Vision, Modeling, and Visualization (2011)
Peter Eisert, Konrad Polthier, and Joachim Hornegger (Eds.)

Extrinsic Self-Calibration of Time-of-Flight Cameras using a Combination of 3D and Intensity Descriptors

J. Schmidt and M. Brückner[†] and J. Denzler

Chair for Computer Vision, Friedrich Schiller University of Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany
Email: {joern.schmidt, marcel.brueckner, joachim.denzler}@uni-jena.de

Abstract

Time-of-Flight (ToF) cameras are able to simultaneously record intensity and depth images at a high frequency. Many applications require images that are recorded from different viewpoints. In order to consolidate the recorded data into a common coordinate system, the extrinsic calibration between the cameras needs to be known. From a practical point of view this calibration should be accomplished without any user interaction or artificial calibration objects. Classical approaches for extrinsic self-calibration fail to extract correct point correspondences and do not exploit the important information provided by the depth images. In this paper we discuss the characteristics of extrinsic ToF camera calibration and present a descriptor combination for the extraction of 3D point correspondences. Several experiments on real data demonstrate the robustness and high accuracy of our approach. Our method outperforms the state-of-the-art approach for point correspondence extraction in classical camera images.

Categories and Subject Descriptors (according to ACM CCS): I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Range data

1. Introduction

In the recent years several new camera technologies for real-time acquisition of depth data have been presented. The most prominent example is Microsoft's Kinect camera which computes the depth by projecting a special infrared pattern. Another technology is the Time-of-Flight (ToF) camera which uses modulated near-infrared light to measure the depth based on the Time-of-Flight principle [XSH*98]. Many different research areas benefit from these cameras e.g. robot navigation [PMS*08], medical image processing [SPH08] or computer graphic topics like light fields or augmented/mixed reality as motivated by Kolb et al. [KBKL10].

For many applications a single camera does not suffice. Instead the object or scene needs to be recorded from several different viewpoints. In order to establish a relationship between the intensity and depth images from the different viewpoints, an accurate extrinsic calibration of the cameras is necessary. This calibration describes the relative camera rotations and translations (the relative poses) and enables the transformation of points between the different camera coordinate systems.

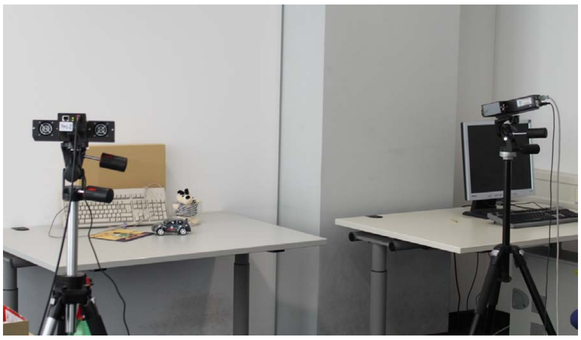

Figure 1: *Two static ToF cameras in a wide-baseline setup. Classical approaches for extrinsic self-calibration fail to calibrate such a configuration due to the low resolution and high noise of the ToF images.*

In the case of moving objects or dynamic scenes, these

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

DOI: 10.2312/PE/VMV/VMV11/269-276

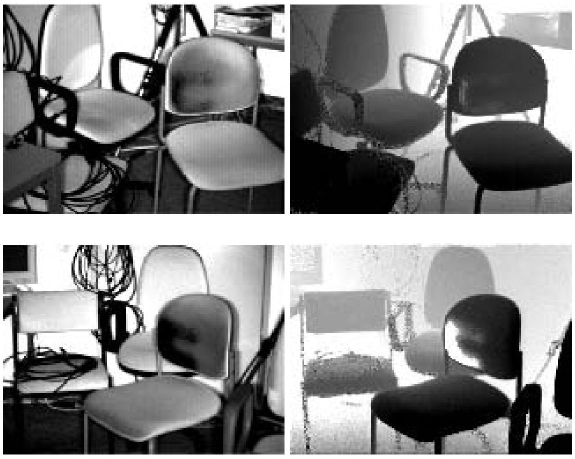

Figure 2: *The intensity (left column) and depth images (right column) recorded by two cameras in a wide-baseline setup.*

images need to be recorded simultaneously which requires the use of multiple cameras. The advantage of ToF cameras is that several of these can operate simultaneously without affecting one another if each camera uses a different modulation frequency. This is why we will focus on ToF cameras in this work. However, the described approach is also applicable to other types of cameras that are able to record depth and intensity images. A multi-camera setup consisting of two ToF cameras is shown in Figure 1.

Classical approaches for extrinsic camera calibration use several images of a calibration pattern [Zha99, SBK08], or track a moving LED in a dark room [CDS00, SHVG02] or some other easily detectable object [GP08] to establish 2D or 3D point correspondences. These approaches work also for ToF cameras. However, from a practical point of view a pure self-calibration is much more appealing. Self-calibration in this context means that no artificial calibration objects or any user interaction are necessary. Instead, the cameras estimate their relative orientation and position only from the images that they record from the scene. Extrinsic self-calibration approaches for classical cameras extract 2D point correspondences, e.g. using SIFT [Low04]. Based on these the relative orientation and translation up to scale can be estimated with methods like the 5-point algorithm [Nis04]. There exist several reasons why these approaches are inappropriate for the extrinsic self-calibration of ToF cameras. The low resolution and high proportion of image noise complicates the extraction of correct point correspondences. Especially in camera setups like in Figure 2 (almost) no correct point correspondence can be extracted. The cameras in this example are in a wide-baseline setup which means that the distance between the cameras is quite big in relation to the scene distance. Another problem is that approaches like the 5-point algorithm [Nis04] estimate the translation only up to scale since 2D point correspondences do not offer enough information to estimate this scale. For the transformation of depth measurements between the different camera coordinate systems, however, the correct translation scale is inevitable.

The acquisition of depth images at high frequency is also of special interest for the robot navigation. Hence several approaches can be found in the literature that cover the topic of estimating the relative pose between the different images of a moving ToF camera. All of these approaches assume a small baseline, i.e. the movement of the cameras between the images is relatively small. Beder et al. [BSK08] present a maximum likelihood approach that estimates the camera motion directly from the depth images. Swadzba et al. [SLP*07] use KLT tracking to establish 3D point correspondences between consecutive images of a moving ToF camera. From these the relative poses are estimated and refined using an iterative closest point (ICP) approach [BM92]. May et al. [MDH*09] compare and benchmark several different approaches for registering small baseline ToF data. Furthermore they suggest an extension to the iterative closest point algorithm that increases the robustness under restricted field of view and under larger displacements. Huhle et al. [HJS08] combine the information of three calibrated sensors (ToF camera, color camera, and inertia sensor) to robustly estimate the relative pose of the moving multi-sensor system.

The remainder of this paper is structured as follows. We will first pay attention to the characteristics of multiple simultaneously operating ToF cameras and how to extract different types of 3D data from the depth measurement (Section 2). In the thereafter following Section 3 we will specify the extrinsic calibration between ToF cameras and how to estimate it from 3D point correspondences. The extraction of these point correspondences is then described in Section 4 where we present our combination of a 3D and an intensity descriptor. The results of various experiments on real data are presented in Section 5. The paper ends in Section 6 with conclusions and problems for future work

2. Time-of-Flight Cameras

Time-of-Flight (ToF) cameras emit modulated near-infrared light to acquire simultaneously three types of images: an intensity image $I(\mathbf{x})$, a depth image $D(\mathbf{x})$ and an amplitude image $A(\mathbf{x})$. The amplitude image provides information about the reliability of the single measurements. Swadzba et al. [SLP*07] calculate the mean amplitude and define a threshold relative to this to reject pixel positions $\mathbf{x}$ with a bad amplitude $A(\mathbf{x})$.

2.1. Configuring Multiple Time-of-Flight Cameras

The image acquisition of the cameras is disturbed if two or more ToF cameras simultaneously emit infrared light with the same modulation frequency f. Consequently each camera needs to operate on a different modulation frequency. Note that the modulation frequency limits the maximum depth that can be measured unambiguously [XSH*98]. Hence the possible modulation frequencies are limited by the application of the camera.

Another important parameter of a ToF camera is the integration time t which specifies the sensor's allocated time for collecting photons. An inappropriate integration time will result in a bad signal-to-noise ratio. Lange [Lan00] describes these physical relations and presents a measure for the inaccuracy of the depth measurement at pixel $\mathbf{x}$

$$e(\mathbf{x}) \overset{\text{def}}{=} \frac{c}{4f\sqrt{8}} \frac{\sqrt{I_t(\mathbf{x})}}{A_t(\mathbf{x})} \quad , \tag{1}$$

where c is the light speed, and $I_t(\mathbf{x})$ and $A_t(\mathbf{x})$ are the intensity and amplitude images recorded with integration time t. We find the optimal integration time

$$\hat{t} \overset{\text{def}}{=} \operatorname*{argmin}_t \sum_{\mathbf{x}} e(\mathbf{x}) \tag{2}$$

by minimizing the accumulated depth measurement inaccuracy for all pixels of the ToF sensor. Note that the integration time directly affects the frame rate of the camera. Hence the application of the ToF camera might constrain the possible integration times. However, changing the integration time does not affect the relative camera pose. Hence we use $\hat{t}$ during the extrinsic calibration.

2.2. Point Cloud Estimation and Surface Triangulation

If the intrinsic calibration of a ToF camera is known [Zha99, SBK08], it is possible to calculate for each homogeneous image point $\mathbf{x} \in \mathbb{P}^2$ the position of the 3D point

$$\mathbf{X} \overset{\text{def}}{=} \frac{D(\mathbf{x})}{\left\|\mathbf{K}^{-1}\mathbf{x}\right\|_2} \mathbf{K}^{-1}\mathbf{x} \quad , \tag{3}$$

where $D(\mathbf{x})$ is the depth measurement at pixel position $\mathbf{x}$ and $\mathbf{K}$ is the pinhole matrix [HZ03]. Since the extracted 3D points correspond directly to an image coordinate, we can also assign an intensity value $I(\mathbf{X}) \overset{\text{def}}{=} I(\mathbf{x})$ to each 3D point.

In order to build a surface triangulation from this 3D point cloud we exploit the regular grid of the pixels in the ToF sensor. Of course this simple method could also be exchanged by a more sophisticated approach from the literature. We obtain the surface triangulation

$$\mathcal{S} = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\} \tag{4}$$

where each triangle $\mathcal{T} \overset{\text{def}}{=} \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ in the set is defined by its three corner points $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3 \in \mathbb{R}^3$. For each set of four adjacent image points (x,y), $(x+1,y)$, $(x,y+1)$ and $(x+1,y+1)$ two triangles $\mathcal{T}_g$ and $\mathcal{T}_h$ are formed from the 3D points that correspond to these image points. Note that these triangles need to cover the complete area of the rectangle defined by the image coordinates and there are only two possibilities to select the points for the triangles in that way.

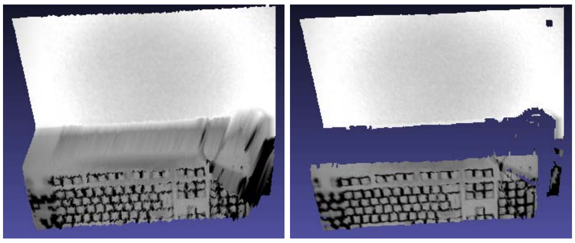

Figure 3: *If the image grid is used to build a triangle mesh from the depth image, separated objects in the scene are connected in the triangle mesh (left). A simple filtering of triangles with long edges separates the objects again (right).*

Nearby objects in the resulting triangulation will be connected with each other, even if there is no physical connection between them in the scene. This is due to the fact that the grid based triangulation does not distinguish any object borders. Shape and position of these connections depends highly on the view direction of the camera and hence complicates the problem of extracting point correspondences. In order to separate the objects from each other, we apply a simple heuristic. We determine the mean μ_l and the standard deviation σ_l of the edge length l of all triangles $\mathcal{T} \in \mathcal{S}$. Triangles where one of the edges has a length $l \geq \mu_l + \sigma_l$ are deleted. Figure 3 shows an example of the triangle mesh before and after the filtering. An alternative approach to avoid the connections between objects in the triangle mesh is to ignore pixel positions with high gradients in the depth image during the triangulation process.

3. Relative Pose Estimation

If two or more ToF cameras simultaneously record images of the same scene from different viewpoints, each camera records 3D data in its own coordinate system. The extrinsic calibration describes the transformation between the different camera coordinate systems. This transformation is basically a similarity transformation that maps a 3D point $\mathbf{X}_i$ in the coordinate system of camera i to its corresponding 3D point

$$\mathbf{X}_j \overset{\text{def}}{=} s_{i,j}\mathbf{R}_{i,j}\mathbf{X}_i + \mathbf{t}_{i,j} \tag{5}$$

in the coordinate system of camera j, where the similarity transformation consists of a rotation $\mathbf{R}_{i,j} \in \mathrm{SO}(3)$, a translation $\mathbf{t}_{i,j} \in \mathbb{R}^3$ and a scale $s_{i,j} \in \mathbb{R}$. Since most ToF cameras measure the depth in metric units, however, only the relative pose $\mathbf{R}_{i,j}, \mathbf{t}_{i,j}$ needs to be estimated and the scale can be assumed as constant $s_{i,j} \overset{\text{def}}{=} 1$.

The two points $(\mathbf{X}_i, \mathbf{X}_j)$ form a 3D point correspondence, since both points describe the same 3D point but in different coordinate systems. With a set of at least three of these point correspondences it is possible to estimate the relative pose by minimizing

$$\underset{\mathbf{R}_{i,j},\mathbf{t}_{i,j}}{\operatorname{argmin}} \sum_k \left\| \mathbf{X}_j^k - \left(\mathbf{R}_{i,j}\mathbf{X}_i^k + \mathbf{t}_{i,j} \right) \right\|^2 \quad , \tag{6}$$

where k runs over all point correspondences. One approach that is capable of minimizing this energy function for a set of 3D point correspondences is the method of Walker and Shao [WS91]. While most other approaches estimate the translation $\mathbf{t}_{i,j}$ and the rotation $\mathbf{R}_{i,j}$ separately in different steps, Walker and Shao use dual number quaternions to estimate both simultaneously which improves the accuracy.

We embed this estimation method into a RANSAC scheme [FB81] to increase the robustness against outliers. An important property of the 3D points can be exploited during the point correspondence sampling of the RANSAC relative pose estimation. The Euclidean distance between the selected 3D points in one camera coordinate system needs to be identical (or at least close) to the distances between the corresponding 3D points in the second camera coordinate system [DWJM98]. If the selected point correspondences do not satisfy this condition, the sample is discarded and a new sample set is drawn.

The obtained calibration is finally refined by applying a variant of the iterative closest point (ICP) algorithm [BM92] similar to the one suggested by May et al. [MDH*09]. Each point in the 3D point set of the first camera is transformed into the coordinate system of the second camera using the current estimate of the relative pose. If it lies in the area of view of the second camera, the nearest neighbor in the second point set is searched. If the distance to this second point is higher than some threshold θ, the point pair is rejected. The resulting 3D point pairs are used to estimate the relative pose. This approach is repeated until convergence. During the iteration the threshold θ is slowly decreased.

4. Point Correspondences Extraction

The difficulty of point correspondence extraction from the image data increases with the baseline between the cameras. Classical approaches for the extraction of point correspondences provide only poor results and collapse even at quite small baselines. This is on the one hand caused by the low image resolution and the high proportion of image noise in the ToF camera images. But also the increasing perspective influences complicate the extraction of correct point correspondences. In this section we will present two descriptors: one that operates on the depth images of the ToF cameras and an intensity difference based descriptor. We describe how both of these descriptors can be used simultaneously to extract 3D point correspondences.

4.1. 3D Descriptor

Trummer et al. [TSD09] present an approach for the registration of 3D surface triangulations based on moment invariants. Since a surface triangulation $\mathcal{S}$ can easily be obtained from the ToF images (as described in Section 2.2), this descriptor is well suited to be applied to our problem. We will now shortly present the basic idea of the descriptor. For further details the reader is referred to [TSD09].

The $(k+l+m)^{\text{th}}$-order 3D surface moment of the surface triangulation $\mathcal{S}$

$$M_{klm}(\mathcal{S}) \overset{\text{def}}{=} \sum_{i=1}^{n} m_{klm}^{i} \tag{7}$$

consists of the accumulated surface moments m_{klm}^{i} of each triangle $\mathcal{T}_i \in \mathcal{S}$. In order to efficiently calculate these surface moments, Trummer et al. [TSD09] suggest to use a minimal parameterization for the points on a triangle $\mathcal{T}$

$$\mathbf{p}_{\mathcal{T}}(u,v) = \left(x_{\mathcal{T}}(u,v), y_{\mathcal{T}}(u,v), z_{\mathcal{T}}(u,v)\right)^{\mathrm{T}} \tag{8}$$

$$\overset{\text{def}}{=} u(\mathbf{c}_1-\mathbf{c}_3)+v(\mathbf{c}_2-\mathbf{c}_3)+\mathbf{c}_3 \quad , \tag{9}$$

where $u,v \geq 0$ are the parameterization scalars with $u+v \leq 1$. Using this parameterization, the surface moments can be written as

$$m_{klm} \overset{\text{def}}{=} C \iint_D x_{\mathcal{T}}^k(u,v)\, y_{\mathcal{T}}^l(u,v)\, z_{\mathcal{T}}^m(u,v)\, \mathrm{d}u\,\mathrm{d}v \quad , \tag{10}$$

where

$$D \overset{\text{def}}{=} \{(u,v) : u,v \geq 0, u+v \leq 1\} \tag{11}$$

is the domain of the triangle parameterization and

$$C \overset{\text{def}}{=} \sqrt{\left(x_u^2+y_u^2+z_u^2\right)\left(x_v^2+y_v^2+z_v^2\right) - \left(x_u x_v + y_u y_v + z_u z_v\right)^2} \tag{12}$$

contains the coefficients of the first fundamental form. The notation

$$x_u \overset{\text{def}}{=} \frac{\partial x_{\mathcal{T}}(u,v)}{\partial u} \tag{13}$$

denotes a partial derivative. Trummer et al. [TSD09] show that the integrals in (10) can be easily resolved and the computation of the surface moments m_{klm} is reduced to a simple equation

$$m_{klm} = C\Big((x_1-x_3)^k (y_1-y_3)^l (z_1-z_3)^m m_{(k+l+m)0} + \ldots + x_3^k y_3^l z_3^m m_{00} \Big) \tag{14}$$

that only contains the coordinates of the 3D triangle corner points and the area moments of the triangle parameterization

$$m_{pq} \overset{\text{def}}{=} \iint_D u^p v^q \mathrm{d}u\,\mathrm{d}v \quad . \tag{15}$$

These area moments m_{pq} have two advantages. First, they are easy to compute. But much more important: they are independent from any specific triangle. Hence, they can be efficiently precomputed.

The 3D surface moments $M_{klm}(\mathcal{S})$ are finally used to compute the 3D descriptor which consists of the eleven 3D moment invariants $I_{22}^2, I_{222}^2, \ldots, I_{1113}^3$ proposed by [LD89].

These invariants include moments up to third order. For further details on these moments and how to compute them, the reader is referred to [LD89].

Since we are interested in estimating a descriptor that distinguishes a single 3D point $\mathbf{X}$, not the complete surface triangulation is used. Instead only the triangles

$$\mathcal{S}(\mathbf{X},r) \overset{\text{def}}{=} \left\{\mathcal{T} : \min\left\{\|\mathbf{c}_i - \mathbf{X}\|_2 : i = 1,2,3\right\} \le r\right\} \quad (16)$$

that lie within a sphere of radius r around $\mathbf{X}$ are used. Triangles that jut out of the sphere are approximated in the way suggested by Trummer et al. [TSD09].

Obviously the radius needs to be selected identical for all cameras and depends on the camera-to-scene distance. However, our experiments (Section 5) show that the choice of the radius is not that critical. A good heuristic used in our experiments is to select several different radii $r \overset{\text{def}}{=} \lambda_r d_{xy}$ relative to the dilation d_{xy} of the 3D point set in x and y direction, where we use three different relative radii $\lambda_r \in \{0.03, 0.06, 0.09\}$. Note that during the matching only descriptors with identical radius are compared.

4.2. Intensity Descriptor

The 3D descriptor described in the previous section is built entirely from the depth estimates of a ToF camera. Thus it is only possible to match the 3D points of the different cameras if there is enough 3D structure in the scene to distinguish the corresponding descriptors. If the scene does not offer enough structural variation or if the structure is very redundant, the simultaneously recorded intensity images can resolve the resulting ambiguities. Hence, we describe in this section a second descriptor based on the intensity images.

Classical image descriptors like [Low04] use histograms of image gradients and scale spaces to create distinguishable and invariant descriptors. Since the 3D descriptor already describes the local structure and we can use the 3D information around the keypoint, we suggest a different approach. For each keypoint $\mathbf{X}$ we consider only 3D points

$$\mathcal{P}(\mathbf{X}) \overset{\text{def}}{=} \left\{\mathbf{X}_i : \|\mathbf{X} - \mathbf{X}_i\|_2 < r\right\} \quad (17)$$

that lie in a sphere with radius r around the keypoint. Note that this is exactly the point set used to calculate the 3D descriptor. Using only points lying in this sphere ensures scale invariance of the descriptors since descriptors with identical radius describe equally sized 3D areas.

Each 3D point $\mathbf{X}$ corresponds to some pixel coordinate, as described in Section 2.2. Hence we can assign an intensity value $I(\mathbf{X})$ to each 3D point in the set. Calculating intensity gradients between these 3D points is not trivial, since the density of the local 3D point cloud might vary and no clear neighborhood relation exists. Instead we build a histogram of intensity differences between the keypoint $\mathbf{X}$ and nearby points $\mathcal{P}(\mathbf{X})$. The histogram ensures the rotation invariance of the descriptor and the intensity differences ensure invariance to additive illumination changes.

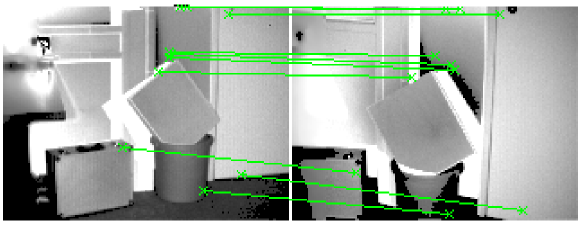

Figure 4: *Ten out of* 109 *inlier point correspondences found using the combined descriptor.*

For each $\mathbf{X}_i \in \mathcal{P}(\mathbf{X})$ the two histogram bins closest to the intensity difference

$$I(\mathbf{X}_i) - I(\mathbf{X}) \quad (18)$$

are increased by a weight ω using a bilinear interpolation to apportion the weight to the two bins. The weight is chosen from a normal distribution

$$\omega(\mathbf{X}_i) \sim \mathcal{N}\left(\mathbf{X}_i \mid \mathbf{X}, \sigma^2\right) \quad (19)$$

that is centered on the keypoint and has a variance of $\sigma^2 \overset{\text{def}}{=} r^2$. The purpose of this Gaussian weighting is to avoid that small changes in the position of the sphere result in severe descriptor changes. The final descriptor is normalized to unit length.

4.3. Matching

The method described in Section 4.1 computes a 3D descriptor for each single 3D point obtained from the depth image. Matching the entire descriptor sets of two images would take much too long. Instead we search for a subset of the point set which includes the points with the most distinctive descriptors. We evaluate the distinctiveness of a point by comparing its descriptor with the descriptors of the neighboring points as proposed by Trummer et al. [TSD09].

The two descriptors described in Section 4.1 and Section 4.2 cannot be easily combined to a single descriptor since they differ in dimension and magnitude. Hence we use two normalized distance measures to determine the descriptor distance

$$d\left(\mathbf{X}_i, \mathbf{X}_j\right) \overset{\text{def}}{=} \frac{1}{\sigma_{3D}} d_{3D}\left(\mathbf{X}_i, \mathbf{X}_j\right) + \frac{1}{\sigma_I} d_I\left(\mathbf{X}_i, \mathbf{X}_j\right) \quad (20)$$

between two 3D points $\mathbf{X}_i$ and $\mathbf{X}_j$ of different ToF camera images, where $d_{3D}(\mathbf{X}_i, \mathbf{X}_j)$ and $d_I(\mathbf{X}_i, \mathbf{X}_j)$ are the Euclidean distances between the 3D and intensity descriptors of the 3D points. The normalization factors σ_{3D} and σ_I are the standard deviations of the nearest neighbor distances of the respective descriptor type. Note that we calculate these distances only between descriptors of different images and not within the descriptors of one image.

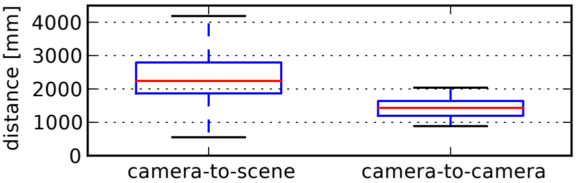

Figure 5: *The camera-to-scene distance (extracted from the measured depth) and the distance between the cameras in our experiments.*

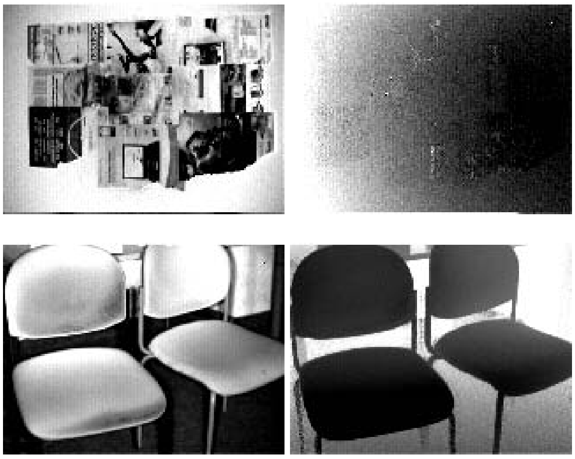

Figure 6: *The intensity (left column) and depth image (right column) of two very different scenes used in the experiments. The first scene (top row) offers almost no 3D structure but a lot of texture. Contrary to this, the second scene (bottom row) consists of various 3D structure but only few texture.*

For each interest point in the first camera, we search for its nearest neighbor in the point set of the second camera. The same procedure is repeated vice versa. The final point correspondence set is the intersection of these two sets. Figure 4 shows some example point correspondences extracted using both descriptors. Note that these point correspondences are not totally accurate due to the heavy noise in the depth and intensity images. However, the accuracy of the extracted correspondences suffices to estimate a good initial relative pose which is then refined using ICP (Section 3).

5. Experimental Evaluation

5.1. Setup

For our experiments we use two PMDTechnologies PMD[vision] 19k cameras. Each of these has a resolution of 160 × 120 pixels. The selected modulation frequencies are 20 MHz and 21 MHz, respectively. The automatically adjusted integration times lie between 30 ms and 60 ms. For the intrinsic calibration of each camera and the extrinsic ground truth calibration between the two ToF cameras we use the calibration pattern based method of Zhang [Zha99]. The point correspondences extracted from the calibration pattern are also used to evaluate the reprojection error of the estimated calibration.

Each camera selects its integration time using (2). The depth image is smoothed by a 3×3 Gaussian. Only the 15% most distinctive points (see Section 4.3 for details) are used which results in 2880 interest points for each image. We use 16 bins for the intensity difference histogram. This value has been determined in additional experiments. We calibrate 10 different setups of the camera pair. In each setup the relative camera orientation and position as well as the scene are changed. Most of our results are presented using boxplots (the line in the middle is the median, the box depicts the 0.25 and 0.75 quantiles, crosses are outliers [MTL78]).

Figure 5 shows the camera-to-scene distance (extracted from the measured depth) and the distance between the cameras. The scenes vary in the amount of available texture and 3D structure. Figure 2 and Figure 6 give an example of the used scenes, reaching from a textured wall to a scene consisting of low textured objects. Since we are using RANSAC, each calibration is repeated 100 times in order to take effects into account that are caused by the random sampling.

5.2. Results

5.2.1. Evaluation of the Relative Descriptor Radius

In Section 4.3 we explained that the selection of the descriptor radius depends on the camera to scene distances of the involved cameras. Since the depth measurements of the ToF cameras are in metric units, a coarse estimate for a good descriptor radius can be derived from the image data. We suggested to select it relative to the dilation d_{xy} of the 3D point set in x and y direction. Figure 7 shows the median reprojection error of our proposed method for a varying relative descriptor radius λ_r. Note that the interval of relative radii that lead to a good calibration is quite big. Hence, the choice of the descriptor radius r is not that critical. For the following experiments we use three relative descriptor radii simultaneously $\lambda_r \in \{0.03, 0.06, 0.09\}$. The resulting absolute radii in our experiments vary between 3.4 cm and 20.8 cm. The median numbers of 3D points that lie inside a sphere are 65, 269 and 588 for the three respective relative radii.

5.2.2. Calibration Accuracy

Figure 8 shows the calibration accuracies achieved by different methods. Three different versions of our approach are evaluated: the Intensity and the 3D descriptor each on its own and the Combined version of these two descriptors, as proposed in the paper. Note that all three descriptors use the same interest points. These are extracted using the method described in Section 4.3. Hence the Intensity descriptor also uses a certain amount of 3D information.

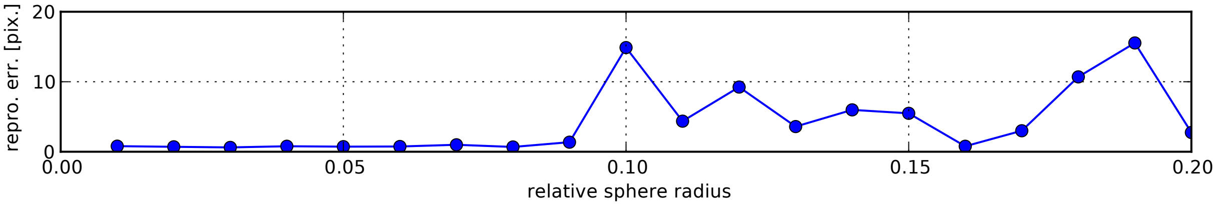

Figure 7: *The achieved median reprojection error for varying relative descriptor radii λ_r. The results show that the selection of this parameter is not that critical. For the later experiment we use three relative descriptor radii simultaneously* $\lambda_r \in \{0.03, 0.06, 0.09\}$.

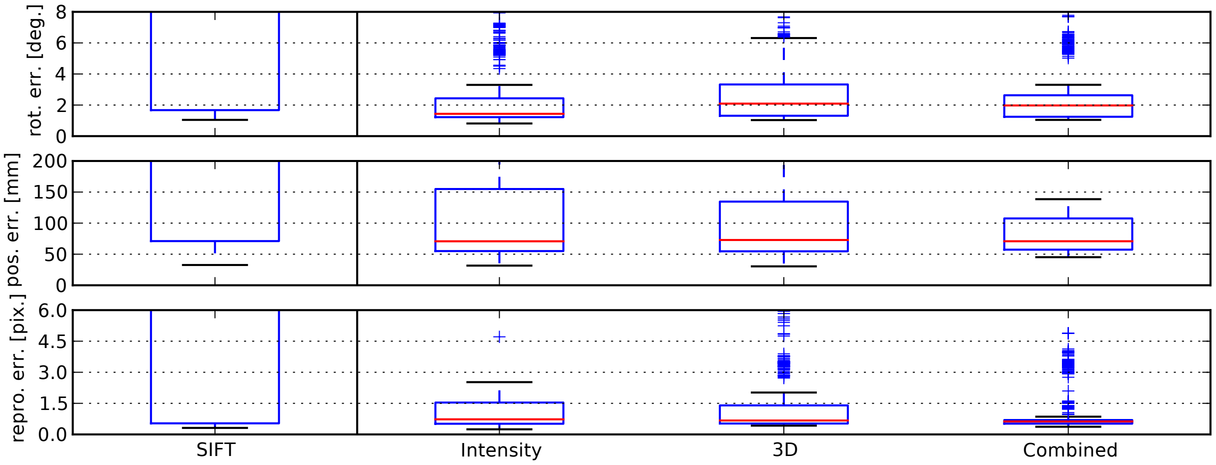

Figure 8: *The calibration errors consisting of the rotation error (top), the position error (center) and the reprojection error (bottom). We present the results using the Intensity and 3D descriptor each on its own and of the Combined descriptor. For comparison we also present the results using SIFT [Low04].*

We also compare our method with SIFT [Low04] applied on the intensity images of the ToF cameras. As motivated in Section 1, when using 2D point correspondences the relative pose can only be estimated up to scale. Hence, we use the 3D points corresponding to the 2D point correspondences extracted by SIFT and use the relative pose estimation described in Section 3. This approach is very similar to the method proposed by May et al. [MDH*09] for 3D mapping.

For all calibrations of each of the different approaches we present the rotation error in degree (top row) and the position error in millimeters (center row) of the estimated relative pose. Furthermore the reprojection error in pixel (bottom row) of both cameras is presented.

Most of the calibrations using SIFT are inaccurate which is caused by the severe amount of outliers in the point correspondences. The low image resolution and the high amount of image noise complicate the extraction of SIFT point correspondences in the wide-baseline setups used for the experiments. The Intensity and the 3D descriptor both achieve a good calibration for most setups. However, it is not surprising that both fail at very different setups. The Combined descriptor, finally, reaches the best results since it is able to resolve the ambiguities of the 3D descriptor in low structured scenes using the additional intensity information. We achieve a median error of 1.97 degree for the rotation and 70.8 millimeters for the position of the cameras. Due to the low median reprojection error of 0.63 pixels, the calibration can be used for many different types of applications.

Note that we do not correct the systematic depth measurement error of the ToF cameras. Estimating a model of this error with one of the calibration pattern based methods [LK06, SBK08] might result in lower rotation and position errors. However, in our future work we want to expand our self-calibration approach in a way that it is also able to estimate a model for this systematic error.

5.2.3. Runtime

Our current implementation takes about 97 seconds on an off-the-shelf quad-core processor for the calibration of two images. About 63 seconds are needed for computing the 3D descriptor since this is done for each single 3D point. Hence, a GPU implementation of this descriptor computation would result in a much better runtime.

6. Conclusions

In this paper we discussed the problem of extrinsic self-calibration of Time-of-Flight (ToF) cameras. Classical approaches for relative pose estimation between cameras are inappropriate since they fail to extract point correspondences in wide-baseline camera setups. Furthermore they do not use the depth measurements of the ToF cameras which are important to estimate the correct scale of the translation. Only if this scale is known, depth estimates of different ToF cameras can be transformed into a common coordinate system.

We suggested to estimate the relative pose between the cameras using 3D point correspondences. For that purpose we described the entire calibration starting with the extraction of 3D data from the depth measurements and how the relative pose is estimated using 3D point correspondences. We presented a descriptor combination consisting of a 3D descriptor that is built from the 3D data obtained from the ToF camera and a descriptor based on intensity differences. The advantage of this descriptor combination is that it is able to extract point correspondences in structured but low textured scenes as well as in low structured but textured scenes. In several experiments on real data we demonstrated the robustness and high accuracy of our approach and that the descriptor combination improves the results compared to each of the single descriptors. We achieve a median error of 1.97 degree for the rotation, 70.8 millimeters for the position and 0.63 pixel for the reprojection.

In our future work we want to improve the localization of interest points in the camera images. Similar to the combined descriptor this interest point localization should be performed in the depth and the intensity image. A more sophisticated selection of the interest points would also improve the runtime. Furthermore we aim to use our extrinsic calibration to estimate a model of the systematic depth measurement error.

References

[BM92] BESL P., MCKAY N.: A method for registration of 3-D shapes. *PAMI 14*, 2 (1992), 239–256.

[BSK08] BEDER C., SCHILLER I., KOCH R.: Real-time estimation of the camera path from a sequence of intrinsically calibrated pmd depth images. In *Int. Arch. of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (2008), pp. 45–50.

[CDS00] CHEN X., DAVIS J., SLUSALLEK P.: Wide area camera calibration using virtual calibration objects. In *CVPR* (2000).

[DWJM98] DORAI C., WANG G., JAIN A., MERCER C.: Registration and integration of multiple object views for 3D model construction. *PAMI 20*, 1 (1998), 83–89.

[FB81] FISCHLER M. A., BOLLES R. C.: Random Sample Consensus: a Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM (CACM) 24*, 6 (1981), 381–395.

[GP08] GUAN L., POLLEFEYS M.: A unified approach to calibrate a network of camcorders and tof cameras. In *Proceedings of the IEEE Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2)* (2008).

[HJS08] HUHLE B., JENKE P., STRAÃ§ER W.: On-the-fly scene acquisition with a handy multi-sensor system. *Int. J. Intelligent Systems Technologies and Applications 5*, 3/4 (2008), 255–263.

[HZ03] HARTLEY R., ZISSERMAN A.: *Multiple View Geometry in computer vision*. Cambridge University Press, 2003.

[KBKL10] KOLB A., BARTH E., KOCH R., LARSEN R.: Time-of-flight cameras in computer graphics. *COMPUTER GRAPHICS forum 29*, 1 (2010), 141–159.

[Lan00] LANGE R.: *3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology*. PhD thesis, University of Siegen, 2000.

[LD89] LO C., DON H.: 3-D moment forms: Their construction and application to object identification and positioning. *PAMI 11*, 10 (1989), 1053–1064.

[LK06] LINDNER M., KOLB A.: Lateral and depth calibration of pmd-distance sensors. In *Proceedings of the International Symposium on Visual Computing (ISVC)* (2006), vol. 2, pp. 524–533.

[Low04] LOWE D. G.: Distinctive image features from scale-invariant keypoints. *IJCV 60*, 2 (2004), 91–110.

[MDH*09] MAY S., DROESCHEL D., HOLZ D., FUCHS S., MALIS E., NÜCHTER A., HERTZBERG J.: Three-dimensional mapping with time-of-flight cameras. *Journal of Field Robotics (JFR) 26*, 11-12 (2009), 934–965.

[MTL78] MCGILL R., TUKEY J., LARSEN W. A.: Variations of Boxplots. *The American Statistician 32* (1978), 12–16.

[Nis04] NISTÉR D.: An efficient solution to the five-point relative pose problem. *PAMI 26* (2004), 756–770.

[PMS*08] PRUSAK A., MELNYCHUK O., SCHILLER I., ROTH H., KOCH R.: Pose estimation and map building with a pmd-camera for robot navigation. *International Journal of Intelligent Systems Technologies and Applications 5*, 3-4 (2008), 355–364.

[SBK08] SCHILLER I., BEDER C., KOCH R.: Calibration of a pmd-camera using a planar calibration pattern together with a multi-camera setup. In *Proc. of the Int. Society for Photogrammetry and Remote Sensing Congress* (2008), pp. 297–302.

[SHVG02] SVOBODA T., HUG H., VAN GOOL L.: ViRoom—low cost synchronized multicamera system and its self-calibration. In *DAGM* (2002), Springer, pp. 515–522.

[SLP*07] SWADZBA A., LIU B., PENNE J., JESORSKY O., KOMPE R.: A comprehensive system for 3D modeling from range images acquired from a 3D ToF sensor. In *Proceedings of the ICVS* (2007), pp. 1–10.

[SPH08] SCHALLER C., PENNE J., HORNEGGER J.: Time-of-flight sensor for respiratory motion gating. *Medical Physics 35*, 7 (2008), 3090–3093.

[TSD09] TRUMMER M., SUESSE H., DENZLER J.: Coarse registration of 3d surface triangulations based on moment invariants with applications to object alignment and identification. In *ICCV* (2009), pp. 1273–1279.

[WS91] WALKER M. W., SHAO L.: Estimating 3-d location parameters using dual number quaternions. *CVGIP: Image Understanding 54*, 3 (1991), 358–367.

[XSH*98] XU Z., SCHWARTE R., HEINOL H., BUXBAUM B., RINGBECK T.: Smart pixel-photonic mixer device (pmd) new system concept of a 3d-imaging camera-on-a-chip. In *Proceedings of the IEEE International Conference on Mechatronics and Machine Vision in Practice* (1998), pp. 259–264.

[Zha99] ZHANG Z.: Flexible Camera Calibration by Viewing a Plane from Unknown Orientations. In *ICCV* (1999), pp. 666–673.