Chapter 4
Alignment of Time-of-Flight and Stereoscopic Data

Abstract An approximately Euclidean representation of the visible scene can be obtained directly from a time-of-flight camera. An uncalibrated binocular system, in contrast, gives only a projective reconstruction of the scene. This chapter analyzes the geometric mapping between the two representations, without requiring an intermediate calibration of the binocular system. The mapping can be found by either of two new methods, one of which requires point correspondences between the range and color cameras, and one of which does not. It is shown that these methods can be used to reproject the range data into the binocular images, which makes it possible to associate high-resolution color and texture with each point in the Euclidean representation. The extension of these methods to multiple time-of-flight system is demonstrated, and the associated problems are examined. An evaluation metric, which distinguishes calibration error from combined calibration and depth error, is developed. This metric is used to evaluate a system that is based on three time-of-flight cameras.

Keywords Depth and color combination · Projective alignment · Time-of-Flight camera calibration · Multicamera systems

4.1 Introduction

It was shown in the preceding chapter that time-of-flight (ToF) cameras can be geometrically calibrated by standard methods. This means that each pixel records an estimate of the scene distance (range) along the corresponding ray, according to the principles described in Chap. 1. The 3-D structure of a scene can also be reconstructed from two or more ordinary images, via the *parallax* 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 color and texture can be obtained.

M. Hansard et al., *Time-of-Flight Cameras*, SpringerBriefs in Computer Science,
DOI: 10.1007/978-1-4471-4658-2_4, © Miles Hansard 2013

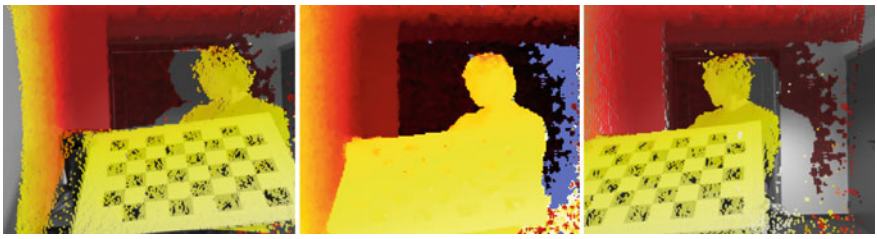

Fig. 4.1 The central panel shows a range image, color-coded according to depth (the *blue* region is beyond the far limit of the device). The *left* and *right* cameras were aligned to the ToF system, using the methods described here. Each 3-D range pixel is reprojected into the high-resolution left and right images (untinted regions were occluded, or otherwise missing, from the range images). Note the large difference between the binocular views, which would be problematic for dense stereo-matching algorithms. It can also be seen that the ToF information is noisy, and of low resolution

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 [23].

The range data, on the other hand, are often very noisy (and, for very scattering surfaces, incomplete), as described in Chap. 1. The spatial resolution of current ToF sensors is relatively low, the depth range is limited, and the luminance signal may be unusable for rendering. It should also be recalled that ToF cameras of the type used here [19] cannot be used in outdoor lighting conditions. These considerations lead to the idea of a *mixed* color and ToF system [18] as shown in Figs. 4.1 and 4.2. Such a system could, in principle, be used to make high-resolution Euclidean reconstructions, with full photometric information [17]. The task of camera calibration would be simplified by the ToF camera, while the visual quality of the reconstruction would be ensured by the color cameras.

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 color images, must be obtained. This chapter 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. The alignment problem has been addressed previously, by fully calibrating the binocular system, and then aligning the two reconstructions by a rigid transformation [6, 12, 27, 28]. This approach can be extended in two ways. First, it is possible to optimize over an explicit parameterization of the camera matrices, as in the work of Beder et al. [3] and Koch et al. [16]. The relative position and orientation of all cameras can be estimated by this method. Second, it is possible to minimize an intensity cost between the images and the luminance signal of the ToF camera. This method estimates the photometric, as well as geometric, relationships between the different

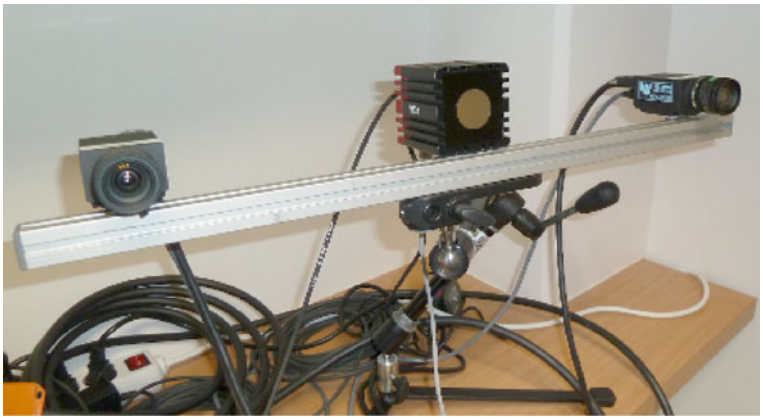

Fig. 4.2 A single ToF+2RGB system, as used in this chapter, with the ToF camera in the center of the rail

cameras [13, 22, 25]. A complete calibration method, which incorporates all of these considerations, is described by Lindner et al. [18].

The approaches described above, while capable of producing good results, have some limitations. First, there may be residual distortions in the range data, that make a rigid alignment impossible [15]. Second, these approaches require the binocular system to be fully calibrated, and recalibrated after any movement of the cameras. This requires, for best results, many views of a known calibration object. Typical view-synthesis applications, in contrast, require only a weak calibration of the cameras. One way to remove the calibration requirement is to perform an essentially 2-D registration of the different images [1, 4]. This, however, can only provide an instantaneous solution, because changes in the scene structure produce corresponding changes in the image-to-image mapping.

An alternative approach is proposed here. It is hypothesized that the ToF reconstruction is approximately Euclidean. This means that an *uncalibrated* binocular reconstruction can be mapped directly into the Euclidean frame, by a suitable 3-D projective transformation. This is a great advantage for many applications, because automatic uncalibrated reconstruction is relatively easy. Furthermore, although the projective model is much more general than the rigid model, it preserves many important relationships between the images and the scene (e.g., epipolar geometry and incidence of points on planes). Finally, if required, the projective alignment can be upgraded to a fully calibrated solution, as in the methods described above.

It is emphasized that the goal of this work is *not* to achieve the best possible photogrammetric reconstruction of the scene. Rather, the goal is to develop a practical way to associate color and texture information to each range point, as in Fig. 4.1. This output is intended to use in view-synthesis applications.

This chapter is organized as follows. Section 4.2.1 briefly reviews some standard material on projective reconstruction, while Sect. 4.2.2 describes the representation

of range data in the present work. The chief contributions of the subsequent sections are as follows: Sect. 4.2.3 describes a *point-based* method that maps an ordinary *projective* reconstruction of the scene onto the corresponding range representation. This does not require the color cameras to be calibrated (although it may be necessary to correct for lens distortion). Any planar object can be used to find the alignment, provided that image features can be matched across all views (including that of the ToF camera). Section 4.2.4 describes a dual *plane-based* method, which performs the same projective alignment, but that does not require any point matches between the views. Any planar object can be used, provided that it has a simple polygonal boundary that can be segmented in the color and range data. This is a great advantage, owing to the very low resolution of the luminance data provided by the ToF camera (176×144 here). This makes it difficult to automatically extract and match point descriptors from these images, as described in Chap. 3. Furthermore, there are ToF devices that do not provide a luminance signal at all. Section 4.2.5 addresses the problem of multisystem alignment. Finally, Sect. 4.3 describes the accuracy than can be achieved with a three ToF+2RGB system, including a new error metric for ToF data in Sect. 4.3.2. Conclusions and future directions are discussed in Sect. 4.4.

4.2 Methods

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.,

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_\triangle \\ P_4 \end{pmatrix} \quad \text{and} \quad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_\triangle \\ U_4 \end{pmatrix} \tag{4.1}$$

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 nonzero 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

$$\boldsymbol{q} \simeq \boldsymbol{C}\boldsymbol{Q} \quad \text{where} \quad \boldsymbol{C} = \big(\boldsymbol{A}_{3\times 3} \,|\, \boldsymbol{b}_{3\times 1}\big). \tag{4.2}$$

The left and right color cameras $\boldsymbol{C}_\ell$ and $\boldsymbol{C}_r$ are similarly defined, e.g., $\boldsymbol{p}_\ell \simeq \boldsymbol{C}_\ell \boldsymbol{P}$. Table 4.1 summarizes the geometric objects that will be aligned.

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

Table 4.1 Summary of notations in the left, right, and range systems

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

$$\boldsymbol{Q} \simeq \boldsymbol{H}\boldsymbol{P} \quad \text{and} \quad \boldsymbol{V} \simeq \boldsymbol{H}^{-\top}\boldsymbol{U}. \tag{4.3}$$

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, e.g., $\boldsymbol{p}_\ell \simeq \left(\boldsymbol{C}_\ell \boldsymbol{H}^{-1}\right)\boldsymbol{Q}$, where $\boldsymbol{H}^{-1}\boldsymbol{Q} \simeq \boldsymbol{P}$.

4.2.1 Projective 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 [10]. In particular, from $\boldsymbol{F}$ and the epipole $\boldsymbol{e}_r$, the cameras can be defined as

$$\boldsymbol{C}_\ell \simeq \left(\boldsymbol{I} \,|\, \boldsymbol{0}_3\right) \text{ and } \boldsymbol{C}_r \simeq \left((\boldsymbol{e}_r)_\times \boldsymbol{F} + \boldsymbol{e}_r \boldsymbol{g}^\top \,\middle|\, \gamma \boldsymbol{e}_r\right). \tag{4.4}$$

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.

4.2.2 Range Fitting

The ToF camera $\boldsymbol{C}$ provides the distance ρ of each scene point from the camera center, as well as its image coordinates $\boldsymbol{q} = (x, y, 1)$. The back projection of this point into the scene is

$$\boldsymbol{Q}_\triangle = \boldsymbol{A}^{-1}\left((\rho/\alpha)\,\boldsymbol{q} - \boldsymbol{b}\right) \quad \text{where} \quad \alpha = \left|\boldsymbol{A}^{-1}\,\boldsymbol{q}\right|. \tag{4.5}$$

Hence, the point $(\boldsymbol{Q}_\triangle, 1)^\top$ is at distance ρ from the optical center $-\boldsymbol{A}^{-1}\boldsymbol{b}$, 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 [2].

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 (4.5) directly. A better approach is to segment the image of the plane in each ToF camera (using 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

$$\rho^\pi = \frac{\boldsymbol{V}_\triangle^\top \boldsymbol{A}^{-1}\boldsymbol{b} - V_4}{(1/\alpha)\, \boldsymbol{V}_\triangle^\top \boldsymbol{A}^{-1}\boldsymbol{q}} \tag{4.6}$$

where $|V_4|$ is the distance of the plane to the camera center, 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 (4.5).

The choice of plane-fitting method is affected by two issues. First, there may be very severe outliers in the data, due to the photometric and geometric errors described in Chap. 1. Second, the noise-model should be based on the pinhole model, which means that perturbations occur radially along visual directions, which are not (in general) perpendicular to the observed plane [11, 24]. Several plane-fitting methods, both iterative [14] and noniterative [20], have been proposed for the pinhole model. The outlier problem, however, is often more significant. Hence, in practice, a RANSAC-based method is often the most effective.

4.2.3 Point-Based Alignment

It is straightforward to show that the transformation $\boldsymbol{H}$ in (4.3) 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 [10], which fits $\boldsymbol{H}$ to *all* of the data. This method is based on the fact that if

$$\boldsymbol{P}' = \boldsymbol{H}\boldsymbol{P} \tag{4.7}$$

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 [5]. This results in $\binom{4}{2} = 6$ interdependent constraints per point. It is convenient to write these homogeneous equations as

$$\begin{pmatrix} Q_4 \boldsymbol{P}'_\triangle - P'_4 \boldsymbol{Q}_\triangle \\ \boldsymbol{Q}_\triangle \times \boldsymbol{P}'_\triangle \end{pmatrix} = \boldsymbol{0}_6. \tag{4.8}$$

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 (4.8) is simply the distance between the points.

Following Förstner [7], the left-hand side of (4.8) can be expressed as $\big(\boldsymbol{Q}\big)_\wedge \boldsymbol{P}'$ where

$$\big(\boldsymbol{Q}\big)_\wedge = \begin{pmatrix} Q_4 \boldsymbol{I}_3 & -\boldsymbol{Q}_\triangle \\ \big(\boldsymbol{Q}_\triangle\big)_\times & \boldsymbol{0}_3 \end{pmatrix} \tag{4.9}$$

is a 6×4 matrix, and $\big(\boldsymbol{Q}_\triangle\big)_\times \boldsymbol{P}_\triangle = \boldsymbol{Q}_\triangle \times \boldsymbol{P}_\triangle$, as usual. The Eq. (4.8) can now be written in terms of (4.7) and (4.9) as

$$\big(\boldsymbol{Q}\big)_\wedge \boldsymbol{H}\boldsymbol{P} = \boldsymbol{0}_6. \tag{4.10}$$

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

$$\Big(\boldsymbol{P}^\top \otimes \big(\boldsymbol{Q}\big)_\wedge\Big)\boldsymbol{h} = \boldsymbol{0}_6 \quad \text{where} \quad \boldsymbol{h} = \mathrm{vec}\big(\boldsymbol{H}\big). \tag{4.11}$$

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 $\big(\boldsymbol{P}_k^\top \otimes (\boldsymbol{Q}_k)_\wedge\big)$ are stacked together, to give the complete system

$$\begin{pmatrix} \boldsymbol{P}_1^\top \otimes \big(\boldsymbol{Q}_1\big)_\wedge \\ \vdots \\ \boldsymbol{P}_{MN}^\top \otimes \big(\boldsymbol{Q}_{MN}\big)_\wedge \end{pmatrix} \boldsymbol{h} = \boldsymbol{0}_{6MN} \tag{4.12}$$

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 (4.12). This solution, which minimizes an *algebraic error* [10], is the singular vector corresponding to the smallest singular value of the matrix. In the minimal case, $M = 1, N = 5$, the matrix would be 30×16. Note that, the point coordinates should be transformed, to ensure that (4.12) is numerically well conditioned [10]. 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 DLT method, in practice, gives a good approximation $\boldsymbol{H}_{\mathrm{DLT}}$ of the homography (4.3). 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,

$$E_\ell(\boldsymbol{C}_\ell) = \sum_{k=1}^{MN} D\big(\boldsymbol{C}_\ell \boldsymbol{Q}_k,\ \boldsymbol{p}_{\ell k}\big)^2 \tag{4.13}$$

where $D(\boldsymbol{p}, \boldsymbol{q}) = |\boldsymbol{p}_\triangle / p_3 - \boldsymbol{q}_\triangle / q_3|$. A 12-parameter optimization of (4.13), starting with $\boldsymbol{C}_\ell \leftarrow \boldsymbol{C}_\ell \boldsymbol{H}_{\mathrm{DLT}}^{-1}$, can be performed by the Levenberg-Marquardt algorithm [21].

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,

$$E\left(\boldsymbol{H}^{-1}\right) = E_\ell\left(\boldsymbol{C}_\ell \boldsymbol{H}^{-1}\right) + E_r\left(\boldsymbol{C}_r \boldsymbol{H}^{-1}\right) \tag{4.14}$$

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

The difference between the separate (4.13) and joint (4.14) 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 [10]. 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 (4.14) should be performed. This will also preserve the original binocular triangulation, provided that a projective-invariant method was used [9]. 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 $(\boldsymbol{C}_\ell^\star)^+$ is the generalized inverse. The right epipole is obtained from $\boldsymbol{e}_r^\star = \boldsymbol{C}_r^\star \boldsymbol{d}_\ell^\star$, where $\boldsymbol{d}_\ell^\star$ represents the nullspace $\boldsymbol{C}_\ell^\star \boldsymbol{d}_\ell^\star = \boldsymbol{0}_3$.

4.2.4 Plane-Based Alignment

The DLT algorithm of Sect. 4.2.3 can also be used to recover $\boldsymbol{H}$ from matched *planes*, rather than matched points. Equation (4.10) becomes

$$\left(\boldsymbol{V}\right)_\wedge \boldsymbol{H}^{-\top} \boldsymbol{U} = \boldsymbol{0}_6 \tag{4.15}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ represent the estimated coordinates of the same plane in the parallax and range reconstructions, respectively. The estimation procedure is identical to that in Sect. 4.2.3, but with $\text{vec}(\boldsymbol{H}^{-\top})$ as the vector of unknowns.

This method, in practice, produces very poor results. The chief reason that obliquely viewed planes are foreshortened, and therefore hard to detect/estimate, in the low-resolution ToF images. It follows that the calibration data set is biased towards fronto-parallel planes.[1] This bias allows the registration to slip sideways, perpendicular to the primary direction of the ToF camera. The situation is greatly

[1] The point-based algorithm is unaffected by this bias, because the scene is ultimately 'filled' with points, regardless of the contributing planes.

improved by assuming that the *boundaries* of the planes can be detected. For example, if the calibration object is rectangular, then the range projection of the plane $\boldsymbol{V}$ is bounded by four edges $\bar{\boldsymbol{v}}_i$, where $i = 1, \ldots 4$. Note that, these are detected as *depth* edges, and so no luminance data are required. The edges, represented as lines $\bar{\boldsymbol{v}}_i$, back project as the faces of a pyramid,

$$\overline{\boldsymbol{V}}_i = \boldsymbol{C}^\top \bar{\boldsymbol{v}}_i = \begin{pmatrix} \overline{\boldsymbol{V}}_{i\triangle} \\ 0 \end{pmatrix}, \quad i = 1, \ldots, L \tag{4.16}$$

where $L = 4$ in the case of a quadrilateral projection. These planes are linearly dependent, because they pass through the center of projection; hence, the fourth coordinates are all zero if, as here, the ToF camera is at the origin. Next, if the corresponding edges $\bar{\boldsymbol{u}}_{\ell i}$ and $\bar{\boldsymbol{u}}_{ri}$ can be detected in the binocular system, using both color and parallax information, then the planes $\overline{\boldsymbol{U}}_i$ can easily be constructed. Each calibration plane now contributes an additional $6L$ equations

$$\left(\overline{\boldsymbol{V}}_i\right)_\wedge \boldsymbol{H}^{-\top} \overline{\boldsymbol{U}}_i = \boldsymbol{0}_6 \tag{4.17}$$

to the DLT system (4.12). Although these equations are quite redundant (any two planes span all possibilities), they lead to a much better DLT estimate. This is because they represent exactly those planes that are most likely to be missed in the calibration data, owing to the difficulty of feature detection over surfaces that are extremely foreshortened in the image.

As in the point-based method, the plane coordinates should be suitably transformed, in order to make the numerical system (4.12) well conditioned. The transformed coordinates satisfy the location constraint $\sum_k \boldsymbol{U}_{k\triangle} = \boldsymbol{0}_3$, as well as the scale constraint $\sum_k |\boldsymbol{U}_{k\triangle}|^2 = 3 \sum_k U_{k4}^2$, where $\boldsymbol{U}_{k\triangle} = (U_{k1}, U_{k2}, U_{k3})^\top$, as usual. A final renormalization $|\boldsymbol{U}_k| = 1$ is also performed. This procedure, which is also applied to the $\boldsymbol{V}_k$, is analogous to the treatment of line coordinates in DLT methods [26].

The remaining problem is that the original reprojection error (4.13) cannot be used to optimize the solution, because no luminance features $\boldsymbol{q}$ have been detected in the range images (and so no 3-D points $\boldsymbol{Q}$ have been distinguished). This can be solved by reprojecting the physical edges of the calibration planes, after reconstructing them as follows. Each edge plane $\overline{\boldsymbol{V}}_i$ intersects the range plane $\boldsymbol{V}$ in a space-line, represented by the 4×4 Plücker matrix

$$\boldsymbol{W}_i = \boldsymbol{V}\overline{\boldsymbol{V}}_i^\top - \overline{\boldsymbol{V}}_i \boldsymbol{V}^\top. \tag{4.18}$$

The line $\boldsymbol{W}_i$ reprojects to a 3×3 antisymmetric matrix [10]; for example

$$\boldsymbol{W}_{\ell i} \simeq \boldsymbol{C}_\ell \boldsymbol{W}_i \boldsymbol{C}_\ell^\top \tag{4.19}$$

in the left image, and similarly in the right. Note that $\boldsymbol{W}_{\ell i}\boldsymbol{p}_\ell = \boldsymbol{0}$ if the point $\boldsymbol{p}_\ell$ is on the reprojected line [10]. The line-reprojection error can therefore be written as

$$E_\ell^\times(\boldsymbol{C}_\ell) = \sum_{i=1}^{L}\sum_{j=1}^{N} D_\times\big(\boldsymbol{C}_\ell \boldsymbol{W}_i \boldsymbol{C}_\ell^\top,\ \bar{\boldsymbol{u}}_{\ell ij}\big)^2. \tag{4.20}$$

The function $D_\times\big(\boldsymbol{M}, \boldsymbol{n}\big)$ compares image lines, by computing the sine of the angle between the two coordinate vectors,

$$D_\times(\boldsymbol{M}, \boldsymbol{n}) = \frac{\sqrt{2}\,|\boldsymbol{M}\boldsymbol{n}|}{|\boldsymbol{M}|\,|\boldsymbol{n}|} = \frac{|\boldsymbol{m}\times\boldsymbol{n}|}{|\boldsymbol{m}|\,|\boldsymbol{n}|}, \tag{4.21}$$

where $\boldsymbol{M} = (\boldsymbol{m})_\times$, and $|\boldsymbol{M}|$ is the Frobenius norm. It is emphasized that the coordinates *must* be normalized by a suitable transformations $\boldsymbol{G}_\ell$ and $\boldsymbol{G}_r$, as in the case of the DLT. For example, the line $\boldsymbol{n}$ should be fitted to points of the form $\boldsymbol{G}\boldsymbol{p}$, and then $\boldsymbol{M}$ should be transformed as $\boldsymbol{G}^{-\top}\boldsymbol{M}$, before computing (4.21). The reprojection error (4.20) is numerically unreliable without this normalization.

The line reprojection (4.21) can either be minimized separately for each camera, or jointly as

$$E^\times\big(\boldsymbol{H}^{-1}\big) = E_\ell^\times\big(\boldsymbol{C}_\ell\boldsymbol{H}^{-1}\big) + E_r^\times\big(\boldsymbol{C}_r\boldsymbol{H}^{-1}\big) \tag{4.22}$$

by analogy with (4.14). Finally, it should be noted that although (4.21) is defined in the *image*, it is an *algebraic* error. However, because the errors in question are small, this measure behaves predictably (see Fig. 4.2).

4.2.5 Multisystem Alignment

The point-based and plane-based procedures, described in Sects. 4.2.3 and 4.2.4 respectively, can be used to calibrate a single ToF+2RGB system. Related methods can be used for the joint calibration of several such systems, as will now be explained, using the *point-based* representation. 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 have the form

$$\boldsymbol{C}_i \simeq (\boldsymbol{A}_i \,|\, \boldsymbol{0}_3), \quad \boldsymbol{C}_{\ell i} \simeq (\boldsymbol{A}_{\ell i} \,|\, \boldsymbol{0}_3) \quad \text{and} \quad \boldsymbol{C}_{ri} \simeq (\boldsymbol{A}_{ri} \,|\, \boldsymbol{b}_{ri}) \tag{4.23}$$

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. This matrix, in the calibrated case, would be a rigid 3-D transformation. However, by analogy with the ToF-to-RGB matrices, each $\boldsymbol{G}_{ij}$ is generalized here to a projective transformation, thereby allowing for spatial distortions in the data. The left and right cameras that project a scene point $\boldsymbol{P}_j$ in coordinate system j

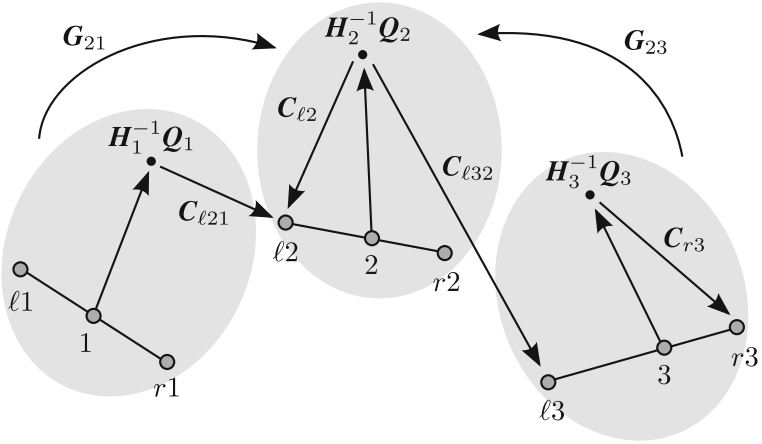

Fig. 4.3 Example of a three ToF+2RGB setup, with ToF cameras labeled 1,2,3. Each *ellipse* represents a separate system, with system 2 chosen as the reference. The *arrows* (with camera-labels) show some possible ToF-to-RGB projections. For example, a point $\boldsymbol{P}_2 \simeq \boldsymbol{H}_2^{-1}\boldsymbol{Q}_2$ in the center projects directly to RGB view $\ell 2$ via $\boldsymbol{C}_{\ell 2}$, whereas the same point projects to $\ell 3$ via $\boldsymbol{C}_{\ell 32} = \boldsymbol{C}_{\ell 3}\boldsymbol{G}_{32}$

to image points $\boldsymbol{p}_{\ell i}$ and $\boldsymbol{p}_{ri}$ in system i are

$$\boldsymbol{C}_{\ell ij} = \boldsymbol{C}_{\ell i}\ \boldsymbol{G}_{ij} \quad \text{and} \quad \boldsymbol{C}_{rij} = \boldsymbol{C}_{ri}\ \boldsymbol{G}_{ij}. \tag{4.24}$$

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 cameras are respectively equal to $\boldsymbol{C}_{\ell i}$ and $\boldsymbol{C}_{ri}$ in (4.23) only when $i = k$, such that $\boldsymbol{G}_{ij} = \boldsymbol{I}$ in (4.24). A typical three-system configuration is shown in Fig. 4.3.

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, the stereo-reconstructed points $\boldsymbol{P}$ are used to estimate these transformations, as they are more reliable than the ToF points $\boldsymbol{Q}$.

4.3 Evaluation

The following sections will describe the accuracy of a nine-camera setup, calibrated by the methods described above. Section 4.3.1 will evaluate *calibration* error, whereas Sect. 4.3.2 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. The importance of this distinction will be discussed.

The setup consists of three rail-mounted ToF+2RGB systems, $i = 1 \ldots 3$, as in Fig. 4.3. The stereo baselines are 17 cm on average, and the ToF cameras are separated by 107 cm on average. The RGB images are 1624×1224, whereas the Mesa Imaging SR4000 ToF images are 176×144, with a depth range of 500 cm. The three stereo systems are first calibrated by standard methods, returning a full Euclidean decomposition of $\boldsymbol{C}_{\ell i}$ and $\boldsymbol{C}_{ri}$, as well as the associated lens parameters. It was established in [8] that projective alignment is generally superior to similarity alignment, and so the transformations $\boldsymbol{G}_{ij}$ and $\boldsymbol{H}_j^{-1}$ will be 4×4 homographies. These transformations were estimated by the DLT method, and refined by LM-minimization of the joint geometric error, as in (4.14).

4.3.1 Calibration Error

The calibration error is measured by first taking ToF points $\boldsymbol{Q}_j^{\pi}$ corresponding to *vertices* on the reconstructed calibration plane π_j in system j, as described in Sect. 4.2.2. These can then be projected into a pair of RGB images in system i, so that the error $E_{ij}^{\text{cal}} = \frac{1}{2}\left(E_{\ell ij}^{\text{cal}} + E_{rij}^{\text{cal}}\right)$ can be computed, where

$$E_{\ell ij}^{\text{cal}} = \frac{1}{|\pi|} \sum_{\boldsymbol{Q}_j^{\pi}} D\left(\boldsymbol{C}_{\ell ij}\, \boldsymbol{H}_j^{-1}\, \boldsymbol{Q}_j^{\pi},\ \ \boldsymbol{p}_{\ell i}\right) \tag{4.25}$$

and E_{rij}^{cal} is similarly defined. The function $D(\cdot, \cdot)$ computes the image distance between inhomogenized points, as in (4.13), and the denominator corresponds to the number of vertices on the board, with $|\pi| = 35$ in the present experiments. The measure (4.25) can of course be averaged over all images in which the board is visible. The calibration procedure has an accuracy of around 1 pixel, as shown in Fig. 4.4.

4.3.2 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 for 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 must be related to the ToF image points $\boldsymbol{q}_j$ by the 2-D *transfer* homographies $\boldsymbol{T}_{\ell ij}$ and $\boldsymbol{T}_{rij}$, where

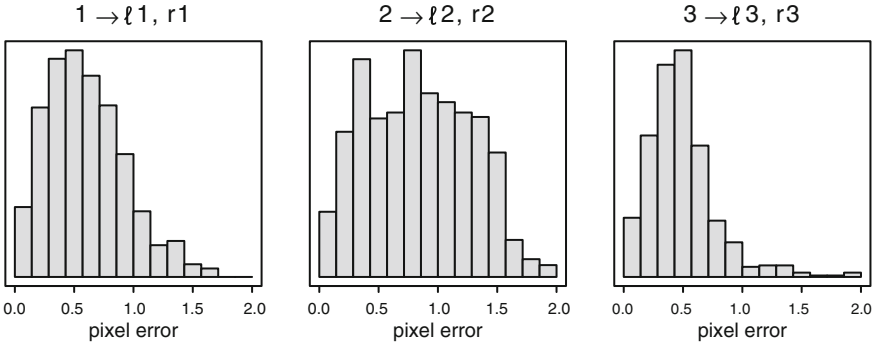

Fig. 4.4 Calibration error (4.25), measured by projecting the fitted ToF points $\boldsymbol{Q}^{\pi}$ to the left and right RGB images (1624×1224) in three separate systems. Each histogram combines *left*-camera and *right*-camera measurements from 15 views of the calibration board. Subpixel accuracy is obtained

$$\boldsymbol{p}_{\ell i} \simeq \boldsymbol{T}_{\ell ij}\, \boldsymbol{q}_j \quad \text{and} \quad \boldsymbol{p}_{ri} \simeq \boldsymbol{T}_{rij}\, \boldsymbol{q}_j. \tag{4.26}$$

These 3×3 matrices can be estimated accurately, because the range data itself is not required. Furthermore, let Π_j be the 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 *reprojected* 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 (4.26). The total error is the average difference between the reprojections and the transfers, $E^{\text{tot}}_{ij} = \frac{1}{2}\big(E^{\text{tot}}_{\ell ij} + E^{\text{tot}}_{rij}\big)$, where

$$E^{\text{tot}}_{\ell ij} = \frac{1}{|\Pi_j|} \sum_{\boldsymbol{q}_j \in \Pi_j} D\Big(\boldsymbol{C}_{\ell ij}\, \boldsymbol{H}_j^{-1}\, \boldsymbol{Q}_j,\ \ \boldsymbol{T}_{\ell ij}\, \boldsymbol{q}_j\Big) \tag{4.27}$$

and E^{tot}_{rij} is similarly defined. The view-dependent denominator $|\Pi_j| \gg |\pi|$ is the number of pixels in the hull Π_j. Hence, E^{tot}_{ij} is the total error, including 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 canceled in the reprojection. This is evident in Fig. 4.5, although it is clear that the tail of each distribution is increased by the range error. However, if the RGB cameras belong to another system, with a substantially different location, then the range errors can be very large in the reprojection. This is clear from Fig. 4.6, which shows that a substantial proportion of the ToF points reproject to the other systems with a total error in excess of 10 pixels.

It is possible to understand these results more fully by examining the distribution of the total error across individual boards. Figure 4.7 shows the distribution for a board reprojected to the same system (i.e., part of the data from Fig. 4.5). There is a relatively smooth gradient 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. The pixels

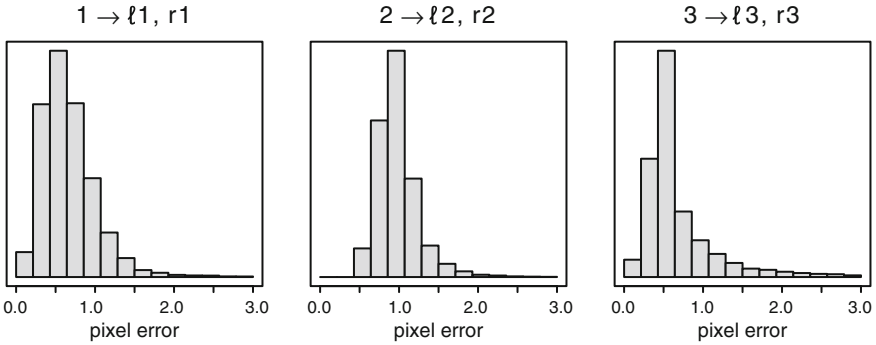

Fig. 4.5 Total error (4.27), measured by projecting the raw ToF points $\boldsymbol{Q}$ to the left and right RGB images (1624 × 1224) in three separate systems. These distributions have longer and heavier tails than those of the corresponding calibration errors, shown in Fig. 4.4

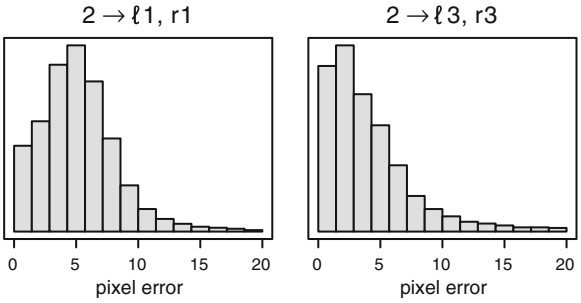

Fig. 4.6 Total error when reprojecting raw ToF points from system 2 to RGB cameras in systems 1 and 3 (*left and right*, respectively). The range errors are emphasized by the difference in viewpoints between the two systems. Average error is now around 5 pixels in the 1624 × 1224 images, and the noisiest ToF points reproject with tens of pixels of error

can be divided into sets from the black and white squares, using the known board geometry and detected vertices. It can be seen in Fig. 4.7 (right) that the total error for each set is comparable. However, when reprojecting to a different system, Fig. 4.8 shows that the total error is correlated with the black and white squares on the board. This is due to significant absorption of the infrared signal by the black squares.

4.4 Conclusions

It has been shown that there is a projective relationship between the data provided by a ToF camera, and an uncalibrated binocular reconstruction. Two practical methods for computing the projective transformation have been introduced; one that requires luminance point correspondences between the ToF and color cameras, and one that

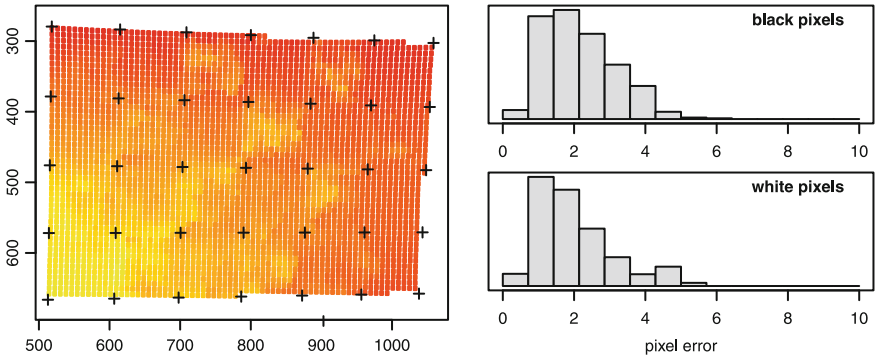

Fig. 4.7 *Left* 3-D ToF pixels ($|\Pi| = 3216$), on a calibration board, reprojected to an RGB image in the same ToF+2RGB system. Each pixel is color coded by the total error (4.27). *Black crosses* are the detected vertices in the RGB image. *Right* histograms of total error, split into pixels on *black* or *white squares*

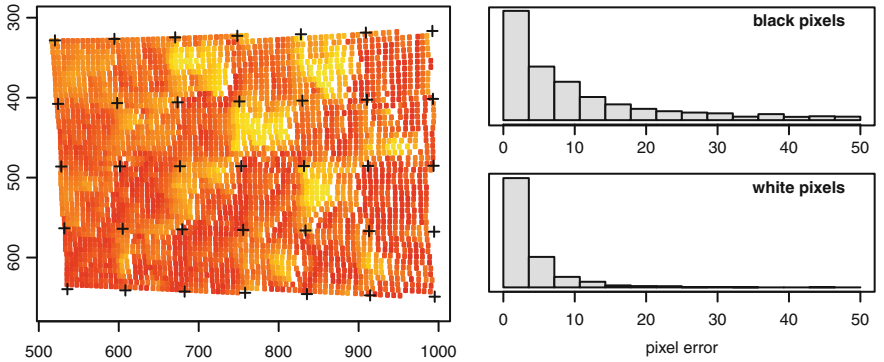

Fig. 4.8 *Left* 3-D ToF pixels, as in Fig. 4.7, reprojected to an RGB image in a different ToF+2RGB system. *Right* histograms of total error, split into pixels on *black or white squares*. The depth of the *black squares* is much less reliable, which leads to inaccurate reprojection into the target system

does not. Either of these methods can be used to associate binocular color and texture with each 3-D point in the range reconstruction. It has been shown that the point-based method can easily be extended to multiple-ToF systems, with calibrated or uncalibrated RGB cameras.

The problem of ToF noise, especially when reprojecting 3-D points to a very different viewpoint, has been emphasized. This source of error can be reduced by application of the denoising methods described in Chap. 1. Alternatively, having aligned the ToF and RGB systems, it is possible to refine the 3-D representation by image matching, as explained in Chap. 5.

References

1. Bartczak, B., Koch, R.: Dense depth maps from low resolution time-of-flight depth and high resolution color views. In: Proceedings of International Symposium on Visual Computing (ISVC), pp. 228–239 (2009)
2. Beder, C., Bartczak, B., Koch, R.: A comparison of PMD-cameras and stereo-vision for the task of surface reconstruction using patchlets. In: Proceedings of Computer Vision and Parallel Recognition (CVPR), pp. 1–8 (2007)
3. Beder, C., Schiller, I., Koch, R.: Photoconsistent relative pose estimation between a PMD 2D3D-camera and multiple intensity cameras. In: Proceedings of Symposium of the German Association for Pattern Recognition (DAGM), pp. 264–273 (2008)
4. Bleiweiss, A., Werman, M.: Fusing time-of-flight depth and color for real-time segmentation and tracking. In: Proceedings of the Dynamic 3D Imaging: DAGM 2009 Workshop, pp. 58–69 (2009)
5. Csurka, G., Demirdjian, D., Horaud, R.: Finding the collineation between two projective reconstructions. Comput. Vis. Image Underst. **75**(3), 260–268 (1999)
6. Dubois, J.M., Hügli, H.: Fusion of time-of-flight camera point clouds. In: Proceedings of European Conference on Computer Vision (ECCV) Workshop on Multi-Camera and Multimodal Sensor Fusion Algorithms and Applications, Marseille (2008)
7. Förstner, W.: Uncertainty and projective geometry. In: Bayro-Corrochano, E. (ed.) Handbook of Geometric Computing, pp. 493–534. Springer, New York (2005)
8. Hansard, M., Horaud, R., Amat, M., Lee, S.: Projective alignment of range and parallax data. In: Proceedings of Computer Vision and Parallel Recognition (CVPR), pp. 3089–3096 (2011)
9. Hartley, R., Sturm, P.: Triangulation. Comput. Vis. Image Underst. **68**(2), 146–157 (1997)
10. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2000)
11. Hebert, M., Krotkov, E.: 3D measurements from imaging laser radars: how good are they? Image Vis. Comput. **10**(3), 170–178 (1992)
12. Horn, B., Hilden, H., Negahdaripour, S.: Closed-form solution of absolute orientation using orthonormal matrices. J. Opt. Soc. Am. A **5**(7), 1127–1135 (1988)
13. Huhle, B., Fleck, S., Schilling, A.: Integrating 3D time-of-flight camera data and high resolution images for 3DTV applications. In: Proceedings of 3DTV Conference, pp. 1–4 (2007)
14. Kanazawa, Y., Kanatani, K.: Reliability of plane fitting by range sensing. In: International Conference on Robotics and Automation (ICRA), pp. 2037–2042 (1995)
15. Kim, Y., Chan, D., Theobalt, C., Thrun, S.: Design and calibration of a multi-view TOF sensor fusion system. In: Proceedings of Computer Vision and Parallel Recognition (CVPR) Workshop on Time-of-Flight Camera based Computer Vision (2008)
16. Koch, R., Schiller, I., Bartczak, B., Kellner, F., Köser, K.: MixIn3D: 3D mixed reality with ToF-camera. In: Proceedings of DAGM Workshop on Dynamic 3D Imaging, pp. 126–141 (2009)
17. Kolb, A., Barth, E., Koch, R., Larsen, R.: Time-of-flight cameras in computer graphics. Comput. Graphics Forum **29**(1), 141–159 (2010)
18. Lindner, M., Schiller, I., Kolb, A., Koch, R.: Time-of-flight sensor calibration for accurate range sensing. Comput. Vis. Image Underst. **114**(12), 1318–1328 (2010)
19. Mesa Imaging AG. http://www.mesa-imaging.ch
20. Pathak, K., Vaskevicius, N., Birk, A.: Revisiting uncertainty analysis for optimum planes extracted from 3D range sensor point-clouds. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), pp. 1631–1636 (2009)
21. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C. Cambridge University Press, 2nd edition (1992)
22. Schiller, I., Beder, C., Koch, R.: Calibration of a PMD camera using a planar calibration object together with a multi-camera setup. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences XXI, pp. 297–302 (2008)

23. Verri, A., Torre, V.: Absolute depth estimate in stereopsis. J. Opt. Soc. Am. A **3**(3), 297–299 (1986)
24. Wang, C., Tanahasi, H., Hirayu, H., Niwa, Y., Yamamoto, K.: Comparison of local plane fitting methods for range data. In: Proceedings of Computer Vision and Parallel Recognition (CVPR), pp. 663–669 (2001)
25. Wu, J., Zhou, Y., Yu, H., Zhang, Z.: Improved 3D depth image estimation algorithm for visual camera. In: Proceedings of International Congress on Image and Signal Processing (2009)
26. Zeng, H., Deng, X., Hu, Z.: A new normalized method on line-based homography estimation. Pattern Recogn. Lett. **29**, 1236–1244 (2008)
27. Zhang Q., Pless, R.: Extrinsic calibration of a camera and laser range finder (improves camera calibration). In: Proceedings of Intenational Conference on Intelligent Robots and Systems, pp. 2301–2306 (2004)
28. Zhu, J., Wang, L., Yang, R.G., Davis, J.: Fusion of time-of-flight depth and stereo for high accuracy depth maps. In: Proceedings of Computer Vision and Parallel Recognition (CVPR), pp. 1–8 (2008)