A HYBRID SYSTEM INTEGRATING CALIBRATION AND REGISTRATION FOR ACCURATE 3D RECONSTRUCTION

Paper 1375

ABSTRACT

With the development of virtual and augmented reality, 3D reconstruction for indoor scenes based on multi-camera systems has become increasingly popular recently. The quality of 3D reconstructions relies on many factors, including the multi-view camera calibration accuracy and the depth estimation quality in each single view. In this paper, we propose a hybrid system which integrates the global camera parameter estimation and 3D model registration instead of traditional pairwise frameworks, to produce high quality 3D models. First, a global bundle adjustment is adopted to increase the accuracy of camera pose estimation. Second, a point cloud registration is used to fuse the inaccurate depth maps to a 3D model. We use the Iterative Closest Point (ICP) algorithm to register the point clouds estimated from different views to minimize the model deviation. The experimental results show the promising effect of our method to produce high-quality models.

Index Terms— Multi-camera system, 3D reconstruction, camera calibration, point cloud registration

1. INTRODUCTION

Indoor 3D reconstruction from multiple views has become more and more prevalent in recent years [1, 2]. High-quality reconstruction has bright future in many applications like telecommunication and VR Games. Compared with the systems based on a single camera, multi-camera systems using RGB cameras have many advantages such as a wider horizon. Meanwhile, to achieve high-quality results, other kinds of sensors are also widely used in the reconstruction system, such as Near Infra-Red (NIR) cameras to estimate depth maps. Hence, the quality of the final reconstruction result depends on many aspects in a multi-camera system, especially on the accuracy of the camera parameters and depth maps.

For a single camera, its intrinsic parameters can be accurately estimated by many internal calibration techniques [3, 4]. For multi-camera calibration, a series of approaches have been proposed as well. One widely used technique for accurate calibration is to use specialized calibration objects, such as planar patterns, rig and so on. Li et al. proposed a specially designed calibration pattern for feature detection and accurate matching [5]. This method works well especially on the systems with a few cameras with small overlapping fields of view. Zhao and Liu proposed an algorithm based on 1D objects for a triple camera system [6]. A 20cm-long stick with 3 markers rotating around a fixed point is used as the calibration object. The algorithm integrates a rank-4 factorization with the standard 1D camera calibration method and is much more convenient than plane-based algorithms. Svoboda et al. proposed a method that only requires a bright-spot object like a laser pointer [7]. Waving the object which can be easily detected in each image through the working space is the only work requested. Kalibr [8] is a free toolbox that solves the multiple camera calibration problem using special 2D barcodes by calibrating neighboring cameras those have overlapping fields of view. However, repeating the sequential calibration process results in accumulated errors. Liu et al. [9] proposed an algorithm for camera parameter adjustment using a checkerboard. They divide a four-camera system into six two-camera subsystems. With a vertically placed checkerboard which is printed on two sides, the corners can be seen in each subsystem and the global adjustment can be done. However, if there are more cameras in the system, it is hard for all the cameras to capture a checkerboard at the same time due to the oblique angle.

Another category of multi-camera calibration techniques is self-calibration, which does not require any specialized objects. However these methods are very sensitive to the textures in the captured scene and usually suffer from the low accuracy. Bundler [10] is a structure-from-motion technique to simultaneously estimate the camera parameters and the 3D point positions from a set of unordered images. It uses SIFT keypoint detector [11] which works well on outdoor scenes but typically fails to find enough point correspondences in indoor cases. Vasconcelos et al. [12] proposed a solution to calibrate a camera with two other calibrated cameras. Bushnevskiy et al. [13] presented a novel approach that enforces constraints arising from the visible epipoles and is especially suitable for dome-like indoor cases.

All these methods use RGB images to calibrate multi-camera systems. Besides of camera parameters, another important influencing factor to the final reconstruction is the depth estimation of each view. Intensive research has been done on the depth estimation problem [14, 15], whereas the error can not be eliminated entirely. Rather than to improve the quality of depth estimation, we use a 3D registration step

to diminish the influence of the depth errors in the fused 3D model. A lot of work has been done in the registration of point sets. Iterative Closest Point (ICP) [16] is the most popular registration method, which performs well with proper initialization. Many registration methods which do not rely on the starting positions of the point clouds have also been proposed [17, 18]. For multiple point-set registration, different from using the sequential pairwise registration strategy, Evangelidis et al. [19] proposed a method that treats all the point clouds on equal terms, and register multiple point clouds globally. However, if two point clouds from different views have the similar shape, for example, the point clouds of the front and the back of a human body, the two point clouds may totally overlap after the registration, which leads to a wrong result. This problem can be solved using the correspondences between the point clouds from the neighboring views by sequential registration.

In this paper, we present a hybrid system integrating both calibration and registration to achieve high-quality reconstruction results. To avoid the accumulative error caused by the repetitive calibration for each pair of cameras, we use a global camera calibration method. Afterwards, we use the point cloud registration method to register the partial point clouds generated from all views. This can effectively reduce the impact of the error in depth estimation and obtain a high-quality 3D reconstruction.

2. ALGORITHM

A multi-camera system is set up to capture objects in an indoor scene, as shown in Fig. 1. K ($K = 8$) camera pods are installed around the working space looking inwards for a full capture. Each camera pod consists of one color camera and two Near Infra-Red (NIR) cameras. A laser pointer is used to produce special patterns. From a pair of images of the projected patterns captured by the two NIR cameras, a depth map D_k can be estimated using the PatchMatch stereo algorithm [15]. There are also several alternative depth cameras or depth estimation algorithm to obtain the depth map of each view. Most of them produce a rough depth map with noise and errors. As Fig. 2 shows, the estimated depth map encounters distortions on the boundary of the human body, and missing data in the head area. How to improve the quality of depth map is beyond the scope of this paper.

Under the pinhole camera assumption, each camera has a group of parameters including the intrinsic matrix K and extrinsic matrix M to project a point in the 3D space to its image plane as

$$z_p \mathbf{x} = \mathbf{K}\mathbf{M}\mathbf{p}, \tag{1}$$

where $\mathbf{p} = (x, y, z, 1)^T$ is the homogeneous coordinate of the 3D point, and $\mathbf{x} = (u, v, 1)^T$ is the homogeneous coordinate of its projected point on the image plane.

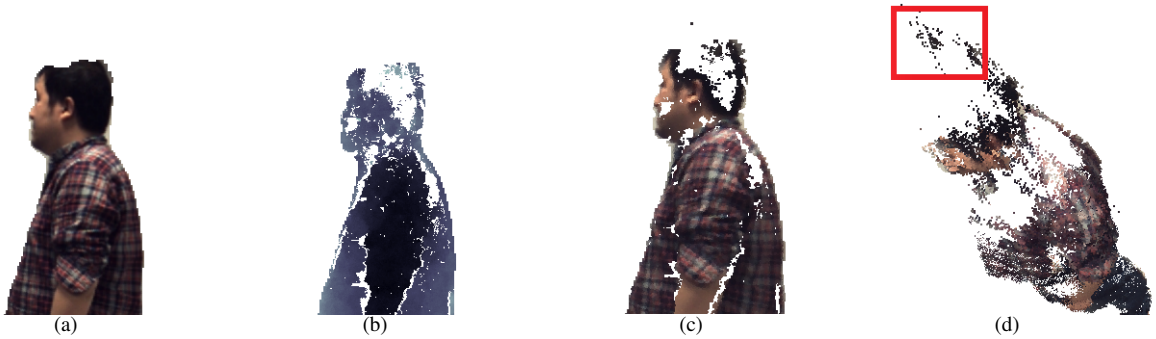
From the depth map estimated by each camera pod, a partial point cloud can be reconstructed by back-projecting each pixel in the depth map into the 3D space according to the parameters of each camera. A 3D model can be obtained by fusing the K point clouds. Ideally, with the accurate depth z_p of each image point $\mathbf{x}$, and accurate camera parameters $\mathbf{K}$ and $\mathbf{M}$, the image pixels captured in different views can be back-projected to the 3D space and well aligned in the 3D space. However, due to the unavoidable error and noises in depth maps, the points cloud obtained in different views deviate a lot even with accurate camera poses. We divide the pipeline into two main parts to reconstruct a high-quality model. The first part is a global camera calibration method to estimate the camera parameters $\{\mathbf{K}_k, \mathbf{M}_k\}_{k=1}^{K}$ as accurate as possible, as described in Sec. 2.1. The second is a registration method to produce 3D point clouds as consistent as possible from inaccurate depth z_p, as described in Sec. 2.2.

Fig. 1. Our multi-camera system with 8 camera pods pointing inwards.

2.1. Global camera calibration

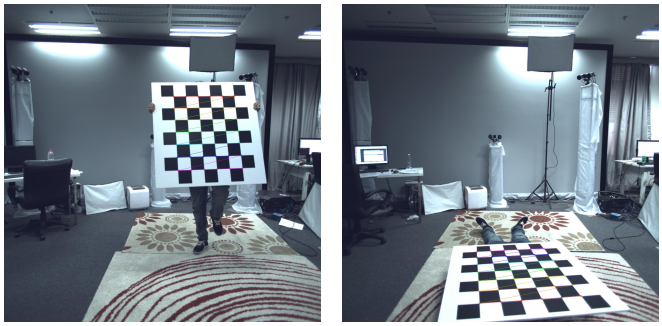
To calibrate the multi-camera system, we follow the common framework that searches for feature point correspondences first and minimizes the re-projection errors. Because of the lack of the features in indoor scenes and the high requirement of the quality of the calibration, we use a 6×7 checkerboard with squares of 117 mm which can be made easily for the accuracy and robustness.

An user holding the checkerboard could move in front of the cameras freely in the working space. At each moment, the checkerboard is simultaneously seen in three or four views usually. Our system is flexible enough because it does not require that the checkerboard must be visible under all points of view. We also use the images with checkerboard on the ground which allows the checkerboard corners can be seen in all the 8 cameras, as shown in Figure 3. This will help to reduce the accumulative error and inconsistence caused by the pairwise calibration. Then the 42 corners on the checkerboard can be detected in each local group. For each time t, denote the group of cameras that can see the checkerboard corners as $\mathcal{C}^t$. The 3D coordinates of the 42 corner points are denoted as $\{\mathbf{p}_i^t\}_{i=1,\ldots,42}$, while their corresponding 2D coordinates detected in each image captured by the camera $c_j \in \mathcal{C}^t$



(a) (b) (c) (d)

Fig. 2. Errors in depth estimation. (a) RGB; (b) Depth; (c) The point cloud reconstructed from RGBD data. (d) Another view of the point cloud. The distortion of the head region (in the red box) is caused by the inaccurate depth estimation.



(a) (b)

Fig. 3. Results of the checkerboard corners detection. (a) Results when the user moves freely. (b) Results when the checkerboard is on the ground.

are $\{\mathbf{x}_{ji}^t\}$. While the user holding the checkerboard moves around, we can capture T groups of checkerboard images and combine them together to globally estimate the camera parameters.

According to the re-projection equation defined in Eq. 1, our calibration method optimizes the camera parameters to minimize the re-projection errors. Considering the T groups of images together, the re-projection error is defined as:

$$E(\mathbf{M}_j, \mathbf{p}_i^t) = \sum_t \sum_{c_j \in \mathcal{C}^t} \sum_i \left(\hat{\mathbf{x}}(\mathbf{K}_j, \mathbf{M}_j, \mathbf{p}_i^t) - \mathbf{x}_{ji}^t\right)^2, \quad (2)$$

where $\hat{\mathbf{x}}(\mathbf{K}_j, \mathbf{M}_j, \mathbf{p}_i)$ is the estimated 2D positions of a 3D point $\mathbf{p}_i^t$ in the image plane of the camera c_j with its parameters $\mathbf{K}_j$ and $\mathbf{M}_j$.

The above optimization could be solved by the Sparse Bundle Adjustment package [20], which uses a Levenberg-Marquardt technique to optimize the parameters iteratively. Inevitably, this algorithm only finds a local optimal. A good initialization is required to get a good results.

Though sequentially calibrating the neighboring cameras leads to accumulated errors, it could provide an initialization for the global optimization. We use the toolbox Kalibr [8] to obtain the intrinsic parameters and the initial extrinsic parameters of neighboring cameras sequentially. Then we estimate the 3D coordinates $\mathbf{p}_{ji}^t$ of each corner from the corresponding 2D coordinates by triangulation. After the initial calibration, we globally optimize the extrinsic parameters and the 3D coordinates of each corner. Typically, $T = 30$ groups produce good enough results.

The re-projection error describes the accuracy of the calibration system. We compute the average re-projection error and compare it to evaluate the effectiveness of our global calibration step. Using the 2D coordinates of the corners detected in the checkerboard images, we can achieve the 3D space coordinates of $\mathbf{p}_i^t$ by triangulation respectively using the camera parameters obtained by Kalibr and our global calibration method (SBA). Then we re-project $\mathbf{p}_i^t$ to each view which the checkerboard can be seen and compute the average re-projection error, as shown in Table 1. We can see that the re-projection error is smaller and consistency using the camera parameters after the global optimization, which proves the effectiveness of our global calibration step.

2.2. Point cloud registration

After the global optimization, we get a set of camera parameters with the minimal re-projection error. However, there are still many problems in the 3D model obtained from directly merging the partial point clouds estimated from different views according to the estimated camera poses, mainly due to the low quality of depth estimation. While our global calibration step provides highly accurate camera poses, the point clouds from neighboring camera pods are close to each other in a coarse level. The misalignment typically occurs at some parts of the human body, such as the head and arms, due to the depth estimation error in theses regions. Therefore, we use the Iterative Closest Point (ICP) algorithm [16] to slightly adjust the rigid transformation between two partial point clouds, and fuse the transformed point clouds to balance out the depth errors.

Table 1. Average re-projection error (pixels) of each camera using different methods. (a) results by Kalibr [8], (b) results by our global calibration method. The re-projection error of (b) is smaller and uniform in each view, while the results of (a) contain the accumulative error and the inconsistence.

Methods	Cam 1	Cam 2	Cam 3	Cam 4	Cam 5	Cam 6	Cam 7	Cam 8
Kalibr	4.54	3.20	11.16	10.74	4.87	6.75	3.64	5.11
SBA	**0.32**	**0.32**	**0.38**	**0.35**	**0.37**	**0.40**	**0.33**	**0.34**

We choose the point cloud of one view as the reference, and compute the transformation of the point clouds from its neighboring views to it using the ICP algorithm. We then merge the point clouds as a new reference. The other views are registered sequentially in a similar way. When all the point clouds of different views are registered, a high-quality 3D model can be achieved, as shown in Fig. 4.

(a) (b)

Fig. 4. A 3D model after ICP registration. (a) The front side of the model. (b) The back side of the model.

3. RESULTS AND DISCUSSION

In order to evaluate the impacts of different components of our hybrid system, a series of experiments are conducted. Firstly, we quantitatively analyze the reconstruction quality using a ground-truth model scanned by an accurate Lidar scanner. Secondly, we show the reconstructed models of human bodies under different poses to demonstrate the robustness of our method.

Reconstruction Quality. The final goal of our system is to achieve a high-quality reconstruction model, we quantitatively compare the results reconstructed by different methods. We generate a ground-truth model by scanning a plaster model of a human head using a 3D laser scanner with the scanning error smaller than 0.04 mm within one meter. The model is placed in our multi-camera system and the RGB and depth images of it can be obtained. We compute the L2 distances of the fused point clouds to the ground-truth mesh for different variants of our method. For each point $\mathbf{p}$ in the reconstructed point cloud, we search for its nearest point $\mathbf{q}$ on the mesh, and compute the distance $\|\mathbf{p}-\mathbf{q}\|$. We divide the distance from 0 mm to 40 mm into 16 intervals and count the total number of the points in each interval. The distribution diagram of the distances between the point clouds to the ground truth using various methods is shown in Fig. 5. We test four variants: (a) the method Kalibr [8] that only uses local calibration of neighboring cameras; (b) Kalibr+ICP, which uses the ICP registration of point clouds reconstructed using the camera parameters estimated by Kalibr; (c) SBA only, which directly merges point clouds after our global calibration using SBA; (d) SBA+ICP, which refines the transformations between point clouds using ICP based on our global camera calibration. As we can see, the number of the points with small distance increases obviously after the optimization and registration step we proposed. We compute the mean and standard deviation of the distances, and show the error bars of the results using various methods in Fig. 6. The reconstructed model combining the global calibration and registration has the smallest error, which proves that the system can produce a high-quality model.

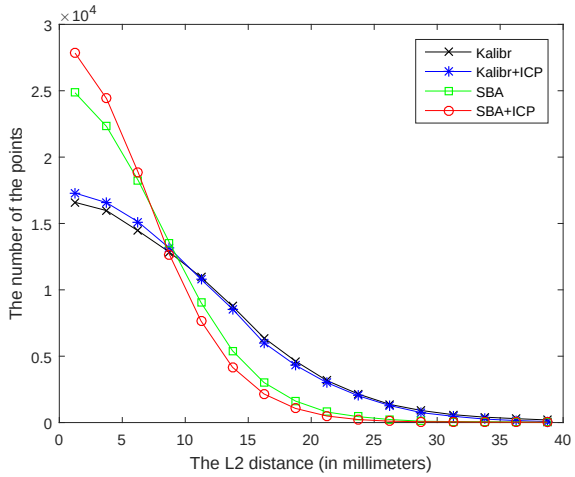


Fig. 5. The distribution of the L2 distances between corresponding points in the point cloud reconstructed using different methods to the ground truth.

More Reconstructed Models. We reconstruct the 3D point clouds of a human body standing in our multi-camera system

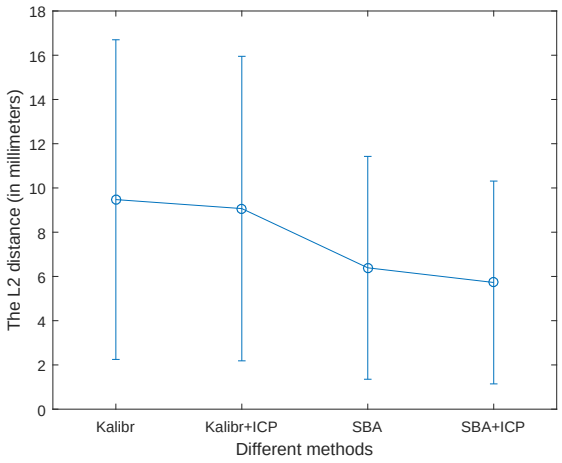
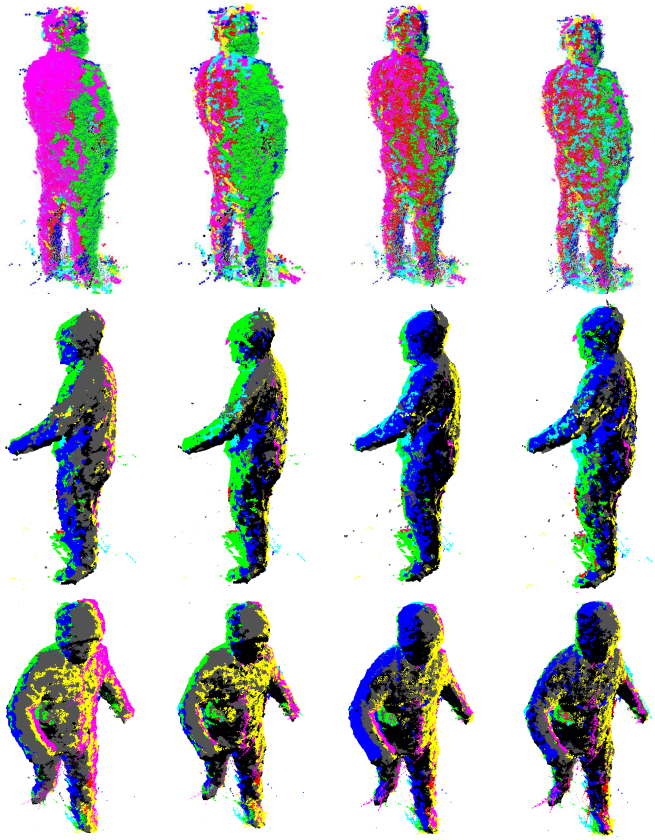


Fig. 6. The error bars of the L2 distances of the point cloud reconstructed using different methods to the ground truth.

using various methods. We mark the point cloud in different views with different colors to show the quality of the fusion results. Fig. 7 shows the reconstruction results in different human pose which demonstrates the effectiveness of our algorithm.

The first row in Fig. 7 shows the back side of a human body model using various methods. The point clouds in result (a) are distinctly separated from other views, as the pink and green views cover the whole surface of the model. Result (b) becomes a little better while the green view still covers the right side of the body. The point clouds align quite well in result (c) with small blemishes, the azure view is inside the model. (d) shows a well-aligned model. The second row in Fig. 7 shows the left side of a human. The grey view in result (a) covers the surface and the black view is nearly inside the model especially on the shoulders and arms. Result (b) becomes a little better in black and grey view, however the green view covers the front of the model. There are similar problems in result (c), as the blue view covers the surface of the arm and head. Result (d) shows a well-aligned model, especially for the leg and arm with less separation between different views. The third row in Fig. 7 shows the front of a human. The grey view covers the most surface of the right arm and the right chest, meanwhile the black view is inside the model in result (a). The point clouds align a little well in result (b), however the yellow view covers the left side of the body, and the blue view is nearly covered by other views. On the contrary, the blue view covers the right side of the body in result (c), especially on the arm and the shoulder. The point clouds more or less separated from other views in these results, while (d) shows a well-aligned model. As we can see, the model reconstructed by our hybrid system has the highest quality.



(a) Kalibr (b) Kalibr+ICP (c) SBA (d) SBA+ICP

Fig. 7. The reconstruction results of a human body in different poses using four methods. We render the point clouds from different views with different colors.

4. CONCLUSION

We present an efficient system integrating a global multi-camera calibration with the 3D registration of point clouds. By a global bundle adjustment, we reduce the accumulative error and the inconsistence caused by the pairwise camera pose estimation, and achieve a set of accurate extrinsic parameters with less re-projection error. With the point cloud registration, the errors in depth estimation are balanced out and a high-quality 3D model is finally obtained. Our calibration algorithm has been tested on both re-projection error and ground truth data. The experimental results has proved that the two steps of our system are both necessary and effective, and more accurate models can be reconstructed using our method.

5. REFERENCES

[1] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim,

Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi, "Fusion4D: Real-time performance capture of challenging scenes," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 114:1–114:13, July 2016.

[2] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi, "Holoportation: Virtual 3D teleportation in real-time," in *UIST*, 2016, pp. 741–754.

[3] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, Nov 2000.

[4] Zhengyou Zhang, "Camera calibration with one-dimensional objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 892–899, July 2004.

[5] B. Li, L. Heng, K. Koser, and M. Pollefeys, "A multiple-camera system calibration toolbox using a feature descriptor-based calibration pattern," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 1301–1307.

[6] Zijian Zhao and Yuncai Liu, "Practical multi-camera calibration algorithm with 1D objects for virtual environments," in *2008 IEEE International Conference on Multimedia and Expo*, June 2008, pp. 1197–1200.

[7] T. Svoboda, D. Martinec, and T. Pajdla, "A convenient multicamera self-calibration for virtual environments," *Presence*, vol. 14, no. 4, pp. 407–422, Aug 2005.

[8] J. Maye, P. Furgale, and R. Siegwart, "Self-supervised calibration for robotic systems," in *2013 IEEE Intelligent Vehicles Symposium (IV)*, June 2013, pp. 473–480.

[9] Jianran Liu, Zaojun Fang, Kun Zhang, and Min Tan, "Algorithm for camera parameter adjustment in multi-camera systems," *Optical Engineering*, vol. 54, no. 10, pp. 104108–104108, 2015.

[10] Noah Snavely, Steven M. Seitz, and Richard Szeliski, "Photo tourism: Exploring photo collections in 3d," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, July 2006.

[11] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[12] Francisco Vasconcelos, João Pedro Barreto, and Edmond Boyer, *A Minimal Solution for Camera Calibration Using Independent Pairwise Correspondences*, pp. 724–737, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[13] A. Bushnevskiy, L. Sorgi, and B. Rosenhahn, "Multi-camera calibration from visible and mirrored epipoles," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3373–3381.

[14] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, 2001, pp. 131–140.

[15] Michael Bleyer, Christoph Rhemann, and Carsten Rother, "Patchmatch stereo - stereo matching with slanted support windows," in *BMVC*, 01 2011, vol. 11, pp. 14.1–14.11.

[16] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, Feb 1992.

[17] Dror Aiger, Niloy J. Mitra, and Daniel Cohen-Or, "4pointss congruent sets for robust pairwise surface registration," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 85:1–85:10, Aug. 2008.

[18] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *2009 IEEE International Conference on Robotics and Automation*, May 2009, pp. 3212–3217.

[19] G.D. Evangelidis, D. Kounades-Bastian, R. Horaud, and Psarakis E.Z., "A generative model for the joint registration of multiple point sets," in *European Conference on Computer Vision (ECCV)*, 2014.

[20] Manolis I. A. Lourakis and Antonis A. Argyros, "Sba: A software package for generic sparse bundle adjustment," *ACM Trans. Math. Softw.*, vol. 36, no. 1, pp. 2:1–2:30, Mar. 2009.