Cross-Calibration of Time-of-flight and Colour Cameras

Miles Hansard, Georgios Evangelidis, Quentin Pelorson, Radu Horaud

▶ **To cite this version:**

HAL Id: hal-01059891

https://hal.inria.fr/hal-01059891

Submitted on 2 Sep 2014

Cross-calibration of Time-of-flight and Colour Cameras

Miles Hansard[a,b], Georgios Evangelidis[a], Quentin Pelorson[a], Radu Horaud[a,*]

[a]*INRIA Grenoble Rhône-Alpes, 38330 Montbonnot Saint-Martin, France*
[b]*School of Electronic Engineering and Computer Science, Queen Mary, University of London, Mile End Road, United Kingdom*

Abstract

Time-of-flight cameras provide depth information, which is complementary to the photometric appearance of the scene in ordinary images. It is desirable to merge the depth and colour information, in order to obtain a coherent scene representation. However, the individual cameras will have different viewpoints, resolutions and fields of view, which means that they must be mutually calibrated. This paper presents a geometric framework for the resulting multi-view and multi-modal calibration problem. It is shown that three-dimensional projective transformations can be used to align depth and parallax-based representations of the scene, with or without Euclidean reconstruction. A new evaluation procedure is also developed; this allows the reprojection error to be decomposed into calibration and sensor-dependent components. The complete approach is demonstrated on a network of three time-of-flight and six colour cameras. The applications of such a system, to a range of automatic scene-interpretation problems, are discussed.

Keywords: Camera networks, time-of-flight cameras, depth cameras, camera calibration, 3D reconstruction, RGB-D data

1. Introduction

The segmentation of multi-view video data, with respect to physically distinct objects of interest, is an essential task in automatic scene-interpretation. Visual segmentation can be based on colour, texture, parallax and motion information (e.g. [1, 2]). The task remains very difficult, however, owing to the combined effects of non-rigid surfaces, variable lighting, and occlusion. It has become clear that depth cameras can make an important contribution to scene understanding, by enabling direct *depth segmentation*, based on the measured scene-structure (see CVIU special issue [3]). This approach is also highly effective for dynamic tasks, such as body tracking and action recognition [4]. Furthermore, if depth and colour information can be merged into a single representation, then a complete 3-D representation is possible, in principle. This is clearly desirable, because colour and texture data are essential to many other aspects of scene-understanding, such as identification and tracking [5].

There are two major obstacles to the construction of a complete scene representation, from a multi-modal camera network. Firstly, typical depth sensors are unable to capture RGB data [6]. This means that the depth and colour cameras will have different viewpoints, and so the raw data are *inconsistent*. Secondly, typical ToF and RGB cameras have limited fields of view, and so the depth and colour data are *incomplete*. This paper addresses both of these problems, by showing how to estimate the geometric relationships in a multi-view, multi-modal camera network. This task will be called *cross-calibration*.

In order to constrain the problem, two practical constraints are imposed from the outset. Firstly, the system will be based on *time-of-flight* (ToF) cameras, in conjunction with ordinary RGB cameras. The ToF cameras are compact, can be properly synchronized, and are industrially specified, e.g., [6]. Secondly, a *modular* network of ToF+RGB units is required. This is so that individual units can be added or removed, in order to optimize the scene-coverage.

1.1. Overview

Time-of-flight cameras can, in principle, be geometrically calibrated by standard methods [7]. This means that each pixel records an estimate of the scene-distance (range) along the corresponding ray. The 3-D structure of a scene can also be reconstructed from two or more ordinary images, via

2

the *parallax* (e.g. binocular disparity) between corresponding image points. There are many advantages to be gained by combining the range and parallax data. Most obviously, each point in a parallax-based reconstruction can be mapped back into the original images, from which colour and texture can be obtained. Parallax-based reconstructions are, however, difficult to obtain, owing to the difficulty of putting the image points into correspondence. Indeed, it may be impossible to find any correspondences in untextured regions. Furthermore, if a Euclidean reconstruction is required, then the cameras must be calibrated. The accuracy of the resulting reconstruction will also tend to decrease with the distance of the scene from the cameras [8].

The range data, on the other hand, are often corrupted by noise and surface-scattering. The spatial resolution of current ToF sensors is relatively low, the depth-range is limited, and the luminance signal may be unusable for rendering and for classical image processing. It should also be recalled that ToF cameras, of the type used here, cannot be used in outdoor lighting conditions. These considerations lead to the idea of a *mixed* colour and time-of-flight system, as described in [9]. Such a system could, in principle, be used to make high-resolution Euclidean reconstructions, including photometric information [10, 11].

In order to make full use of a mixed range/parallax system, it is necessary to find the exact geometric relationship between the different devices. In particular, the reprojection of the ToF data, into the colour images, must be obtained. This paper is concerned with the estimation of these geometric relationships. Specifically, the aim is to align the range and parallax reconstructions, by a suitable 3-D transformation.

1.2. Previous work

Multi-view depth and colour camera-networks, of the kind used here, produce data-streams that are subject to a variety of geometric relationships [12]. These relationships depend on the calibration state, relative orientation, and fields of view of the different cameras. It follows that a variety of calibration strategies can be adopted. These are discussed below, with reference to the literature, and contrasted with the approach presented here.

Perhaps the simplest way to combine RGB and ToF data is to perform an essentially 2-D registration between the images and depth maps, as reviewed in [13]; see also [14, 15, 16]. This 2-D approach, however, can only provide an instantaneous solution, because changes in the scene-structure produce corresponding changes in the image-to-image mapping. Moreover, owing to

3

Figure 1: A single TOF+2RGB system (left), comprising a time-of-flight camera in the centre, plus a pair of ordinary colour cameras. This paper addresses the problem of simultaneously cross-calibrating many such systems (right), as a foundation for scene understanding

the different viewpoints, a complete registration will usually be impossible. If the depth camera also produces a reliable intensity image, then photo-consistency can be used as a 3-D calibration criterion. For example, Beder et al. [17] (see also [18, 19]) reproject the *intensities* of the depth data into the colour images, and optimize the camera parameters with respect to the photo-consistency.

Zhu et al. [20] (see also [21]) present a sensor-fusion framework for the integration of TOF depth and binocular disparity information. This method assumes that a dense disparity map is being computed on-line, which is not required by the method presented in this paper. Furthermore, the geometric calibration method [20] requires manual identification of corresponding points, and is based on a weak perspective camera model. In contrast, our method is automatic, and is based on the more appropriate perspective camera model. However, [20] is complementary to our method, in the sense that their sensor-fusion framework (along with a dense stereo-matcher) could be combined with the projective calibration method described below. Wang and Jia [22] describe a related sensor-fusion framework for Kinect (rather than TOF) depth data and colour images.

Another approach to the multi-modal calibration problem is to apply standard methods, as far as possible, to the depth cameras. Wu et al. [23] describe an example of this approach. Lindner et al. [9] analyze the applicability of standard methods to TOF cameras, as well as characterizing the accuracy of the depth data. Mure-Dubois and Hügli [24] describe the Euclidean alignment of multiple TOF point clouds, having calibrated the cameras

by standard methods.

Silva et al. [25] describe a cross-calibration methodology that is based on the identification of 3-D lines in the TOF data, which are then projected to corresponding 2-D lines in the RGB images. This method does not require a chequerboard or other calibration pattern; it does, however, require the existence and detection of straight depth-edges throughout the scene. The approach of Silva et al. also involves a non-trivial correspondence problem, which in turn influences the calibration accuracy. Our method uses a standard chequerboard pattern, with a known number of vertices, for which the correspondence problem is relatively straightforward.

Zhang and Zhang [26] present cross-calibration methodology that is based on plane constraints, as given in [27]. This has the advantage of not requiring 2-D features to be detected in the (low resolution) TOF images. However, this method cannot address the crucial issue of lens distortion, which is considerable in typical TOF cameras [28]. A related Kinect-based calibration system is described by Herrera et al. [29], again using the plane-based method of [27]. Herrera et al. give a careful analysis of the Kinect intrinsic parameters, including lens and depth distortion. The latter is analyzed in more detail by Teichman et al. [30]. Our method does require features to be detected in the TOF images, but this also makes it straightforward to estimate the lens parameters, using standard techniques.

Mikhelson et al. [31] describe an automatic method for registering a Euclidean point-cloud (obtained from a Kinect device) to its 2-D image-projections. This method, like ours, is based on a chequerboard target. However, Mikhelson et al. perform Euclidean 2-D/3-D registration, in contrast to the more general projective 3-D/3-D registration that is described below. Finally, there are methods that perform 3-D/3-D registration of dense data, subject to pointwise adjustments [32]. This strategy can achieve very close registrations, but introduces a more complex optimization problem, which is not fully compatible with the standard calibration pipeline.

1.3. Paper organization and contributions

This paper is organized as follows. Section 2.1 briefly reviews some standard material on projective reconstruction, while section 2.2 describes the representation of range data in the present work. The chief contributions of the subsequent sections are as follows: Section 2.3 describes a point-based method that maps a classical multi-view reconstruction (projective or Euclidean) of the scene onto the corresponding TOF representation. The data

are obtained from a ToF+2RGB system, as shown in figure 1. This does not require the colour cameras to be calibrated (although it is necessary to correct for lens distortion). It is established that this model includes a projective-linear approximation of the systematic ToF depth-error.

Section 3 addresses the problem of multi-system alignment, which is necessary for complete scene-coverage. It is shown that this can be achieved in a way that is compatible with the individual ToF+2RGB calibrations. The complete cross-calibration pipeline, given a collection of chequerboard images, is fully automatic.

Section 4 contains a detailed evaluation of these methods, using several large data-sets, captured by three ToF+2RGB systems (i.e. a nine-camera network). In particular, section 4.1 extends the usual concepts of reprojection error [12] to the multi-modal case. Section 4.2 then introduces a new metric for mixed ToF/RGB systems, which measures instantaneous sensor noise, as well as calibration error. The appropriateness of the 3-D homography transformation, as opposed to a similarity transformation, is tested in section 4.3. Section 4.4 discusses possible applications of these systems, including some real 3-D reconstruction examples. Conclusions and future directions are discussed in section 5.

The system presented here is based on the approach introduced by Hansard et al. [33, 34]. The earlier work has been improved, and extended to the case of multiple ToF and colour cameras. In addition, a new evaluation methodology has been developed, as described above. The automatic detection of calibration targets in the low-resolution ToF images, which is a pre-requisite for the methods described here, was developed in a separate paper [7].

2. Cross-calibration

This section describes the theory of projective alignment, using the following notation. Bold type will be used for vectors and matrices. In particular, points $\boldsymbol{P}$, $\boldsymbol{Q}$ and planes $\boldsymbol{U}$, $\boldsymbol{V}$ in the 3-D scene will be represented by column-vectors of homogeneous coordinates, e.g.

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_\triangle \\ P_4 \end{pmatrix} \quad \text{and} \quad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_\triangle \\ U_4 \end{pmatrix} \tag{1}$$

where $\boldsymbol{P}_\triangle = (P_1, P_2, P_3)^\top$ and $\boldsymbol{U}_\triangle = (U_1, U_2, U_3)^\top$. The homogeneous coordinates are defined up to a non-zero scaling; for example, $\boldsymbol{P} \simeq (\boldsymbol{P}_\triangle/P_4, 1)^\top$. In particular, if $P_4 = 1$, then $\boldsymbol{P}_\triangle$ contains the ordinary space coordinates of

the point $\boldsymbol{P}$. Furthermore, if $|\boldsymbol{U}_\triangle| = 1$, then U_4 is the signed perpendicular distance of the plane $\boldsymbol{U}$ from the origin, and $\boldsymbol{U}_\triangle$ is the unit normal. The point $\boldsymbol{P}$ is on the plane $\boldsymbol{U}$ if $\boldsymbol{U}^\top\boldsymbol{P} = 0$. The cross product $\boldsymbol{u} \times \boldsymbol{v}$ is often expressed as $(\boldsymbol{u})_\times\boldsymbol{v}$, where $(\boldsymbol{u})_\times$ is a 3×3 antisymmetric matrix. The column-vector of N zeros is written $\boldsymbol{0}_N$. Projective cameras are represented by 3×4 matrices. For example, the range projection is

$$\boldsymbol{q} \simeq \boldsymbol{CQ} \quad \text{where} \quad \boldsymbol{C} = \big(\boldsymbol{A}_{3\times3} \,|\, \boldsymbol{b}_{3\times1}\big) \tag{2}$$

is a block-decomposition of the 3×4 camera matrix. The left and right colour cameras $\boldsymbol{C}_\ell$ and $\boldsymbol{C}_r$ are similarly defined, e.g. $\boldsymbol{p}_\ell \simeq \boldsymbol{C}_\ell\boldsymbol{P}$. Table 1 summarizes the geometric objects that will be aligned.

	Observed	Reconstructed	
	Points	Points	Planes
Binocular $\boldsymbol{C}_\ell,\boldsymbol{C}_r$	$\boldsymbol{p}_\ell, \boldsymbol{p}_r$	$\boldsymbol{P}$	$\boldsymbol{U}$
Range $\boldsymbol{C}$	$(\boldsymbol{q}, \rho)$	$\boldsymbol{Q}$	$\boldsymbol{V}$

Table 1: Summary of notations in the left, right and range systems.

Points and planes in the two systems are related by the unknown 4×4 space-homography $\boldsymbol{H}$, so that

$$\boldsymbol{Q} \simeq \boldsymbol{HP} \quad \text{and} \quad \boldsymbol{V} \simeq \boldsymbol{H}^{-\top}\boldsymbol{U}. \tag{3}$$

This model encompasses all rigid, similarity and affine transformations in 3-D. It preserves *collinearity* and *flatness*, and is linear in homogeneous coordinates. Note that, in the reprojection process, $\boldsymbol{H}$ can be interpreted as a modification of the camera matrices. For example, $\boldsymbol{p}_\ell \simeq \big(\boldsymbol{C}_\ell\boldsymbol{H}^{-1}\big)\boldsymbol{Q}$, where $\boldsymbol{H}^{-1}\boldsymbol{Q} \simeq \boldsymbol{P}$ is the point that would theoretically be reconstructed by triangulation.

The 4×4 homographies $\boldsymbol{H}$ include, as special cases, the rigid transformations that would align a fully-calibrated Euclidean stereo-reconstruction to the ToF measurements. There are two motivations for the generalization. Firstly, it allows uncalibrated binocular reconstructions to be used, as described above. Secondly, it has been shown elsewhere [10, 20] that ToF data are subject to *systematic* depth biases and nonlinear distortions. These are difficult to correct, owing to the lack of a complete parametric model, and

to their dependence on the camera settings (e.g. integration time). Nonetheless, the homography model (3) effectively includes a projective-linear approximation of the depth distortion, which is fitted along with the other transformation parameters. This model is quite powerful: it includes rational depth-distortions, with varying parameters, across each bundle of rays. For example, the two-parameter inverse disparity calibration, as used with Kinect devices [29], is a special case of the homography model described here. Indeed, even if the RGB cameras are fully calibrated, the 3-D homographies are needed to account for depth-distortions and residual reconstruction errors, as demonstrated in [33]. This issue will be explored in section 4.3, below.

2.1. Parallax-based reconstruction

A projective reconstruction of the scene can be obtained from matched points $\boldsymbol{p}_{\ell k}$ and $\boldsymbol{p}_{rk}$, together with the fundamental matrix $\boldsymbol{F}$, where $\boldsymbol{p}_{rk}^\top \boldsymbol{F} \boldsymbol{p}_{\ell k} = 0$. The fundamental matrix can be estimated automatically, using the well-established RANSAC method. The camera matrices can then be determined, up to a four-parameter projective ambiguity [12]. In particular, from $\boldsymbol{F}$ and the epipole $\boldsymbol{e}_r$, the cameras can be defined as

$$\boldsymbol{C}_\ell \simeq \big(\boldsymbol{I} \,|\, \boldsymbol{0}_3\big) \ \ \text{and} \ \ \boldsymbol{C}_r \simeq \big((\boldsymbol{e}_r)_\times \boldsymbol{F} + \boldsymbol{e}_r \boldsymbol{g}^\top \,\big|\, \gamma \boldsymbol{e}_r\big). \tag{4}$$

where $\gamma \neq 0$ and $\boldsymbol{g} = (g_1, \, g_2, \, g_3)^\top$ can be used to bring the cameras into a plausible form. This makes it easier to visualize the projective reconstruction and, more importantly, can improve the numerical conditioning of subsequent procedures.

2.2. Range-based reconstruction

The TOF camera $\boldsymbol{C}$ provides the range (i.e. radial distance) ρ of each scene-point from the camera-centre, as well as the associated image-coordinates $\boldsymbol{q} = (x, y, 1)$. The back-projection of this point into the scene is

$$\boldsymbol{Q}_\triangle = \boldsymbol{A}^{-1}\big((\rho/\alpha)\,\boldsymbol{q} - \boldsymbol{b}\big) \ \ \text{where} \ \ \alpha = \big|\boldsymbol{A}^{-1}\boldsymbol{q}\big|. \tag{5}$$

Hence the point $(\boldsymbol{Q}_\triangle, 1)^\top$ is at distance ρ from the optical centre $-\boldsymbol{A}^{-1}\boldsymbol{b}$, in the direction $\boldsymbol{A}^{-1}\boldsymbol{q}$. The scalar α serves to normalize the direction-vector. This is the standard pinhole model, as used in [35].

The range data are noisy and incomplete, owing to illumination and scattering effects. This means that, given a sparse set of features in the intensity

image (of the ToF device), it is not advisable to use the back-projected point (5) directly. A better approach is to segment the image of the plane in each ToF camera (using the the range and/or intensity data). It is then possible to back-project *all* of the enclosed points, and to robustly fit a plane $\boldsymbol{V}_j$ to the enclosed points $\boldsymbol{Q}_{ij}$, so that $\boldsymbol{V}_j^\top \boldsymbol{Q}_{ij} \approx 0$ if point i lies on plane j. Now, the back-projection $\boldsymbol{Q}_\pi$ of each sparse feature point $\boldsymbol{q}$ can be obtained by intersecting the corresponding ray with the plane $\boldsymbol{V}$, so that the new range estimate ρ^π is

$$\rho^\pi = \frac{\boldsymbol{V}_\triangle^\top \boldsymbol{A}^{-1} \boldsymbol{b} - V_4}{(1/\alpha)\,\boldsymbol{V}_\triangle^\top \boldsymbol{A}^{-1} \boldsymbol{q}} \tag{6}$$

where $|V_4|$ is the distance of the plane to the camera centre, and $\boldsymbol{V}_\triangle$ is the unit-normal of the range plane. The new point $\boldsymbol{Q}^\pi$ is obtained by substituting ρ^π into (5).

The choice of plane-fitting method is affected by two issues. Firstly, there may be very severe outliers in the data, due to the photometric and geometric errors in the depth-estimation process. Secondly, the noise-model should be based on the pinhole geometry, which means that perturbations occur radially along visual directions, which are not (in general) perpendicular to the observed plane [36, 37]. Several plane-fitting methods, both iterative [38] and non-iterative [39] have been proposed for the pinhole model. However, for ToF data, the chief problem is the large number of outliers. This means that a RANSAC-based method is the most effective in this context [33].

2.3. Projective Alignment

It is straightforward to show that the transformation $\boldsymbol{H}$ in (3) could be estimated from five binocular points $\boldsymbol{P}_k$, together with the corresponding range points $\boldsymbol{Q}_k$. This would provide 5×3 equations, which determine the 4×4 entries of $\boldsymbol{H}$, subject to an overall projective scaling. It is better, however, to use the 'Direct Linear Transformation' method [12], which fits $\boldsymbol{H}$ to *all* of the data. This method is based on the fact that if

$$\boldsymbol{P}' = \boldsymbol{H}\boldsymbol{P} \tag{7}$$

is a perfect match for $\boldsymbol{Q}$, then $\mu\boldsymbol{Q} = \lambda\boldsymbol{P}'$, and the scalars λ and μ can be eliminated between pairs of the four implied equations [40]. This results in $\binom{4}{2} = 6$ interdependent constraints per point. It is convenient to write these

9

homogeneous equations as

$$\begin{pmatrix} Q_4 \boldsymbol{P}'_\triangle - P'_4 \boldsymbol{Q}_\triangle \\ \boldsymbol{Q}_\triangle \times \boldsymbol{P}'_\triangle \end{pmatrix} = \boldsymbol{0}_6. \tag{8}$$

Note that if $\boldsymbol{P}'$ and $\boldsymbol{Q}$ are normalized so that $P'_4 = 1$ and $Q_4 = 1$, then the magnitude of the top half of (8) is simply the distance between the points. Following Förstner [41], the left-hand side of (8) can be expressed as $(\boldsymbol{Q})_\wedge \boldsymbol{P}'$ where

$$(\boldsymbol{Q})_\wedge = \begin{pmatrix} Q_4 \boldsymbol{I}_3 & -\boldsymbol{Q}_\triangle \\ (\boldsymbol{Q}_\triangle)_\times & \boldsymbol{0}_3 \end{pmatrix} \tag{9}$$

is a 6×4 matrix, and $(\boldsymbol{Q}_\triangle)_\times \boldsymbol{P}_\triangle = \boldsymbol{Q}_\triangle \times \boldsymbol{P}_\triangle$, as usual. The equations (8) can now be written in terms of (7) and (9) as

$$(\boldsymbol{Q})_\wedge \boldsymbol{H} \boldsymbol{P} = \boldsymbol{0}_6. \tag{10}$$

This system of equations is linear in the unknown entries of $\boldsymbol{H}$, the columns of which can be stacked into the 16×1 vector $\boldsymbol{h}$. The Kronecker product identity $\mathrm{vec}(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Z}) = (\boldsymbol{Z}^\top \otimes \boldsymbol{X})\,\mathrm{vec}(\boldsymbol{Y})$ can now be applied, to give

$$\left(\boldsymbol{P}^\top \otimes (\boldsymbol{Q})_\wedge\right)\boldsymbol{h} = \boldsymbol{0}_6 \quad \text{where} \quad \boldsymbol{h} = \mathrm{vec}(\boldsymbol{H}). \tag{11}$$

If M points are observed on each of N planes, then there are $k = 1, \ldots, MN$ observed pairs of points, $\boldsymbol{P}_k$ from the projective reconstruction and $\boldsymbol{Q}_k$ from the range back-projection. The MN corresponding 6×16 matrices $\left(\boldsymbol{P}_k^\top \otimes (\boldsymbol{Q}_k)_\wedge\right)$ are stacked together, to give the complete system

$$\begin{pmatrix} \boldsymbol{P}_1^\top & \otimes & (\boldsymbol{Q}_1)_\wedge \\ & \vdots & \\ \boldsymbol{P}_{MN}^\top & \otimes & (\boldsymbol{Q}_{MN})_\wedge \end{pmatrix} \boldsymbol{h} = \boldsymbol{0}_{6MN} \tag{12}$$

subject to the constraint $|\boldsymbol{h}| = 1$, which excludes the trivial solution $\boldsymbol{h} = \boldsymbol{0}_{16}$. It is straightforward to obtain an estimate of $\boldsymbol{h}$ from the SVD of the the $6MN \times 16$ matrix on the left of (12). This solution, which minimizes an *algebraic error* [12], is the singular vector corresponding to the smallest singular value of the matrix. Note that the point coordinates should be transformed, to ensure that (12) is numerically well-conditioned [12]. In this case the transformation ensures that $\sum_k \boldsymbol{P}_{k\triangle} = \boldsymbol{0}_3$ and $\frac{1}{MN}\sum_k |\boldsymbol{P}_{k\triangle}| = \sqrt{3}$,

where $P_{k4} = 1$. The analogous transformation is applied to the range points $\boldsymbol{Q}_k$.

The above procedure effectively computes a projective basis in 3-D, which requires five points, no four of which may be co-planar. This gives 5×3 numbers, which are equivalent to the 15 parameters of the homogeneous 4×4 matrix $\boldsymbol{H}$. In practice, however, we require full detection of 35 vertices per board, and around 10 boards per homography (depending on visibility constraints). Hence the procedure operates far from any degenerate cases, which would involve fewer than 15 degrees of freedom in total.

The DLT method, in practice, gives a good approximation $\boldsymbol{H}_{\mathrm{DLT}}$ of the homography (3). This can be used as a starting-point for the iterative minimization of a more appropriate error measure. In particular, consider the *reprojection error* in the left image,

$$E_\ell(\boldsymbol{C}_\ell) = \sum_{k=1}^{MN} D\big(\boldsymbol{C}_\ell \boldsymbol{Q}_k, \, \boldsymbol{p}_{\ell k}\big)^2 \tag{13}$$

where $D(\boldsymbol{p}, \boldsymbol{q}) = |\boldsymbol{p}_\triangle/p_3 - \boldsymbol{q}_\triangle/q_3|$. A 12-parameter minimization of the re-projection error (13), starting with the linear estimate $\boldsymbol{C}_\ell \leftarrow \boldsymbol{C}_\ell \boldsymbol{H}_{\mathrm{DLT}}^{-1}$, is then performed by the Levenberg-Marquardt algorithm [42]. The result will be the camera matrix $\boldsymbol{C}_\ell^\star$ that best reprojects the range data into the left image ($\boldsymbol{C}_r^\star$ is similarly obtained). The solution, provided that the calibration points adequately covered the scene volume, will remain valid for subsequent depth and range data.

Alternatively, it is possible to minimize the *joint* reprojection error, defined as the sum of left and right contributions,

$$E\big(\boldsymbol{H}^{-1}\big) = E_\ell\big(\boldsymbol{C}_\ell \boldsymbol{H}^{-1}\big) + E_r\big(\boldsymbol{C}_r \boldsymbol{H}^{-1}\big) \tag{14}$$

over the (inverse) homography $\boldsymbol{H}^{-1}$. The 16 parameters are again minimized by the Levenberg-Marquardt algorithm, starting from the DLT solution $\boldsymbol{H}_{\mathrm{DLT}}^{-1}$.

The difference between the separate (13) and joint (14) minimizations is that the latter preserves the original epipolar geometry, whereas the former does not. Recall that $\boldsymbol{C}_\ell$ $\boldsymbol{C}_r$, $\boldsymbol{H}$ and $\boldsymbol{F}$ are all defined up to scale, and that $\boldsymbol{F}$ satisfies an additional rank-two constraint [12]. Hence the underlying parameters can be counted as $(12 - 1) + (12 - 1) = 22$ in the separate minimizations, and as $(16 - 1) = 15$ in the joint minimization. The fixed epipolar geometry accounts for the $(9 - 2)$ missing parameters in the joint

minimization. If $\boldsymbol{F}$ is known to be very accurate (or must be preserved) then the joint minimization (14) should be performed. This will also preserve the original binocular triangulation, provided that a projective-invariant method was used [43]. However, if minimal reprojection error is the objective, then the cameras should be treated separately. This will lead to a new fundamental matrix $\boldsymbol{F}^\star = (\boldsymbol{e}_r^\star)_\times \boldsymbol{C}_r^\star (\boldsymbol{C}_\ell^\star)^+$, where $(\cdot)^+$ is the generalized inverse, and $\boldsymbol{C}_\ell^\star$, $\boldsymbol{C}_r^\star$ are the optimized camera-matrices. The epipole in the right-hand image is obtained from $\boldsymbol{e}_r^\star = \boldsymbol{C}_r^\star \boldsymbol{d}_\ell^\star$, where the vector $\boldsymbol{d}_\ell^\star$ represents the nullspace $\boldsymbol{C}_\ell^\star \boldsymbol{d}_\ell^\star = \boldsymbol{0}_3$.

3. Multi-System Alignment

The methods described in section 2.3 can be used to calibrate a single TOF+2RGB system; the joint calibration of *several* such systems will now be explained. In this section the notation $\boldsymbol{P}_i$ will be used for the binocular coordinates (with respect to the left camera) of a point in the i-th system, and likewise $\boldsymbol{Q}_i$ for the TOF coordinates of a point in the same system. Hence the i-th TOF, left and right RGB cameras (sharing the same physical mounting) have the form

$$\boldsymbol{C}_i \simeq \left(\boldsymbol{A}_i \,|\, \boldsymbol{0}_3\right)$$
$$\boldsymbol{C}_{\ell i} \simeq \left(\boldsymbol{A}_{\ell i} \,|\, \boldsymbol{0}_3\right) \quad \text{and} \quad \boldsymbol{C}_{ri} \simeq \left(\boldsymbol{A}_{ri} \,|\, \boldsymbol{b}_{ri}\right) \tag{15}$$

where $\boldsymbol{A}_i$ and $\boldsymbol{A}_{\ell i}$ contain only *intrinsic* parameters, whereas $\boldsymbol{A}_{ri}$ also encodes the relative orientation of $\boldsymbol{C}_{ri}$ with respect to $\boldsymbol{C}_{\ell i}$. Each system has a transformation $\boldsymbol{H}_i^{-1}$ that maps TOF points $\boldsymbol{Q}_i$ into the corresponding RGB coordinate system of $\boldsymbol{C}_{\ell i}$. Furthermore, let the 4×4 matrix $\boldsymbol{G}_{ij}$ be the transformation from system j, mapping *back* to system i (between different physical mountings). This matrix, in the binocularly-calibrated case, is a rigid 3-D transformation. However, by analogy with the TOF-to-RGB matrices $\boldsymbol{H}$, each $\boldsymbol{G}_{ij}$ could be a *projective* transformation in the uncalibrated case; this would allow Euclidean structure to propagate from the TOF measurements, across the entire camera network.

The left and right cameras, in all cases, that project a scene-point $\boldsymbol{P}_j$ in coordinate system j to image-points $\boldsymbol{p}_{\ell i}$ and $\boldsymbol{p}_{ri}$ in system i are

$$\boldsymbol{C}_{\ell ij} = \boldsymbol{C}_{\ell i} \, \boldsymbol{G}_{ij} \quad \text{and} \quad \boldsymbol{C}_{rij} = \boldsymbol{C}_{ri} \, \boldsymbol{G}_{ij}. \tag{16}$$

Note that if a single global coordinate system is chosen to coincide with the k-th RGB system, then a point $\boldsymbol{P}_k$ projects via $\boldsymbol{C}_{\ell ik}$ and $\boldsymbol{C}_{rik}$. These two

12

cameras are respectively equal to $\boldsymbol{C}_{\ell i}$ and $\boldsymbol{C}_{ri}$ in (15) only when $i = k$, such that $\boldsymbol{G}_{ij} = \boldsymbol{I}$ in (16). A typical three-system configuration is shown in fig. 2.
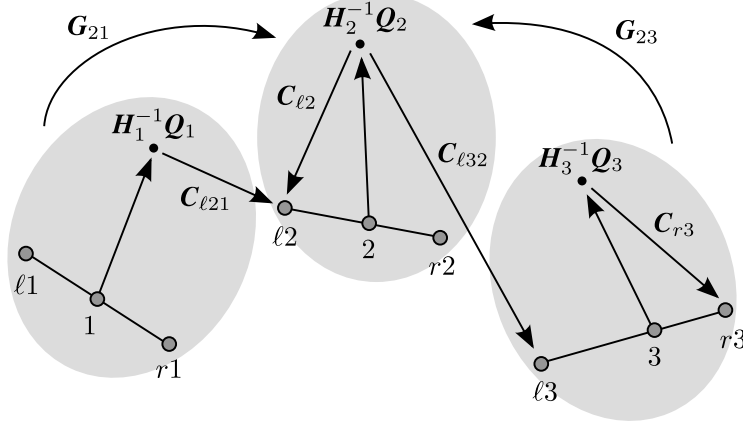

Figure 2: Example of a three ToF+2RGB setup, with ToF cameras labelled 1,2,3; cf. fig. 1. Each ellipse represents a separate system, with system 2 chosen as the reference. The arrows (with camera-labels) show some possible ToF-to-RGB projections. For example, a point $\boldsymbol{P}_2 \simeq \boldsymbol{H}_2^{-1}\boldsymbol{Q}_2$ in the centre projects directly to RGB view $\ell 2$ via $\boldsymbol{C}_{\ell 2}$, whereas the same point projects to $\ell 3$ via $\boldsymbol{C}_{\ell 32} = \boldsymbol{C}_{\ell 3}\boldsymbol{G}_{32}$.

The transformation $\boldsymbol{G}_{ij}$ can only be estimated directly if there is a region of common visibility between systems i and j. If this is not the case (as when the systems face each other, such that the front of the calibration board is not simultaneously visible), then $\boldsymbol{G}_{ij}$ can be computed indirectly. For example, $\boldsymbol{G}_{02} = \boldsymbol{G}_{01}\,\boldsymbol{G}_{12}$ where $\boldsymbol{P}_2 = \boldsymbol{G}_{12}^{-1}\boldsymbol{G}_{01}^{-1}\boldsymbol{P}_0$. Note that, in all cases, the stereo-reconstructed points $\boldsymbol{P}$ are used to estimate these transformations This is because they are always more reliable than the ToF points $\boldsymbol{Q}$, as demonstrated below.

4. Evaluation

The following sections will describe the accuracy of a nine-camera setup, calibrated by the methods described above. Section 4.1 will evaluate *calibration* error, whereas section 4.2 will evaluate *total* error. The former is essentially a fixed function of the estimated camera matrices, for a given scene. The latter also includes the range-noise from the ToF cameras, which varies from moment to moment (due to intrinsic noise, changing illumination, and object motion). The importance of this distinction will be discussed. In

13

section 4.3 we analyze the avantages of using homographies to align the data, as opposed to similarity transformations. Finally, in section 4.4 we demonstrate some applications of the complete system.

The setup consists of three rail-mounted TOF+2RGB systems, $i = 1 \ldots 3$, as in fig. 2. The stereo baselines are 17cm on average, and the TOF cameras are separated by 107cm on average. The RGB images are 1624×1224, whereas the Mesa Imaging SR4000 TOF images are 176×144, with a depth range of 500cm [6]. The three stereo systems are first calibrated by standard methods [12], returning a full Euclidean decomposition of $\boldsymbol{C}_{\ell i}$ and $\boldsymbol{C}_{ri}$, as well as the associated lens parameters. The lenses of the TOF cameras are also calibrated by standard methods [12]. This removes some of the radial depth deformation that has been observed in the TOF data [32]. The matrices $\boldsymbol{G}_{ij}$ are rigid-body transformations, which are estimated by the SVD method of Arun et al. [44]. Projective alignment is preferred for the TOF/RGB alignment, for reasons discussed in sections 2 and 10. Hence the transformations $\boldsymbol{H}_j^{-1}$ are 4×4 homographies, estimated by the method of section 2.3. Specifically, the DLT solutions were refined by Levenberg-Marquardt minimization of the joint geometric error, as in (14).

4.1. Calibration Error

The calibration error is measured by first taking TOF points $\boldsymbol{Q}_j^\pi$ corresponding to chequerboard *vertices* on the reconstructed calibration plane π_j in system j, as described in section 2.2. These can then be projected into a pair of RGB images in system i, so that the geometric image-error $E_{ij}^{\mathrm{cal}} = \frac{1}{2}\big(E_{\ell ij}^{\mathrm{cal}} + E_{rij}^{\mathrm{cal}}\big)$ can be computed, where

$$E_{\ell ij}^{\mathrm{cal}} = \frac{1}{|\pi|} \sum_{\boldsymbol{Q}_j^\pi} D\Big(\boldsymbol{C}_{\ell ij}\boldsymbol{H}_j^{-1}\boldsymbol{Q}_j^\pi, \ \boldsymbol{p}_{\ell i}\Big) \tag{17}$$

and E_{rij}^{cal} is similarly defined. The function $D(\cdot, \cdot)$ computes the image-distance between two inhomogenized points, as in (13), and the denominator corresponds to the number of vertices on the board, with $|\pi| = 35$ in the present experiments. The measure (17) can of course be averaged over all images in which the board is visible. The RGB cameras were calibrated, by standard methods, as described above. Subpixel accuracy was obtained, which confirms the accuracy of the camera matrices $\boldsymbol{C}_{\ell i}$ and $\boldsymbol{C}_{ri}$, as well as the inter-system matrices $\boldsymbol{G}_{ij}$, where $i, j = 1, 2, 3$.

14

For the purpose of evaluation, a *new* set of TOF-vertices $\boldsymbol{Q}_i^\pi$ were reconstructed, fitted and reprojected within each system i. This evaluation effectively tests the quality of the $\boldsymbol{H}_i^{-1}$ matrices, by comparing $\boldsymbol{C}_{\ell i}\boldsymbol{H}_i^{-1}\boldsymbol{Q}_i^\pi$ to $\boldsymbol{p}_{\ell i}$, and analogously to points $\boldsymbol{p}_{ri}$ in the other image. Note that all camera and transformations parameters are now fixed; no optimization was performed with respect to the evaluation data. The whole experiment was performed on three large data-sets, from different capture-sessions, and with different camera configurations, labelled A, B and C. Such a configuration leads to one triplet per system as shown in fig. 3. While an image can be fronto-parallel for one frame, it may appear very slanted to the other frames. The example of fig. 3 shows a calibration image that is almost fronto-parallel to the TOF frame of the central TOF+2RGB unit. Table 2 shows that the average reprojection error remains subpixel in all three data-sets. The corresponding error-distributions are shown as histograms in fig. 4.

Set	Mean	Median	Max	Count
A	0.59	0.52	1.82	1470
B	0.72	0.58	4.86	1470
C	0.45	0.40	1.48	1470

Table 2: Calibration error (17), measured by projecting the fitted TOF vertices $\boldsymbol{Q}_i^\pi$ to the left and right RGB images (1624×1224) of the *respective* systems $i = 1, 2, 3$. The experiment was repeated three times (A–C), with different camera configurations. Each statistic was computed from the left and right RGB-reprojections of 35 vertices in 7 views of the board (total number of 2-D points, per data-set: $3 \times 2 \times 35 \times 7 = 1470$).

It is also interesting to consider how the above TOF-vertices reproject into *different* systems. The error 3-D error distribution, for a given pixel (either TOF or stereo) is highly anisotropic; the *direction* of the corresponding ray is much more reliable than the *distance* along it [32]. In practice, the different TOF+2RGB systems are distributed around the edge of the room, looking inwards. Hence a given system is likely to be seen 'from the side' by at least one other system. This means that any large depth errors in the first system will not cancel-out in the re-projection to the other systems. This effect is seen clearly in figure 5, which shows calibration errors of up to several pixels, from one system to the others.
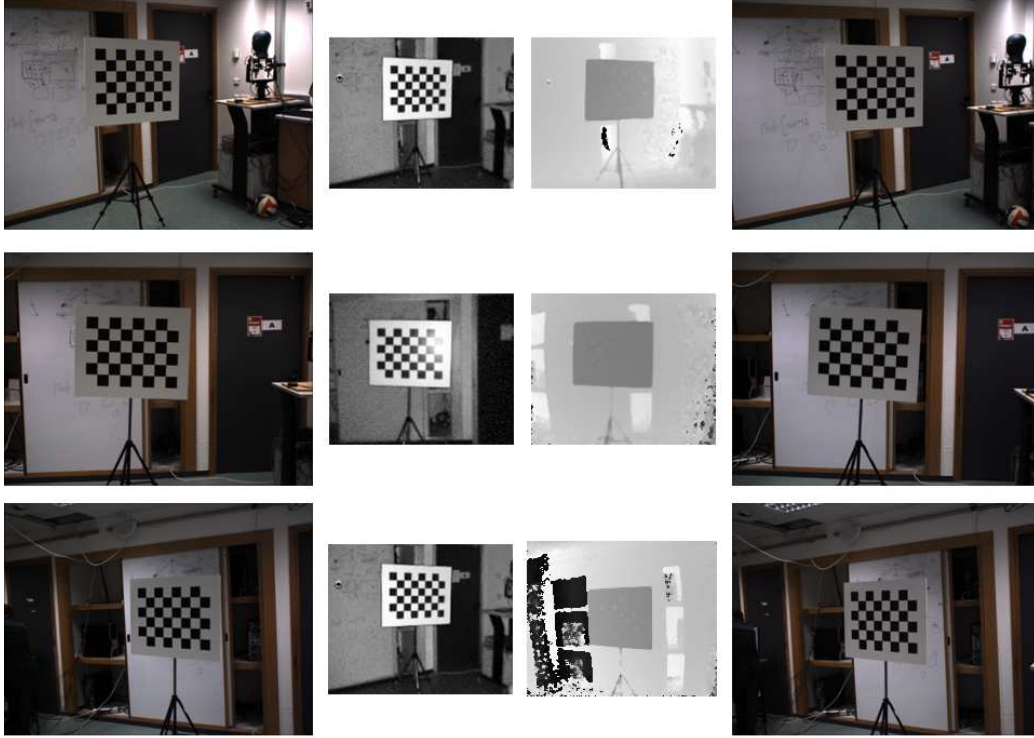
15

Figure 3: Calibration images from synchronized captures. Each row corresponds to the image triplet of a ToF+2RGB system, while the range images are shown as well. Grayscale images of ToF sensors have undergone lens undistortion and contrast enhancement.

4.2. Total Error

The calibration error, as reported in the preceding section, is the natural way to evaluate the estimated cameras and homographies. It is not, however, truly representative of the 'live' performance of the complete setup. This is because the calibration error uses each estimated plane π_j to replace all vertices $\boldsymbol{Q}_j$ with the *fitted* versions $\boldsymbol{Q}_j^\pi$. In general, however, no surface model is available, and so the raw points $\boldsymbol{Q}_j$ must be used as input to segmentation, meshing and rendering processes.

The total error, which combines the calibration and range errors, can be measured as follows. The i-th RGB views of plane π_j are related to the ToF image-points $\boldsymbol{q}_j$ by the 2-D *transfer* homographies $\boldsymbol{T}_{\ell ij}$ and $\boldsymbol{T}_{rij}$, where

$$\boldsymbol{p}_{\ell i} \simeq \boldsymbol{T}_{\ell ij}\,\boldsymbol{q}_j \quad \text{and} \quad \boldsymbol{p}_{ri} \simeq \boldsymbol{T}_{rij}\,\boldsymbol{q}_j. \tag{18}$$
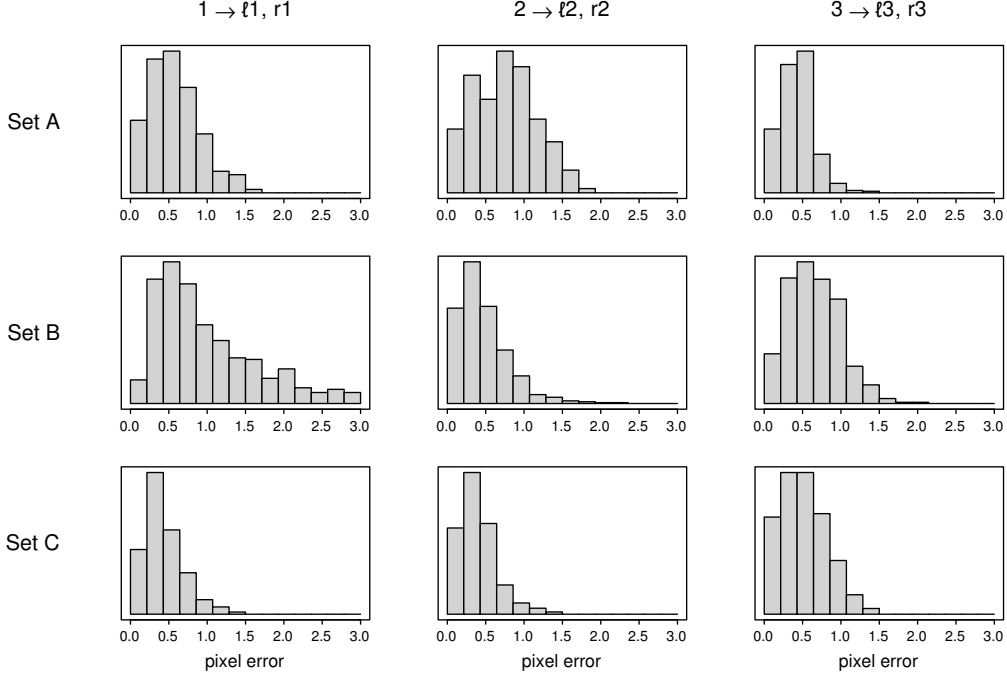
16

Figure 4: Calibration error (17), measured by projecting the fitted TOF points $\boldsymbol{Q}_i^\pi$ to the left and right RGB images (1624×1224) of the *respective* systems $i = 1, 2, 3$. The experiment was repeated three times (A–C), with different camera configurations. Each histogram contains 1470 points, as in table 2.

These 3×3 matrices can be estimated to subpixel accuracy, by using the DLT algorithm [12] to obtain initial homographies, and then applying area-based alignment [45] to produce the final estimates $\boldsymbol{T}_{\ell ij}$ and $\boldsymbol{T}_{rij}$.

Let Π_j be the 2-D hull (i.e. bounding-polygon) of plane π_j as it appears in the TOF image. Any pixel $\boldsymbol{q}_j$ in the hull (including the original calibration vertices) can now be *re-projected* to the i-th RGB views via the 3-D point $\boldsymbol{Q}_j$, or *transferred* directly by $\boldsymbol{T}_{\ell ij}$ and $\boldsymbol{T}_{rij}$ in (18), as shown in figure 6. In other words, we are able to isolate the image-to-image error that is incurred by mapping *via* the 3-D reconstruction, in relation to the direct 2-D to 2-D mapping defined by $\boldsymbol{T}_{\ell ij}$ and $\boldsymbol{T}_{rij}$.

The total error is the average difference between the reprojections and the transfers, $E_{ij}^{\text{tot}} = \frac{1}{2}\big(E_{\ell ij}^{\text{tot}} + E_{rij}^{\text{tot}}\big)$, where

$$E_{\ell ij}^{\text{tot}} = \frac{1}{|\Pi_j|} \sum_{\boldsymbol{q}_j \in \Pi_j} D\Big(\boldsymbol{C}_{\ell ij}\,\boldsymbol{H}_j^{-1}\boldsymbol{Q}_j, \ \ \boldsymbol{T}_{\ell ij}\,\boldsymbol{q}_j\Big) \tag{19}$$
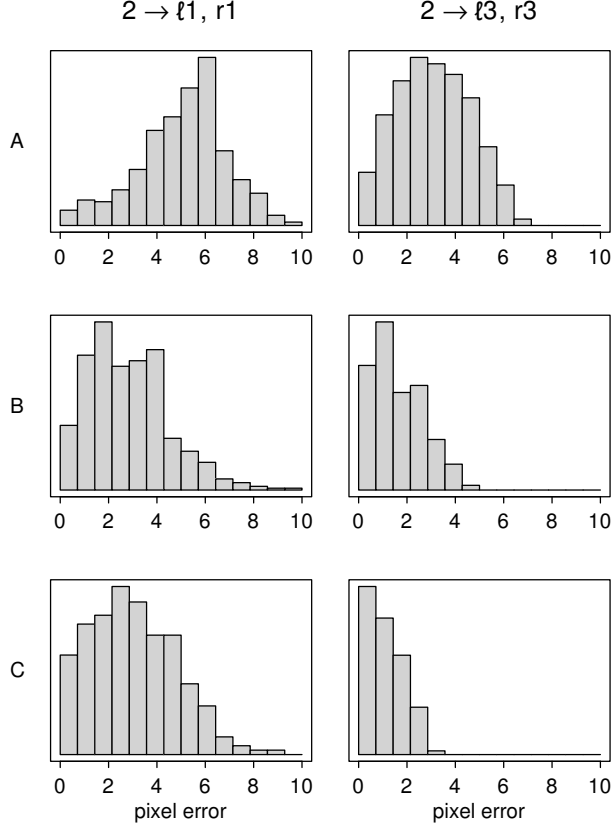
17

Figure 5: Calibration error (17), measured by projecting the fitted ToF points $\boldsymbol{Q}_2^\pi$ to the left and right RGB images (1624×1224) of two *different* systems, $i = 1, 3$. Each histogram combines left-camera and right-camera measurements from 7 views of the calibration board.

and E_{rij}^{tot} is similarly defined. The view-dependent denominator $|\Pi_j| \gg |\pi|$ is the number of *pixels* in the hull Π_j. Hence E_{ij}^{tot} is the total error, including both calibration errors and range-noise, of ToF plane π_j as it appears in the i-th RGB cameras.

If the RGB cameras are not too far from the ToF camera, then the range errors tend to be cancelled in the reprojection. This is evident from table 3, in which the total errors, on average, are not much greater than the calibration errors in table 2. It is, however, clear from fig. 7, that the tails of the total error distributions are greatly increased by outliers in the ToF data-stream. Although errors up to 3 pixels are shown in distributions for the sake of
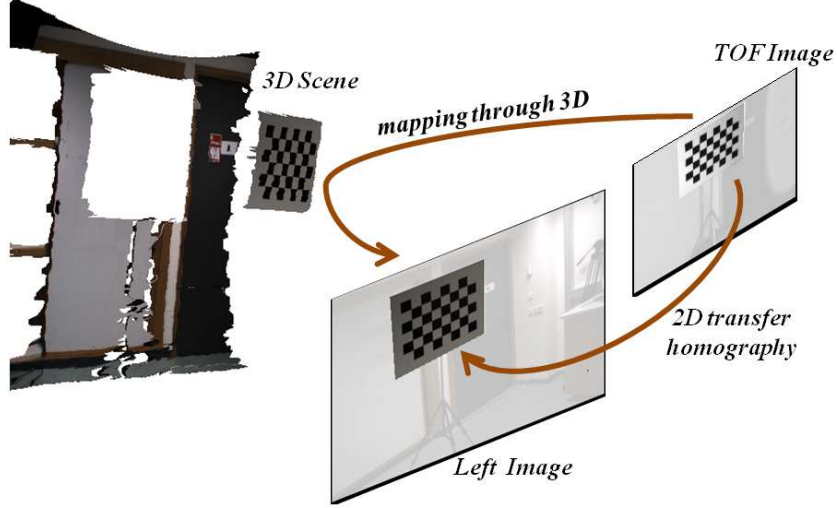
Figure 6: The 2-D transfer homography $\boldsymbol{T}_\ell$ describes the chessboard mapping from the TOF frame to the left frame. In principle, this mapping should be equivalent to a mapping via the range data, i.e. from the TOF frame to the measured 3-D plane, and back to the left frame.

clarity, maximum errors of more than five pixels are common (see table 3).

Set	Mean	Median	Max	Count
A	0.74	0.71	6.07	39859
B	1.48	0.98	20.69	32159
C	0.63	0.54	6.51	29442

Table 3: Total error (19), measured by projecting the raw TOF points $\boldsymbol{Q}_i$ to the left and right RGB images (1624×1224) of the *respective* systems $i = 1, 2, 3$. These figures characterize the raw data that is produced by the live system. This data includes all outliers in the TOF data-stream, which gives rise to at least one error of 20 pixels here.

When the raw TOF points are reprojected to a *different* system, a high total error is expected, because the sensor's range-noise may be viewed 'from the side'. Indeed, fig. 8 shows that a substantial proportion of the TOF points reproject with total errors in excess of ten pixels. These kind of gross errors are characteristic of the TOF data, owing to the inevitable presence of absorbing and scattering surfaces in a typical scene.
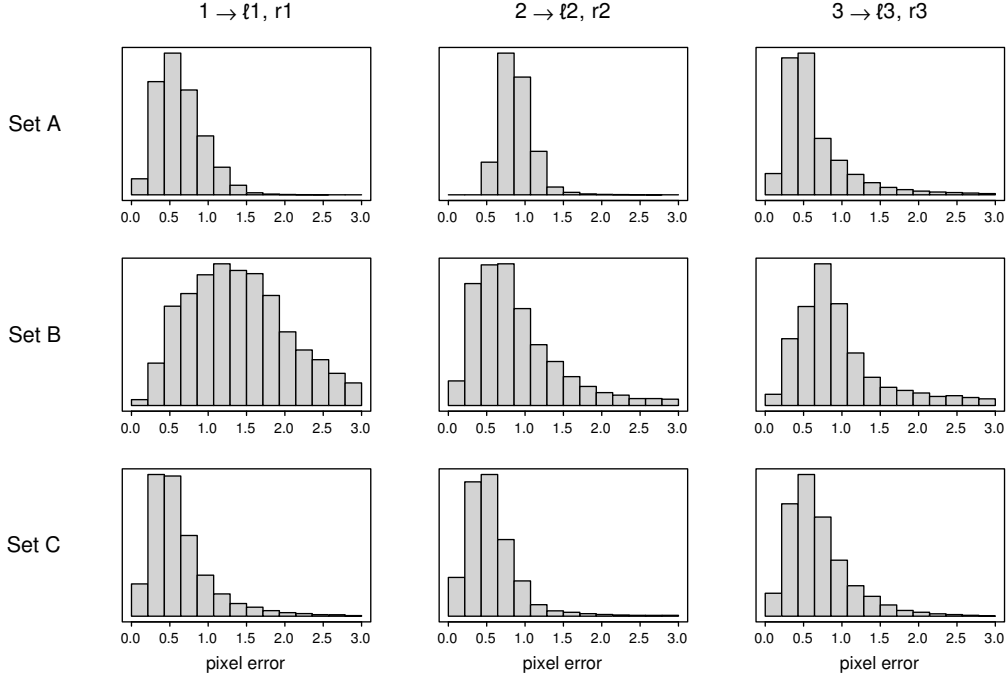
19

Figure 7: Total error (19), measured by projecting the raw ToF points $\boldsymbol{Q}_i$ to the left and right RGB images (1624×1224) of the *respective* systems $i = 1, 2, 3$. These distributions have longer and heavier tails than those of the corresponding calibration errors, shown in fig. 4.

In fact, calibration error and total error are both influenced by surface orientation. For example, the higher errors in Set B of the data (see tables 2 and 3) are due to the presence of some very slanted boards, in both the fitting and evaluation data-sets; this causes two problems, as follows. Firstly, the images of these boards are very foreshortened in some systems. This leads to increased *calibration* error, because the vertices are hard to detect in the corresponding ToF images. Secondly, the strength of the reflected IR signal is reduced whenever the surface is oblique to a given ToF camera. If the surface is also absorbent (like the black squares of the board), then the *total* error is greatly increased, due to the combined effects of scattering and absorption.

It is possible to understand these results more fully by examining the distribution of the total error across individual boards. Figure 9 shows the distribution for a board reprojected to same/different systems (i.e. part of the data from figs. 7 & 8). There is a relatively smooth gradient of error
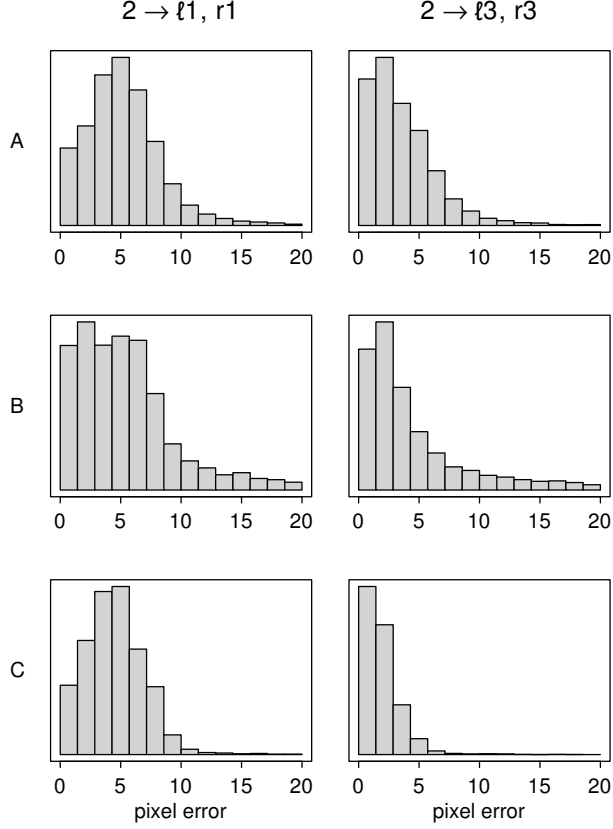
Figure 8: Total error (19), measured by projecting the fitted TOF points $\boldsymbol{Q}_2^\pi$ to the left and right RGB images (1624 × 1224) of two *different* systems, $i = 1, 3$. The range errors are emphasized by the difference in viewpoints between the two systems. Average error is now around five pixels, and the noisiest TOF points reproject with tens of pixels of error.

across the board, which is attributable to errors in the fitting of plane π_j, and in the estimation of the camera parameters. In addition to these effects, it is clear that the gross errors are correlated with the black squares of the board, which reflect too little of the TOF signal. For instance, in the bottom example of figure 9, the mean reprojection error associated with the black squares is 10.05 pixels, whereas the white squares are associated with a mean error of 3.55 pixels. The effect is particularly noticeable, as expected from the histograms, when reprojecting to different camera systems.
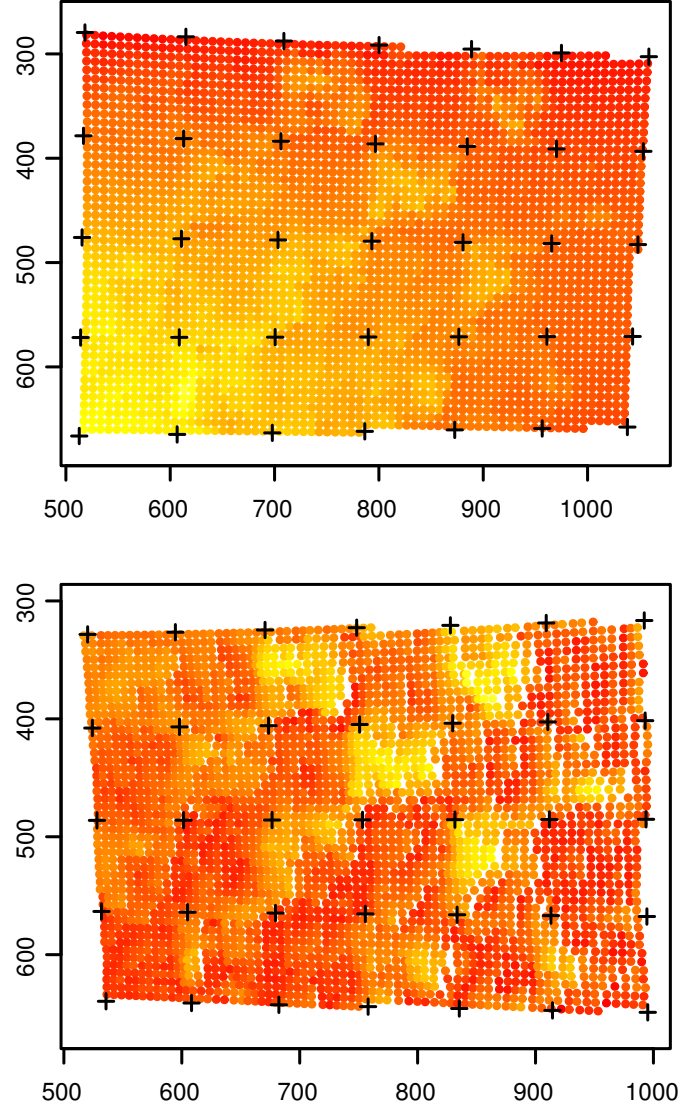
Figure 9: 3-D ToF pixels on an example calibration board, projected into an RGB camera in the same (top) and in a different (bottom) system. Each pixel is colour-coded according to the relative total-error (19) in each case (larger errors are shown in yellow). Black crosses mark the detected vertices in the RGB image. The black squares of the board are associated with larger reprojection errors, particularly after reprojecting to a different system (mean error in the bottom example is 10.05px for black squares, and 3.55px for white).

4.3. Comparison of homography and similarity transformations

It was argued in section 2 that the 3-D homography model is an appropriate way to compensate for miscalibrations of the ToF and stereo systems (and indeed it allows the stereo system to be left uncalibrated, if preferred [33]). This claim is tested in the following experiments, using data images from two new data sets.

The most obvious alternative to the homography is a rigid motion and scaling; i.e. a 3-D *similarity* transformation. Specifically, the 4×4 matrix $\boldsymbol{H}$ in (3) can be replaced by the similarity transformation

$$ \boldsymbol{S} = \begin{pmatrix} \sigma \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{pmatrix} \tag{20} $$

where σ is a (positive) scalar, $\boldsymbol{R}$ is a 3×3 rotation matrix, and $\boldsymbol{t}$ is a 3×1 translation vector. Hence there are only seven degrees of freedom, in contrast to the fifteen of the homography. Geometrically, the similarity model can be interpreted as a rigid transformation between the ToF and stereo systems, where the baseline distance of the latter is unknown.

In principle, a homography should always result in equal or lower error, because it includes the similarity transformation as a special case. In particular, a homography can always be written as $\boldsymbol{H} = \boldsymbol{SD}$ where the deformation $\boldsymbol{D}$ is composed of an affinity and an elation [12]. However, there are two practical issues to consider. Firstly, the quality of the actual estimate in each case; and secondly, the possible danger of over-fitting with the homography.

It was argued in section 2 that the 3-D homography model is an appropriate way to compensate for miscalibrations of the ToF and stereo systems (and also allows the stereo system to be left uncalibrated, if preferred [33]). In particular the deformation $\boldsymbol{D}$ can help to account for locally linear depth bias in the ToF camera [46]. Hence the additional eight degrees of freedom in the homography model allow for an overall approximation to the per-pixel corrections performed by Cui et al. [32].

The procedure to estimate the optimal similarity (20) is analogous to that used to estimate the homography in section 2. Instead of DLT, an initial estimate is obtained from the Procrustes algorithm [47, 48], using three or more correspondences. The reprojection error is then minimized using the Levenberg-Marquardt procedure, as with the homography. Note that the rotation $\boldsymbol{R}$ in (20) is appropriately parameterized by the Rodrigues formula.

The results of these experiments are shown in figure 10. In every TOF+2RGB system, and in both capture sessions, the homography results in lower error than the similarity. The mean reprojections errors (in pixels) were 0.22 vs. 0.65, and 0.46 vs. 1.06, for the respective capture sessions. The generally higher figures from the second set are due to a broader distribution of board-poses in the evaluation data-set.

These experiments show that the 3-D homography transformation is a suitable model for TOF/RGB alignment, as argued in section 2. It may also be noted that the homography is effectively easier to estimate than the similarity, as there is no need to maintain the orthogonality of $\boldsymbol{R}$ during the final minimization.
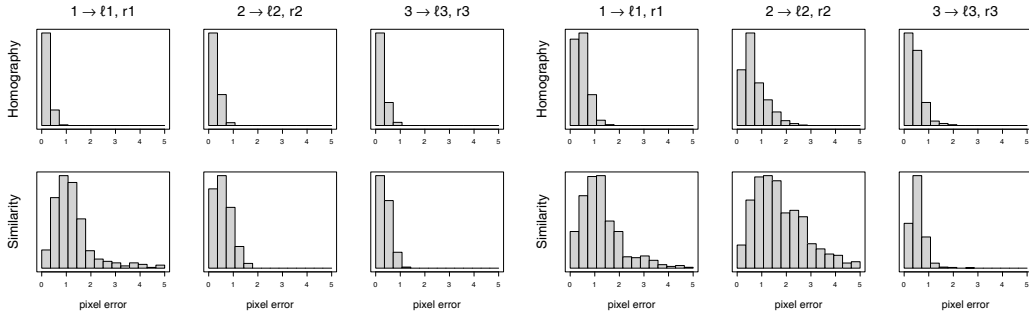

Figure 10: Calibration error (17), measured by projecting the fitted TOF points $\boldsymbol{Q}_i^\pi$ to the left and right RGB images (1624 × 1224) of the *respective* systems $i = 1, 2, 3$. The TOF/stereo alignment was computed via a 3-D homography (top), or a similarity transformation (bottom). It is clear that the former is superior, as described in section 4.3. The left and right 2 × 3 blocks represent different capture-sessions, as described in the text.

4.4. Cross-calibration for 3-D reconstruction and rendering

The proposed cross-calibration methodology also allows a fully-textured 3-D model to be segmented and rendered. This is because the depth-data are automatically assigned to pixels in the RGB images. Figure 11 shows reconstruction instances with and without texture from a 360° reconstruction, obtained from the cross-calibration of four TOF+2RGB systems. Meshing was performed by the standard Poisson reconstruction method [49], followed by simple triangle-based texture mapping.

More dense reconstruction can be achieved by exploiting the full resolution of stereo cameras, as described elsewhere [11]. Figure 12 shows the difference between the raw reconstruction obtained from the cross-calibration

of a TOF+2RGB system, and that obtained after using the TOF data in a subsequent stereo algorithm [11]. A high-resolution depth map is produced, making full use of the high-resolution RGB cameras.
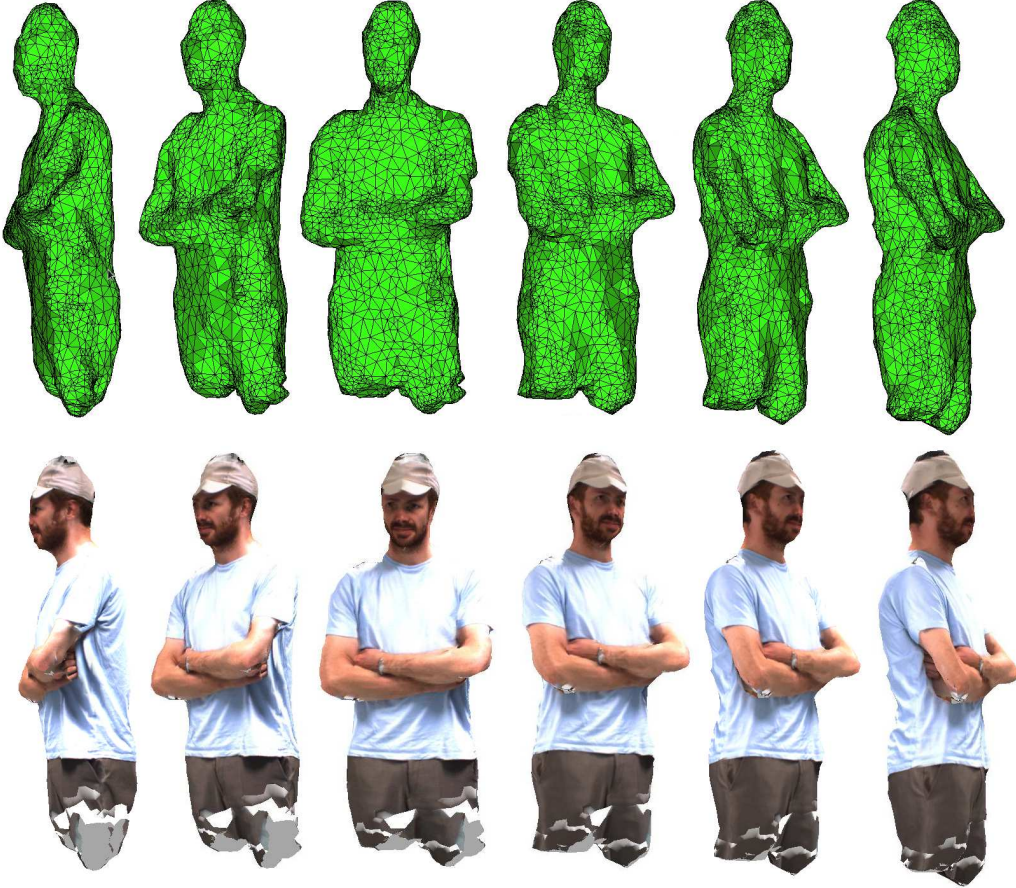

Figure 11: Examples of a segmented 360° reconstruction, after Poisson meshing, with and without RGB texture.

5. Conclusion

It has been shown that 3-D projective transformations can be used to cross-calibrate a system of TOF and stereo reconstructions. A practical method for computing these transformations, based on geometric principles, has been introduced. This calibration procedure has been extended to a
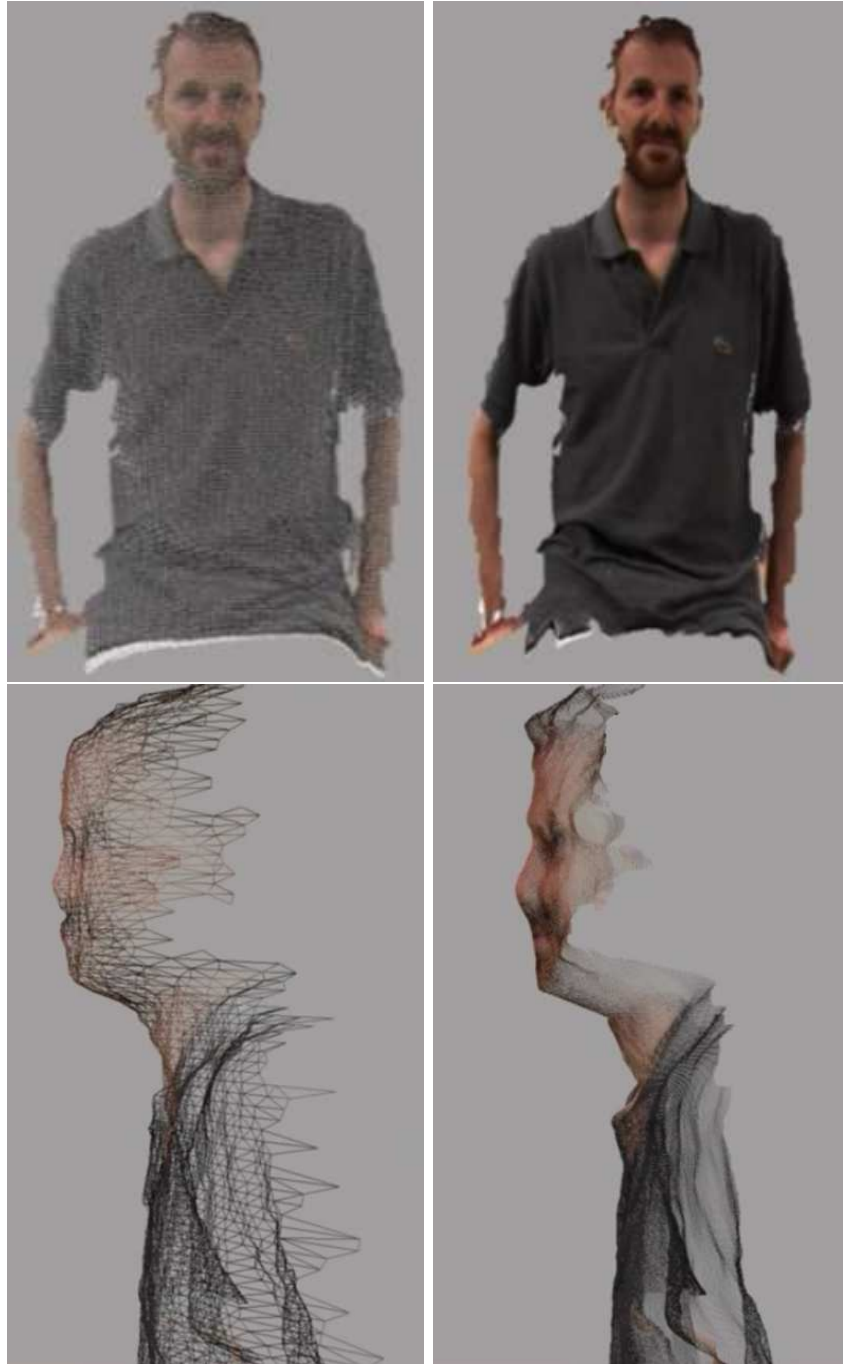
Figure 12: Cross-calibration for 3-D reconstruction. *Left:* The raw mesh, with textured edges, obtained directly from our ToF+2RGB cross-calibration method. *Right:* The textured high-density point cloud, obtained after applying the ToF-stereo fusion method of [11] to the cross-calibration results.

nine-camera network of six RGB and three TOF cameras, and evaluated in detail.

A clear distinction has been made between *calibration error*, which is due to imperfect camera and image models, versus *total error*, which incorporates TOF-specific noise and biases. It has been shown that these errors can be separated geometrically, and used to characterize the performance of a TOF/RGB camera network. The overall performance of the system has been visualized in segmented 360° 3-D reconstructions. This type of reconstruction is challenging to compute, and shows that the system is more than adequate as a basis for scene-segmentation tasks.

The accuracy of the method presented here is somewhat limited by the relatively poor localization of points, even when spatially interpolated, in the TOF images. The detection of standard chequerboard patterns, in TOF images, has been discussed elsewhere [7]. The design of more convenient calibration patterns, for TOF cameras, would be a worthwhile direction for future research. Meanwhile, in mitigation, the spatial resolution of TOF cameras will continue to increase.

Future work should also consider the distribution of range-errors in 3-D, and how this can be used to design custom meshing and surface reconstruction algorithms for time-of-flight data. This, in conjunction with the RGB textures provided by the present method, would lead to a comprehensive approach to 3-D reconstruction, rendering and scene-understanding.

References

[1] D. Hoiem, A. Efros, M. Hebert, Recovering surface layout from an image, International Journal of Computer Vision 75 (1) (2007) 151–172.

[2] R. Tron, R. Vidal, A benchmark for the comparison of 3-d motion segmentation algorithms, in: Proc. CVPR'07, 2007, pp. 1–8.

[3] S. Savarese, T. Tuytelaars, L. V. Gool, Special issue on 3D representation for object and scene recognition, Computer Vision and Image Understanding 113 (12) (2009) 1181–1182.

[4] Z. Liu, M. Beetz, D. Cremers, J. Gall, W. Li, D. Pangercic, J. Sturm, Y.-W. Tai, Special issue on visual understanding and applications with RGB-D cameras, Journal of Visual Communication and Image Representation 25 (1).

[5] A. Alahi, P. Vandergheynst, M. Bierlaire, M. Kunt, Cascade of descriptors to detect and track objects across any network of cameras, Computer Vision and Image Understanding 114 (6) (2010) 624–640.

[6] Mesa Imaging, SR4000 Swiss Ranger manual, `http://www.mesa-imaging.ch` (2010).

[7] M. Hansard, R. Horaud, M. Amat, G. Evangelidis, Automatic detection of calibration grids in time-of-flight images, Computer Vision and Image Understanding 121 (2014) 108–118.

[8] A. Verri, V. Torre, Absolute depth estimate in stereopsis, J. Optical Society of America A 3 (3) (1986) 297–299.

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

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

[11] V. Gandhi, J. Cech, R. Horaud, High-resolution depth maps based on TOF-stereo fusion, in: Proc. ICRA, 2012, pp. 4742–4749.

[12] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2000.

[13] R. Nair, K. Ruhl, S. Meister, H. Schäfer, C. S. Garbe, M. Eisemann, M. Magnor, D. Kondermann, A survey on time-of-flight stereo fusion, in: M. Grzegorzek, C. Theobalt, R. Koch, A. Kolb (Eds.), Time-of-Flight and Depth Imaging, Vol. 8200 of LNCS, Springer, 2013, Ch. A Survey on Time-of-Flight Stereo Fusion, pp. 105–127.

[14] B. Huhle, S. Fleck, A. Schilling, Integrating 3D time-of-flight camera data and high resolution images for 3DTV applications, in: Proc. 3DTV, 2007, pp. 1–4.

[15] B. Bartczak, R. Koch, Dense depth maps from low resolution time-of-flight depth and high resolution color views, in: Proc. Int. Symp. on Visual Computing (ISVC), 2009, pp. 228–239.

[16] A. Bleiweiss, M. Werman, Fusing time-of-flight depth and color for real-time segmentation and tracking, in: Proc. DAGM Workshop on Dynamic 3D Imaging, 2009, pp. 58–69.

[17] C. Beder, I. Schiller, R. Koch, Photoconsistent relative pose estimation between a PMD 2D3D-camera and multiple intensity cameras, in: Proc. Symp. of the German Association for Pattern Recognition (DAGM), 2008, pp. 264–273.

[18] I. Schiller, C. Beder, R. Koch, Calibration of a PMD camera using a planar calibration object together with a multi-camera setup, in: Int. Arch. Soc. Photogrammetry, Remote Sensing and Spatial Information Sciences XXI, 2008, pp. 297–302.

[19] R. Koch, I. Schiller, B. Bartczak, F. Kellner, K. Köser, MixIn3D: 3D mixed reality with ToF-camera, in: Proc. DAGM Workshop on dynamic 3D imaging, 2009, pp. 126–141.

[20] J. Zhu, L. Wang, R. Yang, J. Davis, Z. Pan, Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps, IEEE Trans. on PAMI 33 (7) (2011) 1400–1414.

[21] J. Zhu, L. Wang, R. G. Yang, J. Davis, Fusion of time-of-flight depth and stereo for high accuracy depth maps, in: Proc. CVPR, 2008, pp. 1–8.

[22] Y. Wang, Y. Jia, A fusion framework of stereo vision and kinect for high-quality dense depth maps, in: ACCV 2012 Workshops, 2012, pp. 109–120.

[23] J. Wu, Y. Zhou, H. Yu, Z. Zhang, Improved 3D depth image estimation algorithm for visual camera, in: Proc. International Congress on Image and Signal Processing, 2009, pp. 1–4.

[24] J. Mure-Dubois, H. Hügli, Fusion of time-of-flight camera point clouds, in: ECCV Workshop on Multi-Camera and Multi-modal Sensor Fusion Algorithms and Applications, 2008.

[25] M. Silva, R. Ferreira, J. Gaspar, Camera Calibration using a Color-Depth Camera: Points and Lines Based DLT including Radial Distortion, in: Proc. of IROS Workshop in Color-Depth Camera Fusion in Robotics, 2012.

[26] C. Zhang, Z. Zhang, Calibration between depth and color sensors for commodity depth cameras, in: Proc. of the 2011 IEEE Int. Conf. on Multimedia and Expo, ICME '11, 2011, pp. 1–6.

[27] Z. Zhang, Flexible camera calibration by viewing a plane from unknown orientations, in: Proc. ICCV'99, 1999, pp. 666–673.

[28] Y. Kim, D. Chan, C. Theobalt, S. Thrun, Design and calibration of a multi-view TOF sensor fusion system, in: Proc. CVPR Workshop on time-of-flight Camera based Computer Vision, 2008, pp. 1–7.

[29] D. Herrera, J. Kannala, J. Heikila, Joint Depth and Color Camera Calibration with Distortion Correction, IEEE Trans. PAMI 34 (10) (2012) 2058–2064.

[30] A. Teichman, S. Miller, S. Thrun, Unsupervised intrinsic calibration of depth sensors via slam, in: Proceedings of Robotics: Science and Systems, 2013.

[31] I. V. Mikhelsona, P. G. Leea, A. V. Sahakiana, Y. Wua, A. K. Katsaggelosa, Automatic, fast, online calibration between depth and color cameras, Journal of Visual Communication and Image RepresentationIn Press (available online).
URL http://dx.doi.org/10.1016/j.bbr.2011.03.031

[32] Y. Cui, S. Schuon, S. Thrun, D. Stricker, C. Theobalt, Algorithms for 3d shape scanning with a depth camera, IEEE Trans. PAMI 35 (5) (2013) 1039–1050.

[33] M. Hansard, R. Horaud, M. Amat, S. Lee, Projective Alignment of Range and Parallax Data, in: Proc. CVPR, 2011, pp. 3089–3096.

[34] M. Hansard, S. Lee, O. Choi, R. Horaud, Time-of-Flight Cameras: Principles, Methods and Applications, Springer, 2013.

[35] C. Beder, B. Bartczak, R. Koch, A comparison of PMD-cameras and stereo-vision for the task of surface reconstruction using patchlets, in: Proc. CVPR, 2007, pp. 1–8.

[36] M. Hebert, E. Krotkov, 3D measurements from imaging laser radars: How good are they?, Image and Vision Computing 10 (3) (1992) 170–178.

[37] C. Wang, H. Tanahasi, H. Hirayu, Y. Niwa, K. Yamamoto, Comparison of local plane fitting methods for range data, in: Proc. CVPR, 2001, pp. 663–669.

[38] Y. Kanazawa, K. Kanatani, Reliability of Plane Fitting by Range Sensing, in: Proc. ICRA, 1995, pp. 2037–2042.

[39] K. Pathak, N. Vaskevicius, A. Birk, Revisiting uncertainty analysis for optimum planes extracted from 3D range sensor point-clouds, in: Proc. ICRA, 2009, pp. 1631–1636.

[40] G. Csurka, D. Demirdjian, R. Horaud, Finding the collineation between two projective reconstructions, Computer Vision and Image Understanding 75 (3) (1999) 260–268.

[41] W. Förstner, Uncertainty and projective geometry, in: E. Bayro-Corrochano (Ed.), Handbook of Geometric Computing, Springer, 2005, pp. 493–534.

[42] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, Numerical Recipes in C, 2nd Edition, Cambridge University Press, 1992.

[43] R. Hartley, P. Sturm, Triangulation, Computer Vision and Image Understanding 68 (2) (1997) 146–157.

[44] K. Arun, T. S. Huang, S. D. Blostein, Least-squares fitting of two 3-D point sets, IEEE Trans. PAMI 9 (5) (1987) 698–700.

[45] G. D. Evangelidis, E. Z. Psarakis, Parametric image alignment using enhanced correlation coefficient maximisation, IEEE Trans. PAMI 30 (10) (2010) 1858–1865.

[46] S. Foix, G. Alenya, C. Torras, Lock-in Time-of-Flight (ToF) Cameras: A Survey, IEEE Sensors 11 (9) (2011) 1917–1926.

[47] P. Schönemann, R. Carroll, Fitting one matrix to another under choice of a central dilation and a rigid motion, Psychometrika 35 (2) (1970) 245–255.

[48] B. Horn, H. Hilden, S. Negahdaripour, Closed-form solution of absolute orientation using orthonormal matrices, J. Optical Society of America A 5 (7) (1988) 1127–1135.

[49] M. M. Kazhdan, M. Bolitho, H. Hoppe, Poisson surface reconstruction, in: Symposium on Geometry Processing, 2006, pp. 61–70.