Chapter 5
A Mixed Time-of-Flight and Stereoscopic Camera System

Abstract Several methods that combine range and color data have been investigated and successfully used in various applications. Most of these systems suffer from the problems of noise in the range data and resolution mismatch between the range sensor and the color cameras. High-resolution depth maps can be obtained using stereo matching, but this often fails to construct accurate depth maps of weakly/repetitively textured scenes. Range sensors provide coarse depth information regardless of presence/absence of texture. We propose a novel ToF-stereo fusion method based on an efficient seed-growing algorithm which uses the ToF data projected onto the stereo image pair as an initial set of correspondences. These initial "seeds" are then propagated to nearby pixels using a matching score that combines an image similarity criterion with rough depth priors computed from the low-resolution range data. The overall result is a dense and accurate depth map at the resolution of the color cameras at hand. We show that the proposed algorithm outperforms 2D image-based stereo algorithms and that the results are of higher resolution than off-the-shelf RGB-D sensors, e.g., Kinect.

Keywords Mixed-camera systems · Stereo seed-growing · Time-of-Flight sensor fusion · Depth and color combination

5.1 Introduction

Advanced computer vision applications require both depth and color information. Hence, a system composed of ToF and color cameras should be able to provide accurate *color and depth* information for each pixel and at high resolution. Such a *mixed* system can be very useful for a large variety of vision problems, e.g., for building dense 3D maps of indoor environments.

The 3D structure of a scene can be reconstructed from two or more 2D views via a *parallax* between corresponding image points. However, it is difficult to obtain accurate pixel-to-pixel matches for scenes of objects without textured surfaces, with repetitive patterns, or in the presence of occlusions. The main drawback is that

M. Hansard et al., *Time-of-Flight Cameras*, SpringerBriefs in Computer Science,
DOI: 10.1007/978-1-4471-4658-2_5, © Miles Hansard 2013

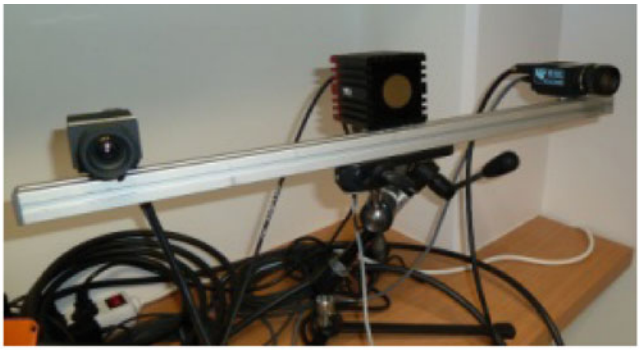
stereo matching algorithms frequently fail to reconstruct indoor scenes composed of untextured surfaces, e.g., walls, repetitive patterns and surface discontinuities, which are typical in man-made environments.

Alternatively, *active-light* range sensors, such as time-of-flight (ToF) or structured-light cameras (see Chap. 1), can be used to directly measure the 3D structure of a scene at video frame rates. However, the spatial resolution of currently available range sensors is lower than high-definition (HD) color cameras, the luminance sensitivity is poorer and the depth range is limited. The range-sensor data are often noisy and incomplete over extremely scattering parts of the scene, e.g., non-Lambertian surfaces. Therefore, it is not judicious to rely solely on range-sensor estimates for obtaining 3D maps of complete scenes. Nevertheless, range cameras provide good initial estimates independently of whether the scene is textured or not, which is not the case with stereo matching algorithms. These considerations show that it is useful to combine the active-range and the passive-parallax approaches, in a *mixed* system. Such a system can overcome the limitations of both the active- and passive-range (stereo) approaches, when considered separately, and provides accurate and fast 3D reconstruction of a scene at high resolution, e.g., 1200×1600 pixels, as in Fig. 5.1.

5.1.1 Related Work

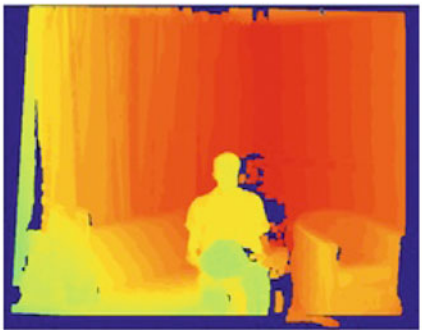
The combination of a depth sensor with a color camera has been exploited in several applications such as object recognition [2, 15, 24], person awareness, gesture recognition [11], simultaneous localization and mapping (SLAM) [3, 17], robotized plant-growth measurement [1], etc. These methods mainly deal with the problem of noise in depth measurement, as examined in Chap. 1, as well as with the low resolution of range data as compared to the color data. Also, most of these methods are limited to RGB-D, i.e., a *single* color image combined with a range sensor. Interestingly enough, the recently commercialized Kinect [13] camera falls in the RGB-D family of sensors. We believe that extending the RGB-D sensor model to RGB-D-RGB sensors is extremely promising and advantageous because, unlike the former type of sensor, the latter type can combine active depth measurement with stereoscopic matching and hence better deal with the problems mentioned above.

Stereo matching has been one of the most studied paradigms in computer vision. There are several papers, e.g., [22, 23] that overview existing techniques and that highlight recent progress in stereo matching and stereo reconstruction. While a detailed description of existing techniques is beyond the scope of this section, we note that algorithms based on greedy local search techniques are typically fast but frequently fail to reconstruct the poorly textured regions or ambiguous surfaces. Alternatively, global methods formulate the matching task as an optimization problem which leads the minimization of a Markov random field (MRF) energy function of the image similarity likelihood and a prior on the surface smoothness.



(a)

(b)



(c)

Fig. 5.1 **a** Two high-resolution color cameras (2.0 MP at 30 FPS) are combined with a single low-resolution ToF camera (0.03 MP at 30 FPS). **b** The 144×177 ToF image (*upper left corner*) and two 1224×1624 *color* images are shown at the true scale. **c** The depth map obtained with our method. The technology used by both these camera types allows simultaneous range and photometric data acquisition with an extremely accurate temporal synchronization, which may not be the case with other types of range cameras such as the current version of Kinect

These algorithms solve some of the aforementioned problems of local methods but are very complex and computationally expensive since optimizing an MRF-based energy function is an NP-hard problem in the general case.

A practical tradeoff between the local and the global methods in stereo is the seed-growing class of algorithms [4–6]. The correspondences are grown from a small set of initial correspondence seeds. Interestingly, they are not particularly sensitive to bad input seeds. They are significantly faster than the global approaches, but they have difficulties in presence of nontextured surfaces; Moreover, in these cases they yield depth maps which are relatively sparse. Denser maps can be obtained by relaxing the matching threshold but this leads to erroneous growth, so there is a natural tradeoff between the accuracy and density of the solution. Some form of regularization is necessary in order to take full advantage of these methods.

Recently, external prior-based generative probabilistic models for stereo matching were proposed [14, 20] for reducing the matching ambiguities. The priors used were based on surface triangulation obtained from an initially matched distinctive interest points in the two color images. Again, in the absence of textured regions, such support points are only sparsely available, and are not reliable enough or are not available at all in some image regions, hence the priors are erroneous. Consequently, such prior-based methods produce artifacts where the priors win over the data, and the solution is biased toward such incorrect priors. This clearly shows the need for more accurate prior models. Wang et al. [25] integrate a regularization term based on the depth values of initially matched *ground control points* in a global energy minimization framework. The ground control points are gathered using an accurate laser scanner. The use of a laser scanner is tedious because it is difficult to operate and because it cannot provide depth measurements fast enough such that it can be used in a practical computer vision application.

ToF cameras are based on an active sensor principle[1] that allows 3D data acquisition at video frame rates, e.g., 30 FPS as well as accurate synchronization with any number of color cameras.[2] A modulated infrared light is emitted from the camera's internal lighting source, is reflected by objects in the scene and eventually travels back to the sensor, where the time of flight between sensor and object is measured independently at each of the sensor's pixel by calculating the precise phase delay between the emitted and the detected waves. A complete depth map of the scene can thus be obtained using this sensor at the cost of very low spatial resolution and coarse depth accuracy (see Chap. 1 for details).

The fusion of ToF data with stereo data has been recently studied. For example, [8] obtained a higher quality depth map, by a probabilistic ad hoc fusion of ToF and stereo data. Work in [26] merges the depth probability distribution function obtained from ToF and stereo. However, both these methods are meant for improvement over the initial data gathered with the ToF camera and the final depth-map result is still limited to the resolution of the ToF sensor. The method proposed in this chapter

[1] All experiments described in this chapter use the Mesa SR4000 camera [18].

[2] http://www.4dviews.com

increases the resolution from 0.03 MP to the full resolution of the color cameras being used, e.g., 2 MP.

The problem of depth-map upsampling has been also addressed in the recent past. In [7] a noise-aware filter for adaptive multilateral upsampling of ToF depth maps is presented. The work described in [15, 21] extends the model of [9], and [15] demonstrates that the object detection accuracy can be significantly improved by combining a state-of-the-art 2D object detector with 3D depth cues. The approach deals with the problem of resolution mismatch between range and color data using an MRF-based superresolution technique in order to infer the depth at every pixel. The proposed method is slow: It takes around 10 s to produce a 320×240 depth image. All of these methods are limited to depth-map upsampling using only a single color image and do not exploit the added advantage offered by stereo matching, which can highly enhance the depth map both qualitatively and quantitatively. Recently, [12] proposed a method which combines ToF estimates with stereo in a semiglobal matching framework. However, at pixels where ToF disparity estimates are available, the image similarity term is ignored. This make the method quite susceptible to errors in regions where ToF estimates are not precise, especially in textured regions where stereo itself is reliable.

5.1.2 Chapter Contributions

In this chapter, we propose a novel method for incorporating range data within a robust seed-growing algorithm for stereoscopic matching [4]. A calibrated system composed of an active-range sensor and a stereoscopic color camera pair, as described in Chap. 4 and [16], allows the range data to be aligned and then projected onto each one of the two images, thus providing an initial sparse set of point-to-point correspondences (seeds) between the two images. This initial seed set is used in conjunction with the seed-growing algorithm proposed in [4]. The projected ToF points are used as the vertices of a mesh-based surface representation which, in turn, is used as a prior to regularize the image-based matching procedure. The novel probabilistic *fusion* model proposed here (between the mesh-based surface initialized from the sparse ToF data and the seed-growing stereo matching algorithm itself) combines the merits of the two 3D sensing methods (active and passive) and overcomes some of the limitations outlined above. Notice that the proposed fusion model can be incorporated within virtually any stereo algorithm that is based on energy minimization and which requires some form initialization. It is, however, particularly efficient and accurate when used in combination with match-propagation methods.

The remainder of this chapter is structured as follows: Sect. 5.2 describes the proposed range-stereo fusion algorithm. The growing algorithm is summarized in Sect. 5.2.1. The processing of the ToF correspondence seeds is explained in Sect. 5.2.2, and the sensor fusion based similarity statistic is described in Sect. 5.2.3. Experimental results on a real data set and evaluation of the method, are presented in Sect. 5.3. Finally, Sect. 5.4 draws some conclusions.

5.2 The Proposed ToF-Stereo Algorithm

As outlined above, the ToF camera provides a low-resolution depth map of a scene. This map can be projected onto the left and right images associated with the stereoscopic pair, using the projection matrices estimated by the calibration method described in Chap. 4. Projecting a single 3D point (x, y, z) gathered by the ToF camera onto the *rectified* images provides us with a pair of corresponding points (u, v) and (u', v') with $v' = v$ in the respective images. Each element (u, u', v) denotes a point in the disparity space.[3] Hence, projecting all the points obtained with the ToF camera gives us a sparse set of 2D point correspondences. This set is termed as the set of initial support points or ToF *seeds*.

These initial support points are used in a variant of the seed-growing stereo algorithm [4, 6] which further grows them into a denser and higher resolution disparity map. The seed-growing stereo algorithms propagate the correspondences by searching in the small neighborhoods of the seed correspondences. Notice that this growing process limits the disparity space to be visited to only a small fraction, which makes the algorithm extremely efficient from a computational point-of-view. The limited neighborhood also gives a kind of implicit regularization, nevertheless the solution can be arbitrarily complex, since multiple seeds are provided.

The integration of range data within the seed-growing algorithm requires two major modifications: (1) The algorithm is using ToF seeds instead of the seeds obtained by matching distinctive image features, such as interest points, between the two images, and (2) the growing procedure is regularized using a similarity statistic which takes into account the photometric consistency as well as the depth likelihood based on disparity estimate by interpolating the rough triangulated ToF surface. This can be viewed as a prior cast over the disparity space.

5.2.1 The Growing Procedure

The growing algorithm is sketched in pseudocode as Algorithm 1. The input is a pair of rectified images (I_L, I_R), a set of *refined* ToF seeds $\mathscr{S}$ (see below), and a parameter τ which directly controls a tradeoff between matching accuracy and matching density. The output is a disparity map D which relates pixel correspondences between the input images.

First, the algorithm computes the prior disparity map D_p by interpolating ToF seeds. Map D_p is of the same size as the input images and the output disparity map, Step 1. Then, a similarity statistic simil $(s|I_L, I_R, D_p)$ of the correspondence, which measures both the photometric consistency of the potential correspondence as well as its consistency with the prior, is computed for all seeds $s = (u, u', v) \in \mathscr{S}$, Step 2. Recall that the seed s stands for a pixel-to-pixel correspondence $(u, v) \leftrightarrow (u', v)$

[3] The disparity space is a space of all potential correspondences [22].

Algorithm 1 Growing algorithm for ToF-stereo fusion

Require: Rectified images (I_L, I_R),
initial correspondence seeds $\mathscr{S}$,
image similarity threshold τ.

1: Compute the prior disparity map D_p by interpolating seeds $\mathscr{S}$.
2: Compute simil$(s|I_L, I_R, D_p)$ for every seed $s \in \mathscr{S}$.
3: Initialize an empty disparity map D of size I_L (and D_p).
4: **repeat**
5: Draw seed $s \in \mathscr{S}$ of the best simil$(s|I_L, I_R, D_p)$ value.
6: **for** each of the four best neighbors $i \in \{1, 2, 3, 4\}$
$q_i^* = (u, u', v) = \underset{q \in \mathscr{N}_i(s)}{\text{argmax}}\ \text{simil}(q|I_L, I_R, D_p)$
do
7: $c := \text{simil}(q_i^*|I_L, I_R, D_p)$
8: **if** $c \geq \tau$ **and** pixels not matched yet **then**
9: Update the seed queue $\mathscr{S} := \mathscr{S} \cup \{q_i^*\}$.
10: Update the output map $D(u, v) = u - u'$.
11: **end if**
12: **end for**
13: **until** $\mathscr{S}$ is empty
14: **return** disparity map D.

between the left and the right images. For each seed, the algorithm searches other correspondences in the surroundings of the seeds by maximizing the similarity statistic. This is done in a 4-neighborhood $\{\mathscr{N}_1, \mathscr{N}_2, \mathscr{N}_3, \mathscr{N}_4\}$ of the pixel correspondence, such that in each respective direction (left, right, up, down) the algorithm searches the disparity in a range of ± 1 pixel from the disparity of the seed, Step 6. If the similarity statistic of a candidate exceeds the threshold value τ, then a new correspondence is found, Step 8. This new correspondence becomes itself a new seed, and the output disparity map D is updated accordingly. The process repeats until there are no more seeds to be grown.

The algorithm is robust to a fair percentage of wrong initial seeds. Indeed, since the seeds compete to be matched based on a best-first strategy, the wrong seeds typically have low score simil(s) associated with them and therefore when they are evaluated in Step 5, it is likely that the involved pixels been already matched. For more details on the growing algorithm, we refer the reader to [4, 6].

5.2.2 ToF Seeds and Their Refinement

The original version of the seed-growing stereo algorithm [6] uses an initial set of seeds $\mathscr{S}$ obtained by detecting interest points in both images and matching them. Here, we propose to use ToF seeds. As already outlined, these seeds are obtained by projecting the low-resolution depth map associated with the ToF camera onto the high-resolution images. Likewise the case of interest points, this yields a sparse set of seeds, e.g., approximately 25,000 seeds in the case of the ToF camera used in

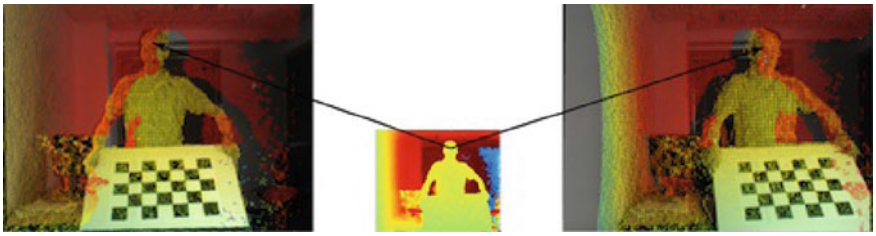


Fig. 5.2 This figure shows an example of the projection of the ToF points onto the left and right images. The projected points are color coded such that the color represents the disparity: cold colors correspond to large disparity values. Notice that there are many wrong correspondences on the computer monitor due to the screen reflectance and to artifacts along the occlusion boundaries

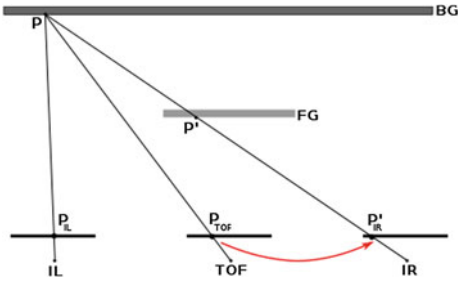


Fig. 5.3 The effect of occlusions. A ToF point P that belongs to a background (BG) objects is only observed in the left image (IL), while it is occluded by a foreground object (FG), and hence not seen in the right image (IR). When the ToF point P is projected onto the *left* and *right* images, an incorrect correspondence ($P_{IL} \leftrightarrow P'_{IR}$) is established

our experiments. Nevertheless, one of the main advantages of the ToF seeds over the interest points is that they are regularly distributed across the images regardless of the presence/absence of texture. This is not the case with interest points whose distribution strongly depends on texture as well as lighting conditions, etc. Regularly distributed seeds will provide a better coverage of the observed scene, i.e., even in the absence of textured areas.

However, ToF seeds are not always reliable. Some of the depth values associated with the ToF sensor are inaccurate. Moreover, whenever a ToF point is projected onto the left and onto the right images, it does not always yield a valid stereo match. There may be several sources of error which make the ToF seeds less reliable than one would have expected, as in Figs. 5.2 and 5.3. In detail:

1. *Imprecision due to the calibration process*. The transformations allowing to project the 3D ToF points onto the 2D images are obtained via a complex sensor calibration process, i.e., Chap. 4. This introduces localization errors in the image planes of up to 2 pixels.

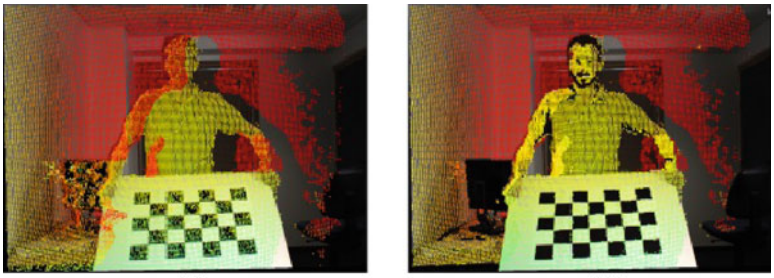


Fig. 5.4 An example of the effect of correcting the set of seeds on the basis that they should be regularly distributed. **a** Original set of seeds. **b** Refined set of seeds

2. *Outliers due to the physical/geometric properties of the scene.* Range sensors are based on active light and on the assumption that the light beams travel from the sensor and back to it. There are a number of situations where the beam is lost, such as specular surfaces, absorbing surfaces (such as fabric), scattering surfaces (such as hair), slanted surfaces, bright surfaces (computer monitors), faraway surfaces (limited range), or when the beam travels in an unpredictable way, such a multiple reflections.
3. *The* ToF *camera and the* 2D *cameras observe the scene from slightly different points of view*. Therefore, it may occur that a 3D point that is present in the ToF data is only seen into the left or right image, as in Fig. 5.3, or is not seen at all.

Therefore, a fair percentage of the ToF seeds are *outliers*. Although the seed-growing stereo matching algorithm is robust to the presence of outliers in the initial set of seeds, as already explained in Sect. 5.2.1, we implemented a straightforward refinement step in order to detect and eliminate incorrect seed data, prior to applying Algorithm 1. First, the seeds that lie in low-intensity (very dark) regions are discarded since the ToF data are not reliable in these cases. Second, in order to handle the background-to-foreground occlusion effect just outlined, we detect seeds which are not uniformly distributed across image regions. Indeed, projected 3D points lying on smooth fronto-parallel surfaces form a regular image pattern of seeds, while projected 3D points that belong to a background surface and which project onto a foreground image region do not form a regular pattern, e.g., occlusion boundaries in Fig. 5.4a.

Nonregular seed patterns are detected by counting the seed occupancy within small 5×5 pixel windows around every seed point in both images. If there is more than one seed point in a window, the seeds are classified as belonging to the background and hence they are discarded. A refined set of seeds is shown in Fig. 5.4b. The refinement procedure typically filters 10–15 % of all seed points.

5.2.3 Similarity Statistic Based on Sensor Fusion

The original seed-growing matching algorithm [6] uses Moravec's normalized cross-correlation [19] (MNCC),

$$\text{simil}(s) = \text{MNCC}(w_L, w_R) = \frac{2\text{cov}(w_L, w_R)}{\text{var}(w_L) + \text{var}(w_R) + \varepsilon} \tag{5.1}$$

as the similarity statistic to measure the photometric consistency of a correspondence $s : (u, v) \leftrightarrow (u', v)$. We denote by w_L and w_R the feature vectors which collect image intensities in small windows of size $n \times n$ pixels centered at (u, v) and $(u'v)$ in the left and right image, respectively. The parameter ε prevents instability of the statistic in cases of low-intensity variance. This is set as the machine floating point epsilon. The statistic has low response in textureless regions and therefore the growing algorithm does not propagate the correspondences across these regions. Since the ToF sensor can provide seeds without the presence of any texture, we propose a novel similarity statistic, $\text{simil}(s|I_L, I_R, D_p)$. This similarity measure uses a different score for photometric consistency as well as an initial high-resolution disparity map D_p, both incorporated into the Bayesian model explained in detail below.

The initial disparity map D_p is computed as follows. A 3D meshed surface is built from a 2D triangulation applied to the ToF image. The disparity map D_p is obtained via interpolation from this surface such that it has the same (high) resolution as of the left and right images. Figure 5.5a, b show the meshed surface projected onto the left high-resolution image and built from the ToF data, before and after the seed refinement step, which makes the D_p map more precise.

Let us now consider the task of finding an optimal high-resolution disparity map. For each correspondence $(u, v) \leftrightarrow (u', v)$ and associated disparity $d = u - u'$ we seek an optimal disparity d^* such that:

$$d^* = \underset{d}{\text{argmax}}\ P(d|I_L, I_R, D_p). \tag{5.2}$$

By applying the Bayes' rule, neglecting constant terms, assuming that the distribution $P(d)$ is uniform in a local neighborhood where it is sought (Step. 6), and considering conditional independence $P(I_l, I_r, D|d) = P(I_L, I_R|d)P(D_p|d)$, we obtain:

$$d^* = \underset{d}{\text{argmax}}\ P(I_L, I_R|d)P(D_p|d), \tag{5.3}$$

where the first term is the color-image likelihood and the second term is the range-sensor likelihood. We define the color-image and range-sensor likelihoods as:

$$\begin{aligned} P(I_L, I_R|d) &\propto \text{EXPSSD}(w_L, w_R) \\ &= \exp\left(-\frac{\sum_{i=1}^{n\times n}(w_L(i) - w_R(i))^2}{\sigma_s^2 \sum_{i=1}^{n\times n}(w_L(i)^2 + w_R(i)^2)} \right), \end{aligned} \tag{5.4}$$

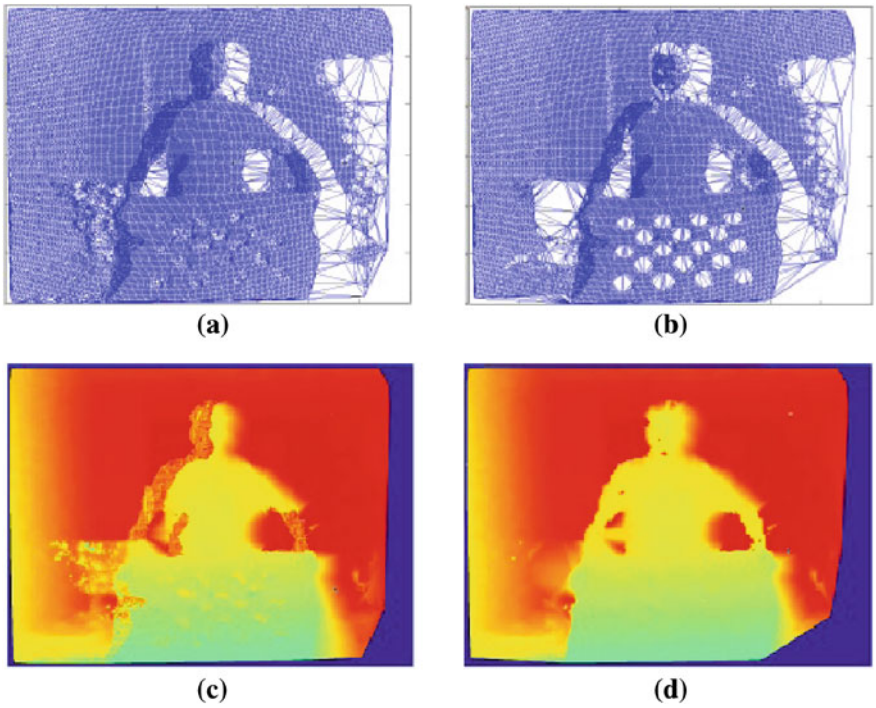


Fig. 5.5 Triangulation and prior disparity map D_p. These are shown using both raw seeds **a**, **c** and refined seeds **b**, **d**. A positive impact of the refinement procedure is clearly visible

and as:

$$P(D_p|d) \propto \exp\left(-\frac{(d-d_p)^2}{2\sigma_p^2}\right) \tag{5.5}$$

respectively, where σ_s are σ_p two normalization parameters. Therefore, the new similarity statistic becomes:

$$\begin{aligned}\text{simil}(s|I_L, I_R, D_p) &= \text{EPC}(w_L, w_R, D_p) \\ &= \exp\left(-\frac{\sum_{i=1}^{n\times n}(w_L(i)-w_R(i))^2}{\sigma_s^2\sum_{i=1}^{n\times n}(w_L(i)^2+w_R(i)^2)} - \frac{(d-d_p)^2}{2\sigma_p^2}\right).\end{aligned} \tag{5.6}$$

Notice that the proposed image likelihood has a high response for correspondences associated with textureless regions. However, in such regions, all possible matches have similar image likelihoods. The proposed range-sensor likelihood regularizes the solution and forces it toward the one closest to the prior disparity map D_p. A tradeoff between these two terms can be obtained by tuning the parameters σ_s and σ_p. We refer to this similarity statistic as the *exponential prior correlation* (EPC) score.

5.3 Experiments

Our experimental setup comprises one Mesa Imaging SR4000 ToF camera [18] and a pair of high-resolution Point Grey[4] color cameras, as shown in Fig. 5.1. The two color cameras are mounted on a rail with a baseline of about 49 cm and the ToF camera is approximately midway between them. All three optical axes are approximately parallel. The resolution of the ToF image is of 144×176 pixels and the color cameras have a resolution of 1224×1624 pixels. Recall that Fig. 5.1b highlights the resolution differences between the ToF and color images. This camera system was calibrated using the alignment method of Chap. 4.

In all our experiments, we set the parameters of the method as follows: Windows of 5×5 pixels were used for matching ($n = 5$), the matching threshold in Algorithm 1 is set to $\tau = 0.5$, the balance between the photometric and range-sensor likelihoods is governed by two parameters in (5.6), which were set to $\sigma_s^2 = 0.1$ and to $\sigma_p^2 = 0.001$.

We show both qualitatively and quantitatively (using data sets with ground truth) the benefits of the range sensor and an impact of particular variants of the proposed fusion model integrated in the growing algorithm. Namely, we compare results of (i) the original stereo algorithm [6] with MNCC correlation and Harris seeds (MNCC-Harris), (ii) the same algorithm with ToF seeds (MNCC-TOF), (iii) the algorithm which uses EXPSSD similarity statistic instead with both Harris (EXPSSD-Harris) and ToF seeds (EXPSSD-TOF), and (iv) the full sensor fusion model of the regularized growth (EPC). Finally, small gaps of unassigned disparity in the disparity maps were filled by a primitive procedure which assigns median disparity in the 5×5 window around the gap (EPC—gaps filled). These small gaps usually occur in slanted surfaces, since Algorithm 1 in Step. 8 enforces one-to-one pixel matching. Nevertheless this way, they can be filled easily, if needed.

5.3.1 Real-Data Experiments

We captured two real-world data sets using the camera setup described above, SET-1 in Fig. 5.6 and SET-2 in Fig. 5.7. Notice that in both of these examples the scene surfaces are weakly textured. Results shown as disparity maps are color coded, such that warmer colors are further away from the cameras and unmatched pixels are dark blue.

In Fig. 5.6d, we can see that the original algorithm [6] has difficulties in weakly textured regions which results in large unmatched regions due to the MNCC statistic (5.1), and it produces several mismatches over repetitive structures on the background curtain, due to erroneous (mismatched) Harris seeds. In Fig. 5.6e, we can see that after replacing the sparse and somehow erratic Harris seeds with uniformly distributed (mostly correct) ToF seeds, the results have significantly been improved. There are no more mismatches on the background, but unmatched regions are still large. In Fig. 5.6f, the EXPSSD statistic (5.4) was used instead of MNCC which

[4] http://www.ptgrey.com/

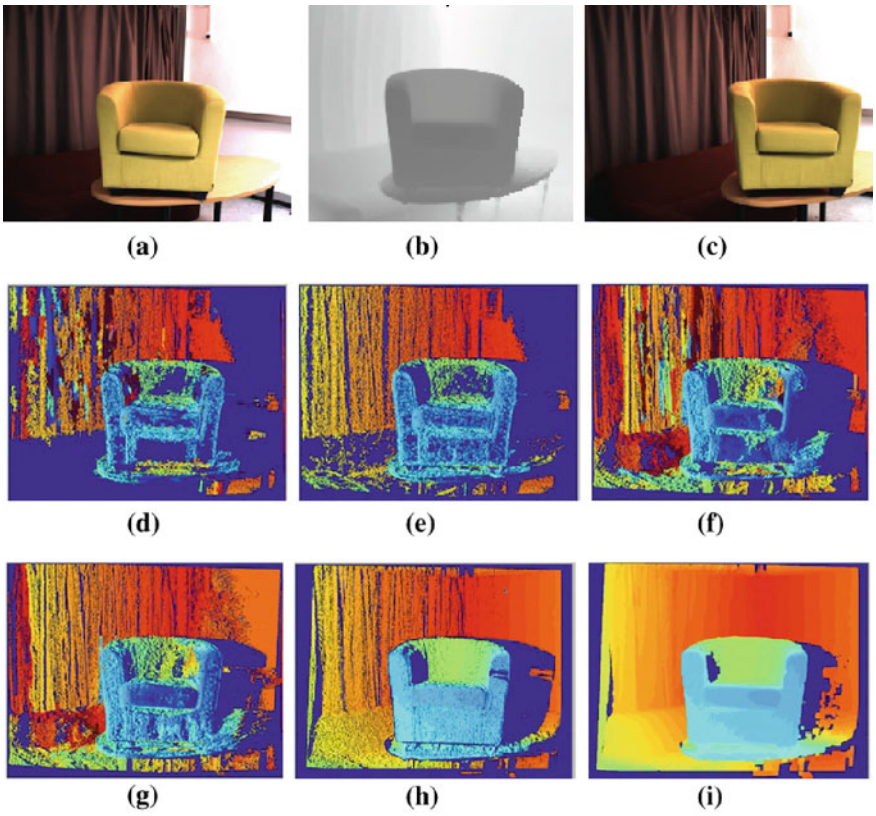


Fig. 5.6 SET-1: **a** left image, **b** ToF image and **c** right image. The ToF image has been zoomed at the resolution of the color images for visualization purposes. Results obtained **d** using the seed-growing stereo algorithm [6] combining Harris seeds and MNCC statistic, **e** using ToF seeds and MNCC statistic , **f** using Harris seeds and EXPSSD statistic, **g** using ToF seeds with EXPSSD statistics. Results obtained with the proposed stereo-ToF fusion model using the EPC (exponential prior correlation) similarity statistic **h**, and EPC after filling small gaps **i**

causes similar mismatches as in Fig. 5.6d, but unlike MNCC there are matches in textureless regions, nevertheless mostly erratic. The reason is that unlike MNCC statistic the EXPSSD statistic has high response in low-textured regions. However, since all disparity candidates have equal (high) response inside such regions, the unregularized growth is random, and produces mismatches. The situation does not improve much using the ToF seeds, as shown in Fig. 5.6g. Significantly better results are finally shown in Fig. 5.6h which uses the proposed EPC fusion model EPC from Eq. (5.6). The EPC statistic, unlike EXPSSD, has the additional regularizing range-sensor likelihood term which guides the growth in ambiguous regions and attracts the solution toward the initial depth estimates of the ToF camera. Results are further refined by filling small gaps, as shown in Fig. 5.6i. Similar observations can be made in Fig. 5.7. The proposed model clearly outperforms the other discussed approaches.

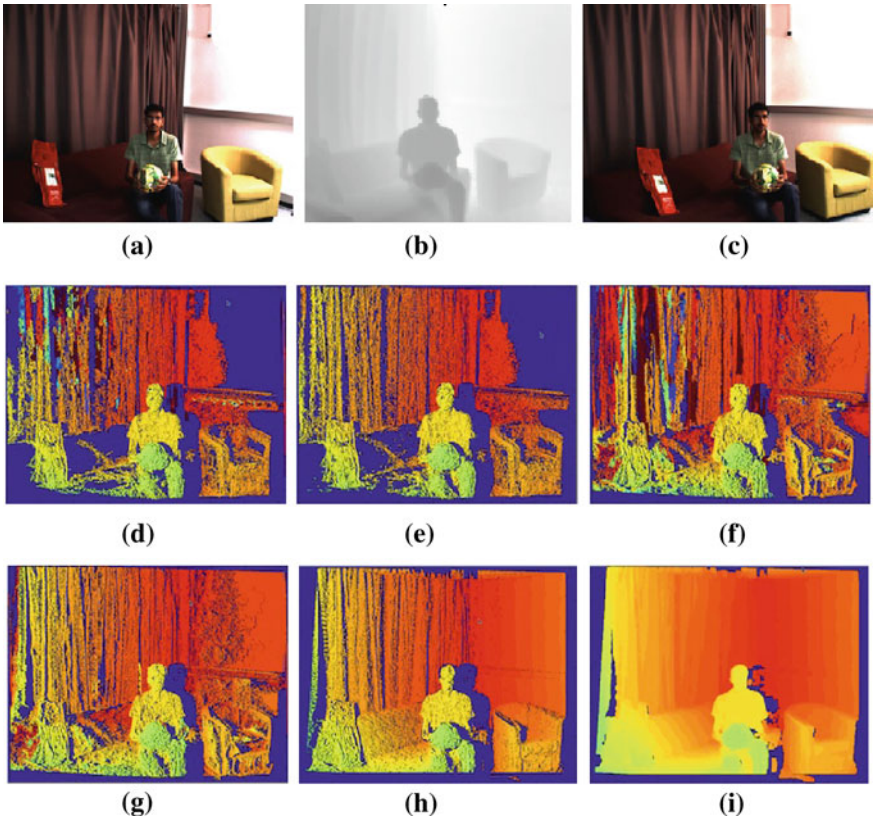


Fig. 5.7 SET-2. Please refer to the caption of Fig. 5.6 for explanations. **a** Left image. **b** ToF image (*zoomed*). **c** Right image. **d** MNCC-Harris. **e** MNCC-TOF. **f** EXPSSD-Harris. **g** EXPSSD-TOF. **h EPC** (proposed). **i EPC** (gaps filled)

5.3.2 Comparison Between ToF Map and Estimated Disparity Map

For the proper analysis of a stereo matching algorithm, it is important to inspect the reconstructed 3D surfaces. Indeed, the visualization of the disparity/depth maps can sometimes be misleading. Surface reconstruction reveals fine details in the quality of the results. This is in order to qualitatively show the gain of the high-resolution depth map produced by the proposed algorithm with respect to the low-resolution depth map of the ToF sensor.

In order to provide a fair comparison, we show the reconstructed surfaces associated with the *dense* disparity maps D_p obtained after 2D triangulation of the ToF data points, Fig. 5.8a, as well as the reconstructed surfaces associated with the disparity map obtained with the proposed method, Fig. 5.8b. Clearly, much more of the sur-

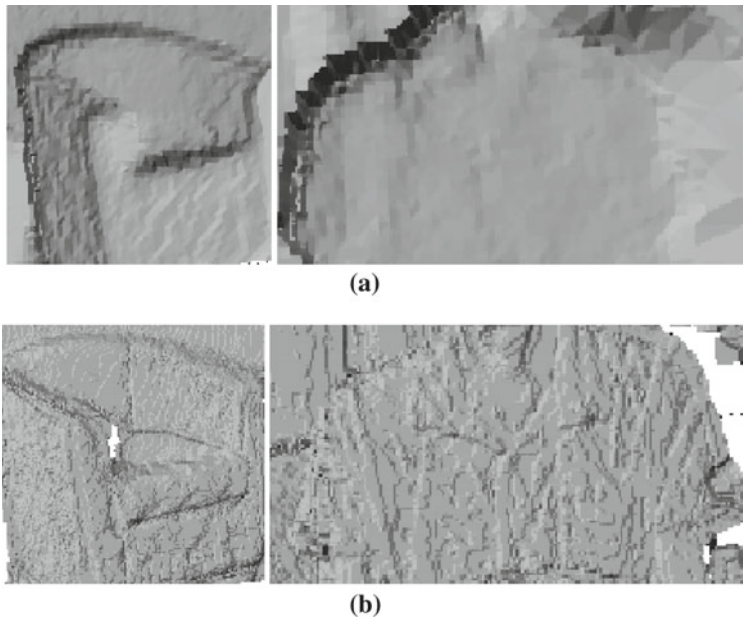


Fig. 5.8 The reconstructed surfaces are shown as relighted 3D meshes for **a** the prior disparity map D_p (2D triangulation on projected and refined ToF seeds), and **b** for the disparity map obtained using the proposed algorithm. Notice the fine surface details which were recovered by the proposed method

face details are recovered by the proposed method. Notice precise object boundaries and fine details, like the cushion on the sofa chair and a collar of the T-shirt, which appear in Fig. 5.8b. This qualitatively corroborates the precision of the proposed method compared to the ToF data.

5.3.3 Ground-Truth Evaluation

To quantitatively demonstrate the validity of the proposed algorithm, we carried out an experiment on data sets with associated ground-truth results. Similarly to [8] we used the Middlebury data set [22] and simulated the ToF camera by sampling the ground-truth disparity map.

The following results are based on the Middlebury-2006 data set.[5] On purpose, we selected three challenging scenes with weakly textured surfaces: Lampshade-1, Monopoly, Plastic. The input images are of size 1330×1110 pixels. We took every 10th pixel in a regular grid to simulate the ToF camera. This gives us about 14k of ToF points, which is roughly the same ratio to color images as for the real sensors. We are aware that simulation ToF sensor this way is naive, since we do not simulate

[5] http://vision.middlebury.edu/stereo/data/scenes2006/

any noise or artifacts, but we believe that for validating the proposed method this is satisfactory.

Results are shown in Fig. 5.9 and Table 5.1. We show the left input image, results of the same algorithms as in the previous section with the real sensor, and the ground-truth disparity map. For each disparity, we compute the percentage of correctly matched pixels in nonoccluded regions. This error statistic is computed as the number of pixels for which the estimated disparity differs from the ground-truth disparity by less than one pixel, divided by number of all pixels in nonoccluded regions. Notice that, unmatched pixels are considered as errors of the same kind as mismatches. This is in order to allow a strict but fair comparison between algorithms which deliver solutions of different densities. The quantitative evaluation confirms the previous observations regarding the real-world setup. The proposed algorithm, which uses the full sensor fusion model, significantly outperforms all other tested variants.

For the sake of completeness, we also report error statistics for the prior disparity map D_p which is computed by interpolating ToF seeds, see step 1 of Algorithm 1. These are 92.9, 92.1, 96.0 % for Lampshade-1, Monopoly, Plastic scene, respectively. These results are already quite good, which means the interpolation we use to construct the prior disparity map is appropriate. These scenes are mostly piecewise planar, which the interpolation captures well. On the other hand, recall that in the real case, not all the seeds are correct due to various artifacts of the range data. Nevertheless in all three scenes, the proposed algorithm (EPC with gaps filled) was able to further improve the precision up to 96.4, 95.3, 98.2 % for the respective scenes. This is again consistent with the experiments with the real ToF sensor, where higher surface details were recovered, see Fig. 5.8.

5.3.4 Computational Costs

The original growing algorithm [6] has low computational complexity due to intrinsic search space reduction. Assuming the input stereo images are of size $n \times n$ pixels, the algorithm has the complexity of $\mathcal{O}(n^2)$, while any exhaustive algorithm has the complexity at least $\mathcal{O}(n^3)$ as noted in [5]. The factor n^3 is the size of the search space in which the correspondences are sought, i.e., the disparity space. The growing algorithm does not compute similarity statistics of all possible correspondences, but efficiently traces out components of high similarity score around the seeds. This low complexity is beneficial especially for high-resolution imagery, which allows precise surface reconstruction.

The proposed algorithm with all presented modifications does not represent any significant extra cost. Triangulation of ToF seeds and the prior disparity map computation is not very costly, and nor is computation of the new EPC statistic (instead of MNCC).

For our experiments, we use an "academic", i.e., a combined Matlab/C implementation which takes approximately 5 s on two million pixel color images. An efficient implementation of the seed-growing algorithm [6] which runs in real time

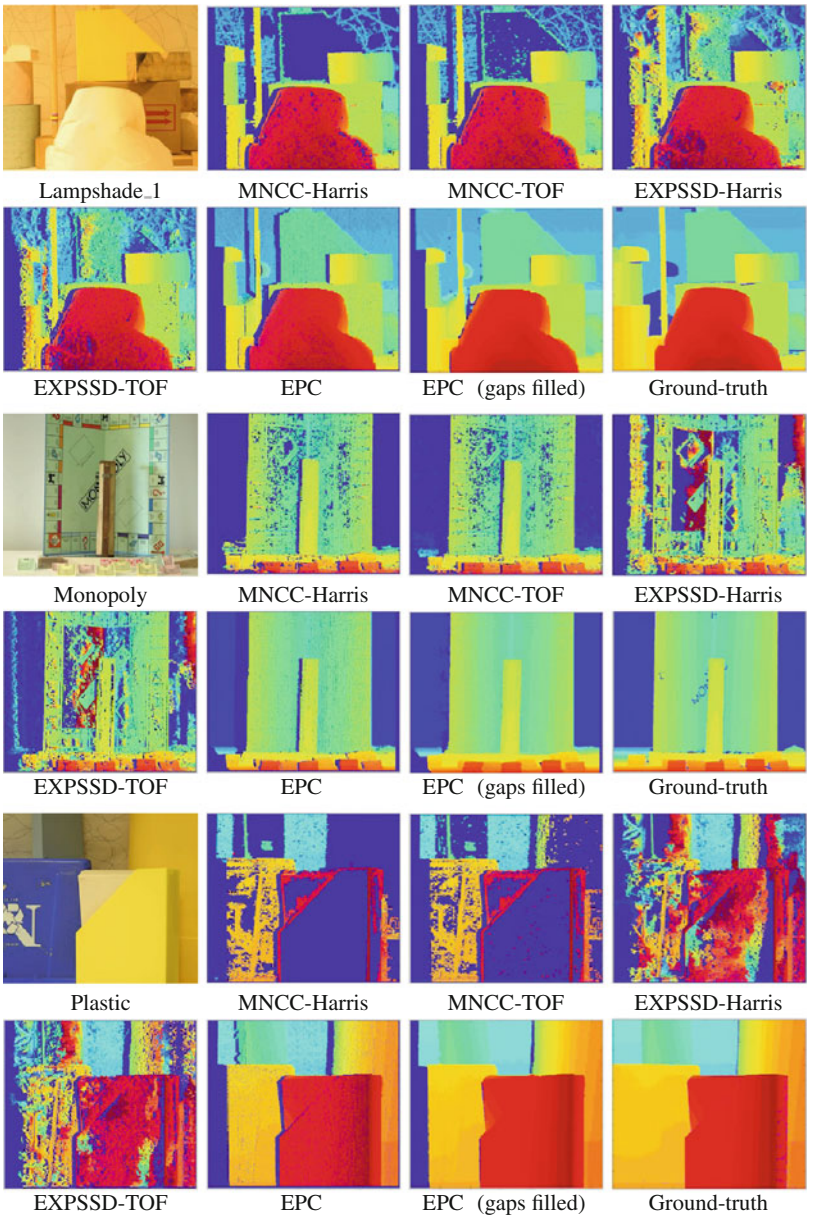


Fig. 5.9 Middlebury data set. *Left-right* and *top-bottom*: the left images, results obtained with the same algorithms as in Figs. 5.6 and 5.7, and the ground-truth disparity maps. This evaluation shows that the combination of the proposed seed-growing stereo algorithm with a prior disparity map, obtained from a sparse and regularly distributed set of 3D points, yields excellent dense matching results

Table 5.1 The error statistics (percentage of correctly matched pixels) associated with the tested algorithms and for three test image pairs from the Middlebury data set

Left image	MNCC-Harris (%)	MNCC-TOF (%)	EXPSSD-Harris (%)	EXPSSD-TOF (%)	**EPC** (%)	**EPC** (gaps filled) (%)
Lampshade-1	61.5	64.3	44.9	49.5	88.8	96.4
Monopoly	51.2	53.4	29.4	32.1	85.2	95.3
Plastic	25.2	28.2	13.5	20.6	88.7	98.2

on a standard CPU was recently proposed [10]. This indicates that a real-time implementation of the proposed algorithm is feasible. Indeed, the modification of the growing algorithm and integration with the ToF data does not bring any significant extra computational costs. The algorithmic complexity remains the same, since we have only slightly modified the similarity score used inside the growing procedure. It is true that prior to the growing process, the ToF data must be triangulated. Nevertheless, this can be done extremely efficiently using computer graphics techniques and associated software libraries.

5.4 Conclusions

We have proposed a novel correspondence growing algorithm, performing fusion of a range sensor and a pair of passive color cameras, to obtain an accurate and dense 3D reconstruction of a given scene. The proposed algorithm is robust, and performs well on both textured and textureless surfaces, as well as on ambiguous repetitive patterns. The algorithm exploits the strengths of the ToF sensor and those of stereo matching between color cameras, in order to compensate for their individual weaknesses. The algorithm has shown promising results on difficult real-world data, as well as on challenging standard data sets which quantitatively corroborates its favorable properties. Together with the strong potential for real-time performance that has been discussed, the algorithm would be practically very useful in many computer vision and robotic applications.

References

1. Alenyà, G., Dellen, B., Torras, C.: 3D Modelling of leaves from color and tof data for robotized plant measuring. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), pp. 3408–3414 (2011)
2. Attamimi, M., Mizutani, A., Nakamura, T., Nagai, T., Funakoshi, K., Nakano, M.: Real-time 3D visual sensor for robust object recognition. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4560–4565 (2010)
3. Castañeda, V., Mateus, D., Navab, N.: SLAM combining ToF and high-resolution cameras. In: IEEE Workshkop on Motion and Video Computing (2011)

4. Cech, J., Matas, J., Perdoch, M.: Efficient sequential correspondence selection by cosegmentation. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1568–1581 (2010)
5. Cech, J., Sanchez-Riera, J., Horaud, R.: Scene flow estimation by growing correspondence seeds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3129–3136 (2011)
6. Čech, J., Šára, R.: Efficient sampling of disparity space for fast and accurate matching. In: Proceedings of BenCOS Workshop CVPR (2007)
7. Chan, D., Buisman, H., Theobalt, C., Thrun, S.: A noise-aware filter for real-time depth upsampling. In: ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (2008)
8. Dal Mutto, C., Zanuttigh, P., Cortelazzo, G.M.: A probabilistic approach to ToF and stereo data fusion. In: Proceedings of 3D Data Processing, Visualization and Transmission, Paris (2010)
9. Diebel, J., Thrun, S.: An application of Markov random fields to range sensing. In: Proceedings on Neural Information Processing Systems (NIPS) (2005)
10. Dobias, M., Sára, R.: Real-time global prediction for temporally stable stereo. In: Proceedings of IEEE International Conference on Computer Vision Workshops, pp. 704–707 (2011)
11. Droeschel, D., Stückler, J., Holz, D., Behnke, S.: Towards joint attention for a domestic service robot—person awareness and gesture recognition using time-of-flight cameras. In: Proceedings of International Conference on Robotic and Animation (ICRA), Shanghai, pp. 1205–1210 (2011)
12. Fischer, J., Arbeiter, G., Verl, A.: Combination of time-of-flight depth and stereo using semiglobal optimization. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), pp. 3548–3553 (2011)
13. Freedman, B., Shpunt, A., Machline, M., Arieli, Y.: Depth Mapping Using Projected Patterns. US Patent No. 8150412 (2012)
14. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. In: Proceedings of the Asian Conference on Computer Vision (ACCV), pp. 25–38 (2010)
15. Gould, S., Baumstarck, P., Quigley, M., Ng, A.Y., Koller, D.: Integrating visual and range data for robotic object detection. In: Proceedings of European Conference on Computer Vision Workshops (2008)
16. Hansard, M., Horaud, R., Amat, M., Lee, S.: Projective alignment of range and parallax data. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3089–3096 (2011)
17. Jebari, I., Bazeille, S., Battesti, E., Tekaya, H., Klein, M., Tapus, A., Filliat, D., Meyer, C., Sio-Hoi, I., Benosman, R., Cizeron, E., Mamanna, J.-C., Pothier, B.: Multi-sensor semantic mapping and exploration of indoor environments. In: Technologies for Practical Robot Applications (TePRA), pp. 151–156 (2011)
18. Mesa Imaging AG. http://www.mesa-imaging.ch
19. Moravec, H.P.: Toward automatic visual obstacle avoidance. In: 5th International Conference Artificial Intelligence (ICAI), pp. 584–594 (1977)
20. Newcombe, R.A., Davison, A.J.: Live dense reconstruction with a single moving camera. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1498–1505 (2010)
21. Park, J., Kim, H., Tai, Y.-W., Brown, M.-S., Kweon, I.S.: High quality depth map upsampling for 3D-TOF cameras. In: Proceedings of IEEE International Conference on Computer Vision (ICCV) (2011)
22. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vision **47**, 7–42 (2002)
23. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proceedings of Computer Vision and Pattern Recognition (CVPR), pp. 519–528 (2006)
24. Stückler J., Behnke, S.: Combining depth and color cues for scale- and viewpoint-invariant object segmentation and recognition using random forests. In: Proceedings of IEEE/ RSJ International Conference on Robots and Systems (IROS), pp. 4566–4571 (2010)

25. Wang, L., Yang, R.: Global stereo matching leveraged by sparse ground control points. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3033–3040 (2011)
26. 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)