Exploiting the Manhattan-world Assumption for Extrinsic Self-calibration of Multi-modal Sensor Networks

Marcel Brückner* Joachim Denzler
Chair for Computer Vision, Friedrich Schiller University of Jena
Ernst-Abbe-Platz 2, 07743 Jena, Germany
{marcel.brueckner, joachim.denzler}@uni-jena.de

Abstract

Many new applications are enabled by combining a multi-camera system with a Time-of-Flight (ToF) camera, which is able to simultaneously record intensity and depth images. Classical approaches for self-calibration of a multi-camera system fail to calibrate such a system due to the very different image modalities. In addition, the typical environments of multi-camera systems are man-made and consist primary of only low textured objects. However, at the same time they satisfy the Manhattan-world assumption. We formulate the multi-modal sensor network calibration as a Maximum a Posteriori (MAP) problem and solve it by minimizing the corresponding energy function. First we estimate two separate 3D reconstructions of the environment: one using the pan-tilt unit mounted ToF camera and one using the multi-camera system. We exploit the Manhattan-world assumption and estimate multiple initial calibration hypotheses by registering the three dominant orientations of planes. These hypotheses are used as prior knowledge of a subsequent MAP estimation aiming to align edges that are parallel to these dominant directions. To our knowledge, this is the first self-calibration approach that is able to calibrate a ToF camera with a multi-camera system. Quantitative experiments on real data demonstrate the high accuracy of our approach.

1. Introduction

Multi-camera systems can be found in many man-made environments. Various computer vision applications like object tracking or 3D reconstruction use such multi-camera systems. Most of these systems consist of classical CCD cameras recording 2D (color) images. A calibrated multi-camera system can be used to extract 3D information from the camera images. However, this (wide baseline) 3D reconstruction is computationally expensive and works only for textured objects.

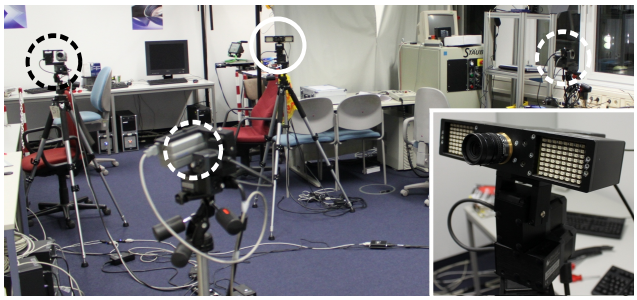

Figure 1. A multi-sensor system consisting of a ToF camera (solid circle and bottom right) and several classical cameras (dashed circles). Each camera is mounted on a pan-tilt unit.

In recent years a new type of camera became increasingly popular: the *Time-of-Flight* (ToF) camera [18] (Figure 1, bottom right). This type of camera uses modulated infrared light to simultaneously record intensity (grayscale) and depth images at a frequency of about 20 Hz. In contrast to stereo cameras it is able to extract depth information even from untextured objects. Another advantage is that the depth is measured in metric units, which endows the correct scale of the 3D reconstruction. Drawbacks of these cameras are their low resolution (200×200 or lower) and that they are not able to record color information. The combination of a ToF camera with a classical multi-camera system overcomes the drawbacks of both. Many computer vision applications benefit from such a *multi-sensor system* (Figure 1). To this end, an accurate calibration between the multi-camera system and the ToF camera is necessary.

State of the art for calibration of such multi-sensor systems are approaches that use a calibration pattern [11, 15, 19] or some other artificial calibration object [4, 10]. From a practical point of view, however, a pure self-calibration is most appealing. Self-calibration in this context means that no artificial landmarks or user interaction are necessary.

*Marcel Brückner would like to thank the Carl Zeiss Foundation (Carl-Zeiss-Stiftung) for supporting his research.

Figure 2. Intensity (left) and depth image (middle) recorded by a ToF camera. Right: Same scene recorded by a CCD camera.

The cameras estimate their rotation and position only from the images they record.

The appearance of an object in the ToF images depends strongly on its material and color. Figure 2 shows an intensity and depth image recorded by a ToF camera and an image of the same scene recorded by a classical CCD camera. Note that the stripes on the sweater on the left are not visible in the ToF intensity image. Another important property is that the depth of dark objects like the black folder in the middle is measured incorrectly. Due to these very different image modalities, the extraction of point correspondences with state of the art methods results in (almost) no correct point correspondences. Hence, point correspondence based approaches for multi-camera calibration [1, 2, 17] fail to calibrate the multi-sensor system.

Another difficulty for point correspondence based approaches is the fact that most multi-sensor systems are placed in man-made environments (*e.g.* building interiors), that primary consist of low textured objects. However, these environments also share the property that most of the surfaces are piecewise planar and aligned to three orthogonal dominant directions. Environments and objects with this property satisfy the so-called Manhattan-world assumption [5]. In this paper we present a method that exploits this assumption to create hypotheses for the calibration between a ToF camera and a calibrated multi-camera system. We then use these hypotheses to formulate the multi-modal sensor network calibration as a MAP problem and solve it by minimizing the corresponding energy function. We assume that the ToF camera is able to record its environment *e.g.* by being mounted on a pan-tilt unit (Figure 1). To our knowledge, this is the first self-calibration approach which is able to calibrate a ToF camera with a multi-camera system.

The Manhattan-world assumption has also been exploited by Furukawa *et al*. [6, 7] for dense 3D reconstruction of planar, non-textured surfaces like buildings or building interiors. Note that even though we are adapting some of their ideas we are aiming to solve a totally different problem. They use a calibrated multi-camera system to estimate a dense 3D reconstruction of planar, non-textured surfaces. In contrast we aim to estimate the transformation between the coordinate systems of a ToF camera and a multi-camera system.

The remainder of this paper is structured as follows. We state the problem and give an overview of our method in Section 2. Section 3 shows how to estimate the initial hypotheses. The MAP estimation of the final calibration is described in Section 4. Section 5 presents our experiments and results. Conclusions are given in Section 6.

2. Problem Statement and Method Overview

Starting point of our approach is a calibrated multi-camera system [1, 2, 17] and a ToF camera. They both have their own coordinate system. We assume that the ToF camera is able to change its point of view to record images from its environment. In our experiments we mount the ToF camera on a pan-tilt unit, but it could also be mounted *e.g.* on some mobile robot. The setup that we have in mind is a multi-sensor system consisting of pan-tilt unit mounted CCD and ToF cameras, similar to the one in Figure 1.

Normally, if one likes to add some camera to an already calibrated multi-camera system, the rotation $\mathbf{R}$ and the translation $\mathbf{t}$ need to be estimated. However, as the ToF camera is able to measure metric distances, we formulate the calibration as a similarity transformation from the coordinate system of the the multi-camera system. This allows us to correct the typically unknown scale of the multi-camera calibration [1, 17]. A similarity transformation

$$\mathbf{S} \stackrel{\text{def}}{=} \begin{pmatrix} a^s\mathbf{R}^s & \mathbf{t}^s \\ \mathbf{0} & 1 \end{pmatrix} \tag{1}$$

consists of a rotation $\mathbf{R}^s$, a translation $\mathbf{t}^s$ and a scale factor a^s. Written in this matrix form it can be used to transform homogeneous 3D coordinates from one coordinate system to the other.

Our approach aims to estimate this similarity transformation. To this end we exploit the Manhattan-world assumption [5]. It assumes that most surfaces in the surrounding world of the camera are piecewise planar and aligned to three orthogonal dominant directions. This assumption is satisfied for the typical environment of multi-camera systems like building interiors or urban scenes.

Figure 3 gives a schematic overview of our proposed method. It consists of two steps. First we estimate two 3D reconstructions of the complete environment: one using the multi-camera system and another one using the ToF camera (Section 3.1). For each of these 3D reconstructions we estimate the three orthogonal dominant directions by computing a histogram over the surface normal directions (Section 3.2). Given these dominant directions we can calculate hypotheses for the similarity transformation between the two coordinate systems.

In the second step we use these hypotheses as prior for a MAP estimation (Section 4) that aims to align 3D points of the multi-camera system, that are on edges parallel to the dominant directions, with edges in the ToF intensity images, that are parallel to the same dominant direction.

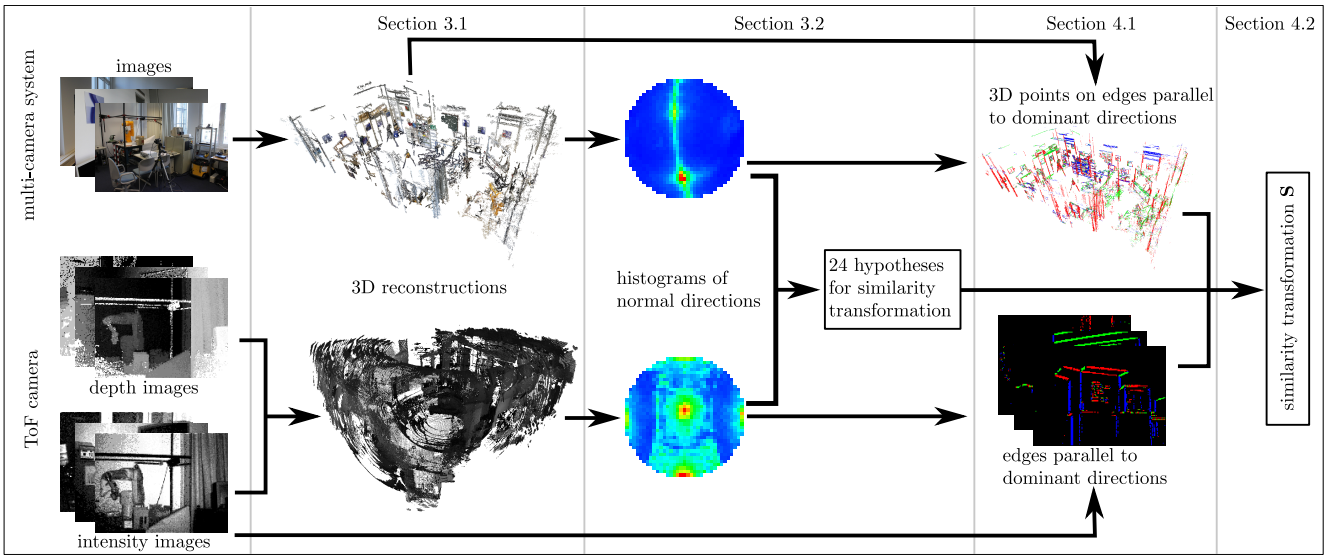

Figure 3. Overview of the proposed method. Each step is explained in the indicated Section.

Note that we use a superscript ^{f} for all ToF camera related terms and a superscript ^{c} for terms related to the multi-camera system throughout the paper.

3. Estimating the Initial Candidates

3.1. 3D Reconstruction of the Environment

The input of our approach are n intensity images $\mathcal{I}^f \stackrel{\text{def}}{=} \{I_1^f, \ldots, I_i^f, \ldots, I_n^f\}$ and n depth images $\mathcal{D}^f \stackrel{\text{def}}{=} \{D_1^f, \ldots, D_i^f, \ldots, D_n^f\}$ from the environment recorded by the ToF camera. For each of these images the pinhole matrix $\mathbf{K}_i^f$ and the relative pose consisting of a rotation $\mathbf{R}_i^f$ and translation $\mathbf{t}_i^f$ are assumed to be known. With these data it is possible to estimate a 3D point cloud. The position of a 3D point is calculated by

$$\mathbf{X}^f \stackrel{\text{def}}{=} \frac{D_i^f\left(\mathbf{x}^f\right)}{\left\|\mathbf{K}_i^{f\,-1}\mathbf{x}^f\right\|_2}\mathbf{R}_i^{f\,\mathrm{T}}\mathbf{K}_i^{f\,-1}\mathbf{x}^f - \mathbf{R}_i^{f\,\mathrm{T}}\mathbf{t}_i^f\,, \quad (2)$$

where $D_i^f(\mathbf{x}^f)$ is the measured depth in the i-th depth image at image coordinate $\mathbf{x}^f$. This 3D point estimation is done for each image coordinate $\mathbf{x}^f$ in each depth image $D_i^f \in \mathcal{D}^f$. We estimate the surface normal $\mathbf{n}^f$ of each 3D point $\mathbf{X}^f$ by local plane fitting.

From the multi-camera system we get m images $\mathcal{I}^c = \{I_1^c, \ldots, I_m^c\}$ with known relative poses and pinhole matrices. Given these data we estimate a 3D reconstruction using the multi-view stereo approach of Furukawa and Ponce [9]. This results in a set of 3D points $\mathcal{P}^c$. The approach also estimates the surface normal $\mathbf{n}^c$ of each 3D point $\mathbf{X}^c \in \mathcal{P}^c$.

The modalities of the two 3D reconstructions are very different. The images in Figure 3 give an impression of these. While the ToF reconstruction is very dense but quite noisy, the reconstruction using the multi-camera system is very accurate but since 3D points can only be estimated near textured objects it is also quite sparse.

3.2. Estimating the Dominant Directions

The surfaces in a Manhattan-world are aligned to three orthogonal dominant directions. For each of our 3D reconstructions we want to estimate these directions. Similar to [6], we first build two histograms h^c and h^f of surface normal directions $\mathbf{n}$ (with $\|\mathbf{n}\|_2 = 1$) over a unit hemisphere. Since we are only interested in the directions of the surface normals, we do not need a histogram over a complete sphere. Hence the normal directions $-\mathbf{n}$ and $\mathbf{n}$ are mapped to the same histogram bin.

Note, that the reconstructions of many realistic scenes consist only of two orthogonal dominant directions (see example histograms in Figure 3). This is why we search the histograms for the *two* unit normals $\hat{\mathbf{n}}_q, \hat{\mathbf{n}}_r$ that

$$\underset{\mathbf{n}_q,\mathbf{n}_r}{\operatorname{argmax}} \left(h\left(\mathbf{n}_q\right) + h\left(\mathbf{n}_r\right)\right) \text{ with } \mathbf{n}_q^{\mathrm{T}}\mathbf{n}_r = 0\,. \quad (3)$$

These correspond to the three orthogonal dominant directions $\mathbf{d}_1 \stackrel{\text{def}}{=} \hat{\mathbf{n}}_q$, $\mathbf{d}_2 \stackrel{\text{def}}{=} \hat{\mathbf{n}}_r$ and $\mathbf{d}_3 \stackrel{\text{def}}{=} \hat{\mathbf{n}}_q \times \hat{\mathbf{n}}_r$. At the end of this procedure, we have the three dominant directions of the ToF camera reconstruction $\mathcal{V}^f \stackrel{\text{def}}{=} \{\mathbf{d}_1^f, \mathbf{d}_2^f, \mathbf{d}_3^f\}$ and of the multi-camera reconstruction $\mathcal{V}^c \stackrel{\text{def}}{=} \{\mathbf{d}_1^c, \mathbf{d}_2^c, \mathbf{d}_3^c\}$.

The estimated directions are basically the same but estimated in different coordinate systems. There are exactly 24 rotations $\{\tilde{\mathbf{R}}_1^s \ldots, \tilde{\mathbf{R}}_k^s, \ldots, \tilde{\mathbf{R}}_{24}^s\}$ that align the three orthogonal directions of one set with the directions of the other set (also considering negative directions).

We now form a set of 24 initial similarity transformations $\{\tilde{\mathbf{S}}_1, \ldots, \tilde{\mathbf{S}}_k, \ldots, \tilde{\mathbf{S}}_{24}\}$ using the rotations $\tilde{\mathbf{R}}_k^s$ and calculating the translation and scale by minimizing the distance between the depth measured by the ToF camera and the depth of the transformed 3D point set $\mathcal{P}^c$

$$\underset{\mathbf{t}^s, a^s}{\operatorname{argmin}} \sum_{\mathbf{X}^c \in \mathcal{P}^c} \min_i \left| D_i^f\left(\pi_i^{\mathbf{S}}\left(\mathbf{X}^c\right)\right) - \left\| c_i^{\mathbf{S}}\left(\mathbf{X}^c\right)\right\|_2 \right| . \quad (4)$$

The function

$$c_i^{\mathbf{S}}\left(\mathbf{X}^c\right) \stackrel{\text{def}}{=} \mathbf{R}_i^f\left(a^s \mathbf{R}^s \mathbf{X}^c + \mathbf{t}^s\right) + \mathbf{t}_i^f \quad (5)$$

transforms a 3D point $\mathbf{X}^c$ from the coordinate system of the multi-camera system to the coordinate system of the i-th image of the ToF camera using the similarity transformation $\mathbf{S}$. The function

$$\pi_i^{\mathbf{S}}\left(\mathbf{X}^c\right) \stackrel{\text{def}}{=} \mathbf{K}_i^f c_i^{\mathbf{S}}\left(\mathbf{X}^c\right) \quad (6)$$

projects a 3D point $\mathbf{X}^c$ from the multi-camera coordinate system into the i-th image of the ToF camera using the similarity transformation $\mathbf{S}$. We use the downhill simplex method [13] for the optimization of (4).

One might assume that an energy minimization similar to (4) (optimizing also the rotation) should suffice to estimate the correct similarity transformation. Unfortunately this is not the case. This is due to the different modalities of the two 3D data sets. The 3D reconstruction obtained by the multi-camera system consists only of few 3D points on planar surfaces, most points are along object edges. Exactly at these object edges the depth measurement of ToF cameras is very inaccurate. The accuracy also suffers from a systematic depth measurement error of the ToF camera [11, 15]. Furthermore the cost function of (4) is not reliable for deciding which of the similarity transformations $\tilde{\mathbf{S}}_k$ is the correct one.

4. Estimating the Similarity Transformation

We avoid using the depth images of the ToF camera in our final step. Instead we use the 24 similarity transformation hypotheses as prior for a Maximum a Posteriori (MAP) estimation. This MAP estimation aims to align 3D points of the multi-camera system, which are on edges parallel to the dominant directions, to edges in the ToF intensity images, which are parallel to the same dominant direction.

4.1. Edges Parallel to Dominant Directions

Similar to [6], we use the estimated dominant directions $\mathcal{V}^f$ and $\mathcal{V}^c$ to find edges in the camera images $\mathcal{I}^f$ and $\mathcal{I}^c$ that are parallel to one of these directions. We search the images for image coordinates lying on image edges [3] that are parallel to one of the dominant directions $\mathbf{d} \in \mathcal{V}$. An image edge at image point $\mathbf{x}$ is parallel to a dominant direction $\mathbf{d}$ if it passes through $\mathbf{x}$ and the vanishing point $\mathbf{v_d}$ of the direction $\mathbf{d}$. All the image points $\mathbf{x}_{i,\mathbf{d}}^f$ from the ToF intensity images that fulfill these constraints for the direction $\mathbf{d}$ are stored in the set $\mathcal{J}_{\mathbf{d}}^f$. For each of these image points the index of its image i is known. The set $\mathcal{J}_{\mathcal{V}}^f \stackrel{\text{def}}{=} \{\mathcal{J}_{\mathbf{d}_1}^f, \mathcal{J}_{\mathbf{d}_2}^f, \mathcal{J}_{\mathbf{d}_3}^f\}$ contains the image point sets of the three directions. Figure 4 shows a ToF intensity image and the extracted edges. Each edge direction is colored differently.

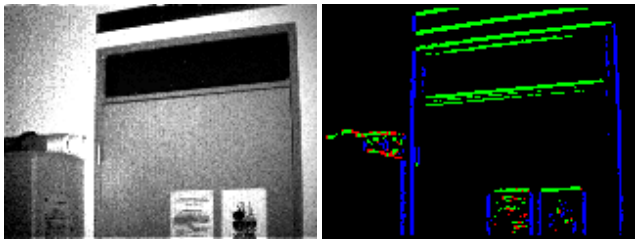

Figure 4. Intensity image of the ToF camera and the resulting edge image. Each color encodes a different dominant direction.

The same procedure is repeated for the images of the multi-camera system $\mathcal{I}^c$. But instead of storing the image points that lie on an edge parallel to one of the dominant directions $\mathbf{d}$, we store the 3D points, that are projected to one of these edges, in the point set $\mathcal{P}_{\mathbf{d}}^c$. The set $\mathcal{P}_{\mathcal{V}}^c \stackrel{\text{def}}{=} \{\mathcal{P}_{\mathbf{d}_1}^c, \mathcal{P}_{\mathbf{d}_2}^c, \mathcal{P}_{\mathbf{d}_3}^c\}$ contains the 3D point sets of the three directions.

4.2. MAP Estimation

Given $\mathcal{J}_{\mathcal{V}}^f$ and $\mathcal{P}_{\mathcal{V}}^c$ we want to estimate the similarity transformation $\mathbf{S}$. In order to increase the robustness against noise and outliers we formulate a Maximum a Posteriori (MAP) problem

$$\underset{\mathbf{S}}{\operatorname{argmax}}\, p\left(\mathbf{S} \mid \mathcal{J}_{\mathcal{V}}^f, \mathcal{P}_{\mathcal{V}}^c\right) \sim p\left(\mathcal{J}_{\mathcal{V}}^f, \mathcal{P}_{\mathcal{V}}^c \mid \mathbf{S}\right) p\left(\mathbf{S}\right) , \quad (7)$$

where $p(\mathcal{J}_{\mathcal{V}}^f, \mathcal{P}_{\mathcal{V}}^c \mid \mathbf{S})$ is the likelihood and $p\left(\mathbf{S}\right)$ is the prior. We define the likelihood as

$$p\left(\mathcal{J}_{\mathcal{V}}^f, \mathcal{P}_{\mathcal{V}}^c \mid \mathbf{S}\right) \stackrel{\text{def}}{\sim} \quad (8)$$
$$\prod_{\mathbf{d} \in \mathcal{V}} \prod_{\mathbf{X}_{\mathbf{d}}^c \in \mathcal{P}_{\mathbf{d}}^c} \max_{\mathbf{x}_{i,\mathbf{d}}^f \in \mathcal{J}_{\mathbf{d}}^f} \mathrm{e}^{-\lambda d_f\left(\mathbf{x}_{i,\mathbf{d}}^f, \mathbf{X}_{\mathbf{d}}^c\right)}$$

where λ is the parameter of the exponential distribution and

$$d_f\left(\mathbf{x}_{i,\mathbf{d}}^f, \mathbf{X}_{\mathbf{d}}^c\right) \stackrel{\text{def}}{=} \min\left(d_c, \left\|\mathbf{x}_{i,\mathbf{d}}^f - \pi_i^{\mathbf{S}}\left(\mathbf{X}_{\mathbf{d}}^c\right)\right\|_2\right) \quad (9)$$

is the minimum between the length of the image diagonal d_c in pixels and the Euclidean distance between the projection of $\mathbf{X}_{\mathbf{d}}^c$ and the image point $\mathbf{x}_{i,\mathbf{d}}^f$, both lying on an edge parallel to the same dominant direction $\mathbf{d}$. This distance function increases the robustness against outliers. The likelihood (8) is high if each 3D point is projected on an image

Figure 5. Example images of the two rooms used for the experiments: a low textured office (left) and a laboratory with several occlusions and ambiguities (right).

edge with the same dominant direction. Note that the assignment of the dominant directions is given by the rotation of the similarity transformation.

The prior should be high if the distance of $\mathbf{S}$ to one of the initial 24 similarity transformations $\tilde{\mathbf{S}}_k$ is small. The distance between two similarity transformations is actually defined by three distances: the rotation distance $d_{\mathbf{R}}(\mathbf{S}, \tilde{\mathbf{S}}_k)$ in degree, the translation distance $d_{\mathbf{t}}(\mathbf{S}, \tilde{\mathbf{S}}_k)$ in meters and the relative scale difference $d_a(\mathbf{S}, \tilde{\mathbf{S}}_k)$. We model the prior

$$p(\mathbf{S}) \stackrel{\text{def}}{=} \min_k e^{-\lambda_{\mathbf{R}} d_{\mathbf{R}}(\mathbf{S}, \tilde{\mathbf{S}}_k)} e^{-\lambda_{\mathbf{t}} d_{\mathbf{t}}(\mathbf{S}, \tilde{\mathbf{S}}_k)} e^{-\lambda_a d_a(\mathbf{S}, \tilde{\mathbf{S}}_k)} \tag{10}$$

as the product of three exponential distributions, where $\lambda_{\mathbf{R}}$, $\lambda_{\mathbf{t}}$ and λ_a are the parameters of these distributions.

In order to find the optimum of this MAP estimation, we initialize an optimization at each of the initial similarity transformations. The optimization is done using the downhill simplex method [13]. The similarity transformation with the highest probability is the final MAP estimate.

5. Experiments and Results

5.1. Experimental Setup

In our experiments we use a PMDTechnologies PMD[vision] 19K ToF camera (resolution 160×120) that is mounted on a Directed Perception PTU-46-17.5 pan-tilt unit (Figure 1, bottom right). The poses $\mathbf{R}_i^f, \mathbf{t}_i^f$ of these ToF images are provided by the pan-tilt unit.

The CCD images are recorded with two statically mounted AVT Pike cameras (resolution 1388×1038) and a handheld Canon 500D camera (resolution 2352×1568). For the calibration of the CCD images we use the Bundler software of Snavely [16]. 3D reconstruction is done using the PMVS software of Furukawa and Ponce [8]. For the ground truth calibration between the two static cameras and the ToF camera we use the calibration pattern based MultiCamera-Calibration software of Schiller [14]. This software is also used to estimate the intrinsic parameters of the ToF camera. We use the point correspondences extracted from the calibration pattern to calculate the reprojection error of our calibration. The calibration errors presented in Section 5.2 are between the ToF camera and the two static cameras.

Our method is tested on a total of 12 calibrations using different camera setups and two different rooms. Figure 5 shows some example images of the two rooms. The rooms are typical office/laboratory environments. They differ in size, inventory and amount of texture. Depending on the camera setup $90 - 160$ images are recorded, about half of these are ToF images.

Note, that in none of our experiments the ToF and multi-camera 3D reconstructions cover the *complete* or exactly the *same* part of the room. However, the 3D reconstructions need to provide enough information to estimate the dominant directions. This is why in our experiments at least two walls of the room are included in the 3D reconstructions.

In our experiments we use 1005 bins for the normal histograms and the parameter of the exponential distribution (8) is set to $\lambda = 1.0$. The three parameters of the prior (10) are set to $\lambda_{\mathbf{R}} = 0.01$, $\lambda_{\mathbf{t}} = 0.025$ and $\lambda_a = 10.0$. The choice of these parameters is not crucial as additional experiments show.

5.2. Results

We present the results of our experiments in Figure 6 using box plots. The box depicts the 0.25 and 0.75 quantiles, the line in the middle is the median, the dashed lines represent the 1.5 interquartile range, and crosses are outliers. For further details please refer to [12]. The plot shows the rotation and translation direction error in degree, the position error in millimeters and the reprojection error of the ToF camera and the two static CCD cameras in pixels. Note, that the reprojection error is calculated using the estimated calibration and point correspondences which are extracted from a calibration pattern.

We achieve a median error of 0.87 degree for th rotation and 57 mm for the position of the ToF camera relative to the multi-camera system. Particularly the achieved median ToF reprojection error of 1.23 pixels is very low considering the difficult image modalities of this camera. However, many calibration pattern based methods [11, 15] also estimate a model to correct systematic errors in the depth measurement of the ToF camera. A focus of our future research will be to estimate such a model using our calibration.

A qualitative impression of the calibration is given in Figure 7: the images on the left and in the middle show the 3D reconstruction using the multi-camera system and the ToF camera, respectively. The 3D reconstruction in the right image uses the estimated calibration to combine the dense 3D point cloud of the ToF camera with the color images of the multi-camera system. The white stripes in the two 3D point clouds on the right result from areas where the ToF camera did not record any images.

In the current OpenMP implementation, the calibration takes about 15 minutes on an off-the-shelf quad-core CPU.

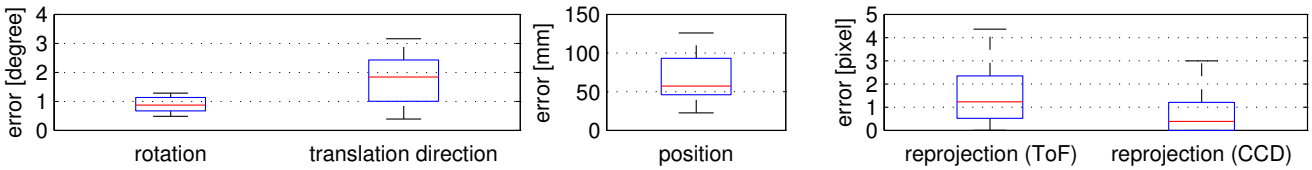

Figure 6. Left and center: The calibration error of the ToF camera relative to the multi-camera system. Right: Separate reprojection errors for the ToF camera and the static CCD cameras of the multi-camera system.

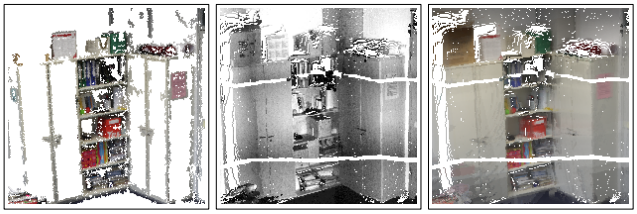

Figure 7. 3D point clouds of the same scene using the multi-camera system (left), the ToF camera (middle) and the calibrated multi-sensor system (right).

Since many parts of our approach can run in parallel, a GPU implementation will improve the runtime significantly.

6. Conclusions and Future Work

We presented a novel extrinsic self-calibration method for calibrating a Time-of-Flight (ToF) camera with a calibrated multi-camera system. Such a calibrated multi-sensor system combines the advantages of a ToF camera, which is able to record depth information in real-time, and a multi-camera system, which records high resolution color images. Our approach does not need any artificial calibration object or user interaction. Instead we exploit the Manhattan-world assumption. Even though this assumption limits the application areas of our approach, it is satisfied by most of the environments of typical multi-camera system applications. To our knowledge this is the first self-calibration approach for such a multi-sensor system. Although the individual steps in our approach seem straightforward the effective combination and MAP based calibration lead to a high accuracy in different experiments covering environments with low texture, occlusions and ambiguities. We achieved a median error of $0.87°$ for the rotation, $57\,\text{mm}$ for the position and 1.23 pixels for the reprojection of the ToF camera. In our future work we aim to estimate a model to correct systematic errors in the depth measurement of the ToF camera.

References

[1] F. Bajramovic and J. Denzler. Global uncertainty-based selection of relative poses for multi camera calibration. In *BMVC*, volume 2, pages 745–754, 2008. 2

[2] M. Brückner and J. Denzler. Active self-calibration of multi-camera systems. In *DAGM*, pages 31–40, 2010. 2

[3] J. Canny. A computational approach to edge detection. *PAMI*, 8(6):679–698, 1986. 4

[4] X. Chen, J. Davis, and P. Slusallek. Wide area camera calibration using virtual calibration objects. In *CVPR*, volume 2, pages 520–527, 2000. 1

[5] J. Coughlan and A. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *ICCV*, volume 2, pages 941–947, 1999. 2

[6] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Manhattan-world stereo. In *CVPR*, pages 1422–1429, 2009. 2, 3, 4

[7] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *ICCV*, pages 80–87, 2009. 2

[8] Y. Furukawa and J. Ponce. Patch-based multi-view stereo software. http://grail.cs.washington.edu/software/pmvs/. 5

[9] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 32(8):1362–1376, 2010. 3

[10] L. Guan and M. Pollefeys. A Unified Approach to Calibrate a Network of Camcorders and ToF Cameras. In *ECCV Workshop M2SFA2*, 2008. 1

[11] M. Lindner and A. Kolb. Lateral and depth calibration of pmd-distance sensors. In *Advances in Visual Computing*, volume 2, pages 524–533, 2006. 1, 4, 5

[12] R. McGill, J. Tukey, and W. A. Larsen. Variations of Boxplots. *The American Statistician*, 32:12–16, 1978. 5

[13] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965. 4, 5

[14] I. Schiller. http://www.mip.informatik.uni-kiel.de/tiki-index.php?page=Calibration. 5

[15] I. Schiller, C. Beder, and R. Koch. Calibration of a pmd-camera using a planar calibration pattern together with a multi-camera setup. In *ISPRS*, pages 297–302, 2008. 1, 4, 5

[16] N. Snavely. Bundler: SfM for Unordered Image Collections. http://phototour.cs.washington.edu/bundler/. 5

[17] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. In *SIGGRAPH*, pages 835–846, 2006. 2

[18] Z. Xu, R. Schwarte, H. Heinol, B. Buxbaum, and T. Ringbeck. Smart pixel-photonic mixer device (pmd) new system concept of a 3d-imaging camera-on-a-chip. In *M2VIP*, pages 259–264, 1998. 1

[19] Z. Zhang. Flexible Camera Calibration by Viewing a Plane from Unknown Orientations. In *ICCV*, pages 666–673, 1999. 1