A toolbox for easily calibrating omnidirectional cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5695–5701, 2006. 5

[33] M. Schmeing and X. Jiang. Depth image based rendering. In *Pattern Recognition, Machine Intelligence and Biometrics*, pages 279–310, 2011. 1

[34] B. Tippetts, D. J. Lee, K. Lillywhite, and J. Archibald. Review of stereo vision algorithms and their suitability for resource-limited systems. *Journal of Real-Time Image Processing*, 11(1):5–25, 2016. 1

[35] T.-C. Wang, A. A. Efros, and R. Ramamoorthi. Depth estimation with occlusion modeling using light-field cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(11):2170–2181, 2016. 1

[36] O. Wasenmüller and D. Stricker. Comparison of Kinect V1 and V2 depth images in terms of accuracy and precision. In *Asian Conference on Computer Vision Workshops (ACCVW)*, pages 34–45, 2016. 1, 3, 4

[37] C. Zach. Robust bundle adjustment revisited. In *European Conference on Computer Vision (ECCV)*, pages 772–787, 2014. 4

[38] 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)*, pages 1–6, 2015. 1

[39] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(11):1330–1334, 2000. 2, 5

[40] J. Zhu, L. Wang, J. Gao, and R. Yang. Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(5):899–909, 2010. 2

[41] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 2

[42] S. Zinger, L. Do, and P. de With. Free-viewpoint depth image based rendering. *Journal of Visual Communication and Image Representation (JVCIR)*, 21(5):533–541, 2010. 1, 4