High-Resolution Depth Maps Based on TOF-Stereo Fusion

Vineet Gandhi, Jan Cech, Radu Horaud

HAL Id: hal-00725616
https://hal.inria.fr/hal-00725616

Submitted on 27 Aug 2012

High-Resolution Depth Maps Based on TOF-Stereo Fusion

Vineet Gandhi[†], Jan Čech, and Radu Horaud
INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France
{vineet.gandhi, jan.cech, radu.horaud}@inria.fr

Abstract— The combination of range sensors with color cameras can be very useful for robot navigation, semantic perception, manipulation, and telepresence. Several methods of combining range- and color-data have been investigated and successfully used in various robotic 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, since the resolution of current range sensors is much less than the resolution of 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, or if the scene exhibits complex self-occlusions. Range sensors provide coarse depth information regardless of presence/absence of texture. The use of a calibrated system, composed of a time-of-flight (TOF) camera and of a stereoscopic camera pair, allows data fusion thus overcoming the weaknesses of both individual sensors. 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 based on a Bayesian model which combines an image similarity score 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 color-range sensors, e.g., Kinect. Moreover, the algorithm potentially exhibits real-time performance on a single CPU.

(a) A TOF-stereo setup

(b) The TOF image is shown in the upper-left corner of a color image.   (c) The proposed method delivers a high-resolution depth map.

Fig. 1. (a) Two high-resolution color cameras (2.0MP at 30FPS) are combined with a single low-resolution time-of-flight camera (0.03MP at 30FPS). (b) A 144×177 TOF image and a 1224×1624 color image 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.
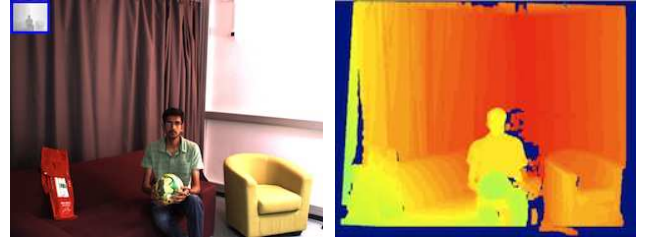
I. INTRODUCTION

An advanced computer vision system should be able to provide both accurate *color and depth* information for each pixel at high resolution. Such a system can be very useful for automated vision problems especially in the context of robotics, 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 pattern, or in the presence of occlusions. The main drawback is that 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, 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 at 30 frames/second, as in Fig. 1.

A. Related Work

The combination of a depth sensor with a color camera has been exploited in several robotic applications such as object recognition [14], [21], [2], person awareness, gesture recognition [11], simultaneous localization and mapping (SLAM) [3], [16], robotized plant-growth measurement [1], etc. These methods have to deal with the difficulty of noise in depth measurement and the inferior resolution of range data as compared to the color data. Also, most of these systems are limited to RGB-D, i.e., a *single* color image combined with the range data. Interestingly enough, the recently commercialized Kinect[1] camera falls in the RGB-D family of sensors. We believe that extending this model to an RGB-D-RGB sensor is extremely advantageous because it can incorporate stereoscopic matching and hence better deal with the problems mentioned above.

Stereo matching has been one of the most studied paradigms in computer vision. Several papers, e.g., [19], [20] present an overview of existing techniques and highlight recent progress in stereo matching and stereo reconstruction. Algorithms based on a greedy local search are typically fast but frequently fail to reconstruct the poorly textured regions or ambiguous surfaces. Global methods formulate the matching task as a single optimization problem which leads to minimization of an Markov random field (MRF) energy function of the image similarity likelihood and a prior on the surface smoothness. 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], [5], [6]. The correspondences are grown from a small set of initial correspondence seeds. Interestingly, they are not particularly sensitive to wrong input seeds. They are significantly faster than the global approaches, but they have difficulties in presence of non textured 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, which 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 an external prior-based generative probabilistic model for stereo matching was proposed in [13], [18] for reducing the matching ambiguities. The prior used was based on surface-triangulation on initially-matched distinctive interest points in the images. Again, in the absence of textured regions, such support points are either not available or are not reliable enough and the priors are erroneous. Consequently, the methods produce artifacts in cases the priors win over the data and the solution is biased towards the incorrect priors. This clearly shows the need for more accurate prior models. [22] integrates 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. A laser scanner is difficult to operate and cannot provide range information fast enough such that it can be used in a practical robotic application.

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

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

The problem of depth map upsampling has been previously addressed. In [7] a noise-aware filter for adaptive multi-lateral upsampling of TOF depth maps is presented. The work described in [14] extends the model of [9] and demonstrates that the object detection accuracy can be significantly improved by combining a state-of-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 super-resolution technique in order to infer the depth at every pixel. The proposed method is slow: It takes around 10 seconds 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.

B. Contributions

In this paper we propose a novel Bayesian 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 [15], e.g., Fig. 1, allows the range data to be 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 novelty is that the range data 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 Bayesian *fusion*, between the mesh-based surface (initialized from the sparse range data) and the seed-growing stereo matching algorithm itself, combines the merits of the two 3D sensing methods and overcomes the limitations outlined above. The proposed fusion model can be incorporated within virtually any stereo algorithm that is based on energy minimization and which requires proper initialization. It is, however, particularly efficient and accurate when used in combination with match-propagation methods.

The remainder of this paper is structured as follows: Section II describes the proposed range-stereo fusion algorithm. Experimental results on a real dataset and evaluation of the method, are presented in section III. Finally, section IV draws some conclusions.

II. THE PROPOSED 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 [15]. 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) in the respective images. Each element (u, u', v) denotes a point in the disparity space[4]. 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 neighborhood of given seed correspondences. Hereby, only a small fraction of disparity space is visited, 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 required 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

[4]The disparity space is a space of all potential correspondences [19].

Algorithm 1 Growing algorithm with sensor fusion

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

1: Compute the prior disparity map D_p by interpolating seeds $\mathcal{S}$.
2: Compute $\text{simil}(s|I_L, I_R, D_p)$ for each every seed $s \in \mathcal{S}$.
3: Initialize empty disparity map D of size I_L (and D_p).
4: **repeat**
5:   Draw seed $s \in \mathcal{S}$ of the best $\text{simil}(s|I_L, I_R, D_p)$ value.
6:   **for** each of the four best neighbors $i \in \{1, 2, 3, 4\}$
$$q_i^* = (u, u', v) = \underset{q \in \mathcal{N}_i(s)}{\text{argmax}} \; \text{simil}(q|I_L, I_R, D_p)$$
   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 $\mathcal{S} := \mathcal{S} \cup \{q_i^*\}$.
10:      Update the output map $D(u, v) = u - u'$.
11:    **end if**
12: **end for**
13: **until** $\mathcal{S}$ is empty
14: **return** disparity map D.

---

likelihood based on disparity estimate by interpolating the rough triangulated TOF surface. This can be viewed as a prior cast over the disparity space.

The growing algorithm is summarized in Sec. II-A. The processing of the TOF correspondence seeds is explained in Sec. II-B, and the sensor fusion based similarity statistic is described in Sec. II-C.

A. The Growing Procedure

The growing algorithm is sketched in pseudo-code as Alg. 1. The input is a pair of rectified images (I_L, I_R), a set of (refined) TOF seeds $\mathcal{S}$, and a parameter τ which directly controls a trade-off between accuracy and density of the matching. 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 $\text{simil}(s|I_L, I_R, D_p)$ of the correspondence, which measures both the photometric consistency of the possible correspondence and consistency with the prior, is computed for all seeds $s = (u, u', v) \in \mathcal{S}$, Step 2. Recall that the seed s stands for a correspondence $(u, v) \leftrightarrow (u', v)$ between pixels in 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 $\{\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3, \mathcal{N}_4\}$ of the pixel correspondence, such that in each respective direction (left, right, up, down) the algorithm searches the disparity in a range ± 1 pixel from the disparity of the seed, Step 6. If the similarity statistic of a candidate exceeds threshold τ, then a new correspondence is found, Step 8. It becomes a new seed, and output disparity map D is updated.
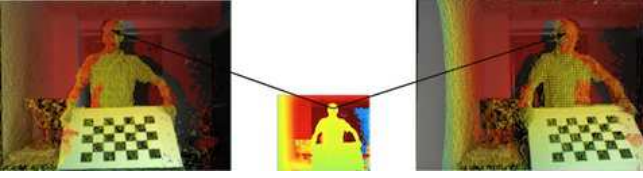
Fig. 2. Projection of TOF sensor data on left and right images. The points are color coded and the color represents disparity such that colder colors are closer to the cameras. The images are not in the true scale. Notice wrong correspondences on the computer screen due to low reflectance and artifacts along occlusion boundaries.
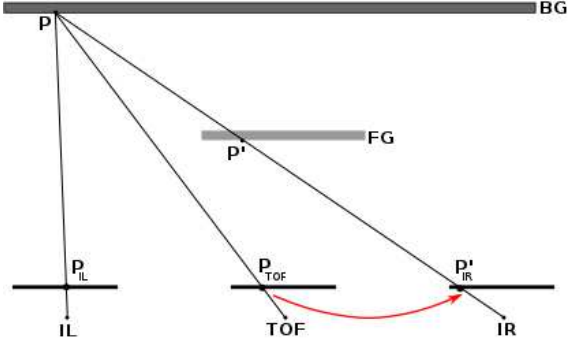

Fig. 3. The effect of occlusions. A background (BG) point P is seen in the left image (IL) and in the TOF image, while it is occluded by a foreground object (FG) and hence not seen in the right image (IR). In the process of reprojection of 3D TOF points, a wrong correspondence ($P_{IL} \leftrightarrow P'_{IR}$) is produced.

The process repeats until there are no more seeds to be grown.

The algorithm is fairly insensitive to wrong initial seeds. Since the seeds compete to be matched in the best first strategy, the wrong seeds typically have low score $\text{simil}(s)$ and therefore when they are drawn in Step 5, the involved pixels are likely to have been matched already. For more details on the growing algorithm, we refer to [6], [4].

B. TOF Seeds and Their Refinement

The original version of the seed-growing stereo algorithm [6] uses an initial set of seeds $\mathcal{S}$ 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 in 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 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. Obviously, regularly distributed seeds will provide a better coverage of the observed scene.

However, TOF seeds are not always accurate. Indeed, when a 3D point is projected onto the left- and the right-image, it does not always yield a valid stereo match. There
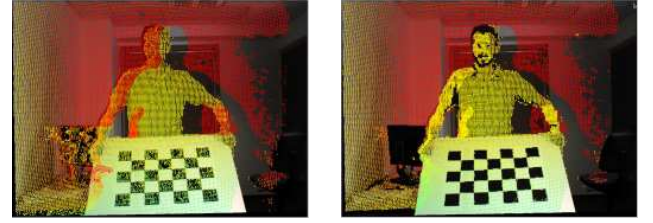
(a) Original set of seeds   (b) Refined set of seeds

Fig. 4. An example of the effect of correcting the set of seeds on the basis that they should be regularly distributed.

may be several sources of error which make the TOF seeds less reliable than one would have expected, e.g., Fig. 2 and Fig. 3. In detail:

1) *Imprecision due to the calibration data.* The transformations allowing to project the 3D TOF points onto the 2D images are obtained via a complex sensor calibration process [15]. This introduces a localization error up to 2-3 pixels.
2) *Outliers due to the physical/geometric properties of the scene.* Range sensors are based on active light and on the assumption that the active beam of light travels 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- and 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 image is only seen into the left or right image, e.g., Fig. 3.

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 section II-A, we implemented a straightforward refinement step in order to detect and eliminate these kind of bad seed data, prior to applying Alg. 1. Firstly, the seeds that lie in low-intensity (very dark) regions are discarded since TOF-based range data are not reliable in these cases. Secondly, 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. See occlusion boundary in Fig. 4(a).

Such seeds are detected by counting the occupancy of small 5×5 pixel windows around every seed point in both images. If there is more than one seed point, the seeds belonging to the background are discarded. Refined set of seeds is shown in Fig. 4(b). The refinement procedure typically filters 10-15% of all seed points.
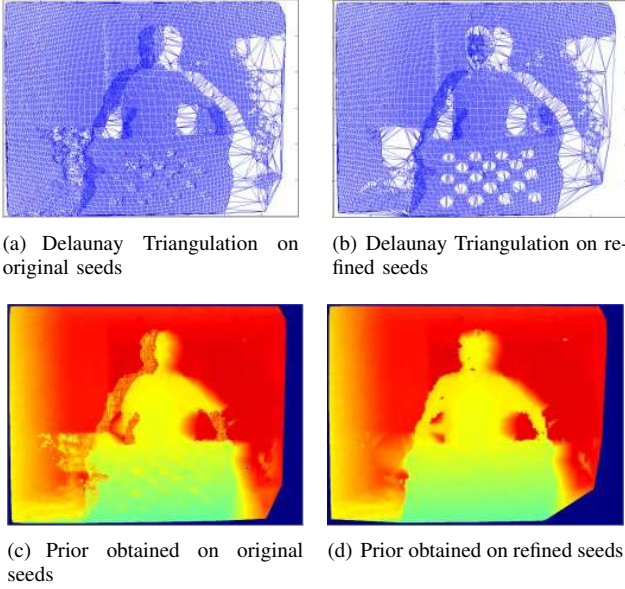
(a) Delaunay Triangulation on original seeds

(b) Delaunay Triangulation on refined seeds

(c) Prior obtained on original seeds

(d) Prior obtained on refined seeds

Fig. 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.

C. Similarity Statistic Based on Sensor Fusion

The original seed-growing matching algorithm [6] uses Moravec's normalized cross correalation [17] (MNCC),

$$\text{simil}(s) = \text{MNCC}(w_L, w_R) = \frac{2\text{cov}(w_L, w_R)}{\text{var}(w_L) + \text{var}(w_R) + \epsilon} \quad (1)$$

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. Fig. 5(a) and 5(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:

$$d^* = \underset{d}{\text{argmax}} \ P(d|I_L, I_R, D_p). \quad (2)$$

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:

$$d^* = \underset{d}{\text{argmax}} \ P(I_L, I_R|d) \ P(D_p|d), \quad (3)$$

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

$$P(I_L, I_R|d) \propto \text{EXPSSD}(w_L, w_R) =$$
$$= \exp\left(-\frac{\sum_{i=1}^{n \times n}(w_L(i) - w_R(i))^2}{\sigma_s^2 \sum_{i=1}^{n \times n}(w_L(i)^2 + w_R(i)^2)}\right), \quad (4)$$

and as

$$P(D_p|d) \propto \exp\left(-\frac{(d - d_p)^2}{2\sigma_p^2}\right) \quad (5)$$

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

$$\text{simil}(s|I_L, I_R, D_p) = \text{EPC}(w_L, w_R, D_p) =$$
$$= \exp\left(-\frac{\sum_{i=1}^{n \times n}(w_L(i) - w_R(i))^2}{\sigma_s^2 \sum_{i=1}^{n \times n}(w_L(i)^2 + w_R(i)^2)} - \frac{(d - d_p)^2}{2\sigma_p^2}\right). \quad (6)$$

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 towards 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.

III. EXPERIMENTS

Our experimental setup comprises one Mesa Imaging Swissranger™ SR4000 TOF camera and a pair of high-resolution Point Grey[5] color cameras, e.g., Fig. 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 144×177 and the color cameras have a resolution of 1224×1624. Recall that Fig. 1(b) highlights the resolution differences between the TOF and color images.This camera system was calibrated using the method described in [15].

In all our experiments, we set the parameters of the method as follows: Windows of 5×5 pixels were used for matching ($n = 5$), matching threshold in Alg. 1 to $\tau = 0.5$, the balance between the photometric and range sensor likelihoods was set to $\sigma_s^2 = 0.1$ and to $\sigma_p^2 = 0.001$ in (6).

We show both qualitatively and quantitatively (using datasets with ground-truth) benefits of the range sensor and
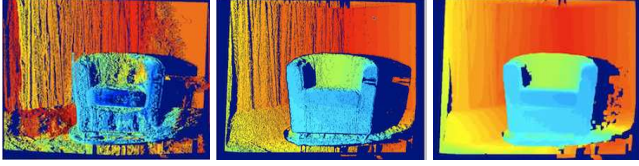
[5]http://www.ptgrey.com/

(a) Input data RGB-TOF-RGB (the true scale is shown in Fig. 1(b)).
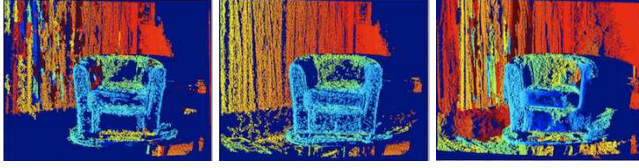
(b) MNCC-Harris   (c) MNCC-TOF   (d) EXPSSD-Harris

(e) EXPSSD-TOF   (f) **EPC** (proposed)   (g) **EPC** (gaps filled)

Fig. 6. SET-1. (a) A triplet composed of a pair of color images (left and right) and a TOF image (middle), results obtained (b) using the seed growing stereo algorithm [6] combined Harris seeds and MNCC statistic, (c) using TOF seeds and MNCC statistic , (d) using Harris seeds and EXPSSD statistic , (e) using TOF seeds with EXPSSD statistics. Results obtained with proposed full stereo-TOF fusion model using EPC similarity statistic (f) and full model EPC after filling small gaps (g).

(a) Input data RGB-TOF-RGB (the true scale is shown in Fig. 1(b)).


(b) MNCC-Harris   (c) MNCC-TOF   (d) EXPSSD-Harris

(e) EXPSSD-TOF   (f) **EPC** (proposed)   (g) **EPC** (gaps filled)

Fig. 7. SET-2. Please refer to the caption of Fig. 6 for explanation.

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 Alg. 1 in Step. 8 enforces one-to-one pixel matching. Nevertheless this way, they can be filled easily if needed.

A. Real-Data Experiments

We captured two real-world datasets using the camera setup described above, SET-1 in Fig. 6 and SET-2 in 7. Notice that in both of these examples the scene surfaces are weakly textured.

1) Comparisons between disparity maps: Results as disparity maps are shown color-coded, such that warmer colors are further away from the cameras and unmatched pixels are dark blue.

In Fig. 6(b), we can see that the original algorithm [6] has difficulties in low textured regions which results in large unmatched regions due to MNCC statistic (1), and it

produces several mismatches over repetitive structures on the background curtain, due to erroneous (mismatched) Harris seeds. In Fig. 6(c), we can see that after replacing the sparse erratic Harris seeds with uniformly distributed mostly correct TOF seeds, results improve significantly. There are no more mismatches on the background, but unmatched regions are still large. In Fig. 6(d), the EXPSSD statistic (4) was used instead of MNCC which causes similar mismatches as in Fig. 6(b), 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. 6(e). Significantly better results are finally shown in Fig. 6(f) which is using the full proposed fusion model EPC (6). The EPC statistic compared to EXPSSD has the additional regularizing range sensor likelihood term which guides the growth in ambiguous regions and attracts the solution towards the rough estimate of the TOF camera. Results are further refined by filling small gaps in Fig. 6(g). Similar observations can be made in Fig. 7. The proposed model clearly outperforms the other discussed approaches.

2) Comparisons between reconstructed surfaces: 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 low-resolution depth map
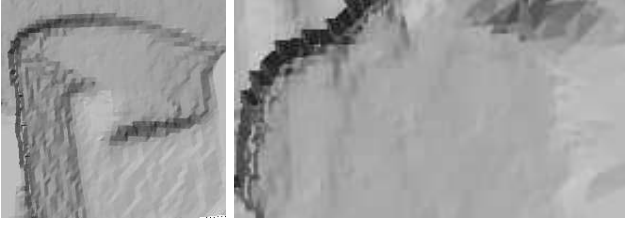
(a) Dense surface reconstruction using the disparity map D_p corresponding to a 2D triangulation of the TOF data points. Zoomed sofa chair and zoomed T-shirt from SET-2 in Fig. 7(a).
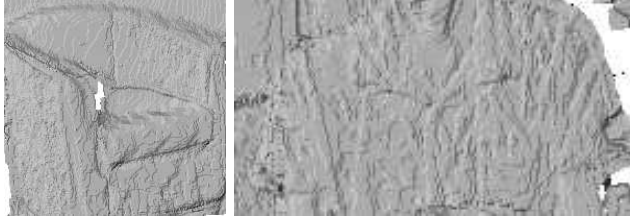
(b) Surface reconstruction using the proposed algorithm (EPC) shown on the same zoomed areas as above, i.e., Fig. 7(g).

Fig. 8. The reconstructed surfaces are shown as relighted 3D meshed 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.

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. 8(a), as well as the reconstructed surfaces associated with the disparity map obtained with the proposed method, Fig. 8(b). Clearly, much more of the surface details are recovered by the proposed method. Notice precise object boundaries and fine details, like a pillow on the sofa chair and a collar of the T-shirt, which appear in Fig. 8(b). This qualitatively corroborates the precision of the proposed method compared to the TOF data.

B. Ground-Truth Evaluation

To quantitatively demonstrate the validity of the proposed algorithm, we carried out an experiment on datasets with associated ground-truth results. Similarly to [8] we use Middlebury dataset [19] and simulated the TOF camera by sampling the ground-truth disparity map.

We used the Middlebury-2006 dataset[6]. 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 any noise or artifacts, but we believe that for validating the proposed method this is satisfactory.

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

Results are shown in Fig. 9. We show 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 non-occluded regions. This error statistic is computed as 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 non-occluded 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 solution of different density. The quantitative evaluation confirms the observation from 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 Alg.1. This is 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 a range sensor. 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 respective scenes. This is again consistent with the experiments with the real TOF sensor, where higher surface details were recovered, see Fig. 8.

C. Computational Costs

The original growing algorithm [6] has low computational complexity due to intrinsic search space reduction. Assuming the input stereo images 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 $\mathcal{O}(n^3)$, [5]. Factor n^3 is the size of the search space where 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 really expensive as well as the computing the new EPC statistic instead of MNCC.

For our experiments, we use an academic (combined Matlab and C) implementation which takes about 5 seconds with 2 MP images. Nevertheless, recently [10] presented an implementation of the growing algorithm [6] which runs in real-time in normal CPU, without parallel hardware. This indicates that a real-time implementation of the proposed algorithm would be feasible.
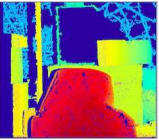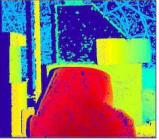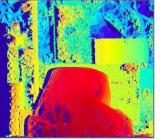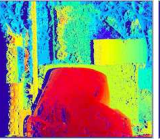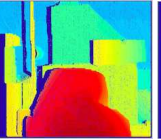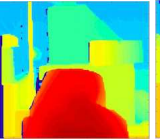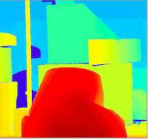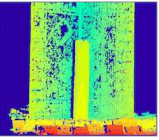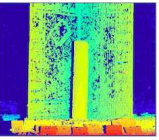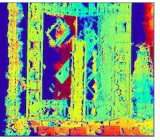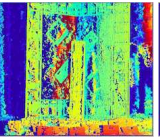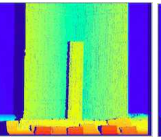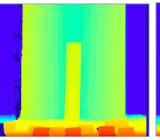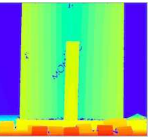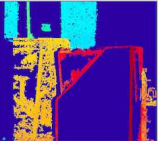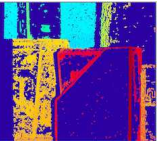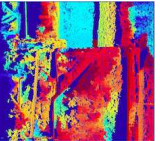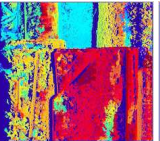
Left image	MNCC-Harris	MNCC-TOF	EXPSSD-Harris	EXPSSD-TOF	**EPC**	**EPC** (gaps filled)	Ground-truth
Lampshade-1	61.5%	64.3%	44.9%	49.5%	88.8%	96.4%	–
Monopoly	51.2%	53.4%	29.4%	32.1%	85.2%	95.3%	–
Plastic	25.2%	28.2%	13.5%	20.6%	88.7%	98.2%	–

Fig. 9. Middlebury dataset. We show left images, results of the same algorithms as in Fig. 6 and 7, and the ground-truth disparity maps. There are error statistics (percentage of correctly matched pixels) below the disparity maps. Observations from the real-world cases are confirmed quantitatively. The algorithm using the full model (EPC) gives clearly the best results.

IV. CONCLUSIONS

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

REFERENCES

[1] G. Alenyà, B. Dellen, and C. Torras, "3d modelling of leaves from color and tof data for robotized plant measuring," in *ICRA*, 2011, pp. 3408–3414.

[2] M. Attamimi, A. Mizutani, T. Nakamura, T. Nagai, K. Funakoshi, and M. Nakano, "Real-time 3D visual sensor for robust object recognition," in *IROS*, 2010, pp. 4560–4565.

[3] V. Castaneda, D. Mateus, and N. Navab, "SLAM combining tof and high-resolution cameras," *IEEE Workshkop on Motion and Video Computing*, 2011.

[4] J. Čech and R. Šára, "Efficient sampling of disparity space for fast and accurate matching," in *In Proc. BenCOS Workshop CVPR*, 2007.

[5] J. Cech, J. Matas, and M. Perdoch, "Efficient sequential correspondence selection by cosegmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1568–1581, 2010.

[6] J. Cech, J. Sanchez-Riera, and R. Horaud, "Scene flow estimation by growing correspondence seeds," in *CVPR*, 2011, pp. 3129–3136.

[7] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "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] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo, "A probabilistic approach to tof and stereo data fusion," in *3DPVT*, May 2010.

[9] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *NIPS*, 2005.

[10] M. Dobias and R. Sára, "Real-time global prediction for temporally stable stereo," in *ICCV Workshop on Live Dense Reconstruction with Moving Cameras*, 2011, pp. 704–707.

[11] D. Droeschel, J. Stückler, D. Holz, and S. Behnke, "Towards joint attention for a domestic service robot - person awareness and gesture recognition using time-of-flight cameras," in *ICRA*, 2011, pp. 1205–1210.

[12] J. Fischer, G. Arbeiter, and A. Verl, "Combination of time-of-flight depth and stereo using semiglobal optimization," in *ICRA*, 2011, pp. 3548–3553.

[13] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *ACCV*, 2010, pp. 25–38.

[14] S. Gould, P. Baumstarck, M. Quigley, A. Y. Ng, and D. Koller, "Integrating visual and range data for robotic object detection," in *ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2)*, 2008.

[15] M. E. Hansard, R. Horaud, M. Amat, and S. Lee, "Projective alignment of range and parallax data," in *CVPR*, 2011, pp. 3089–3096.

[16] I. Jebari, S. Bazeille, E. Battesti, H. Tekaya, M. Klein, A. Tapus, D. Filliat, C. Meyer, Sio-Hoi, Ieng, R. Benosman, E. Cizeron, J.-C. Mamanna, and B. Pothier, "Multi-sensor semantic mapping and exploration of indoor environments," in *TePRA*, 2011, pp. 151 –156.

[17] H. P. Moravec, "Toward automatic visual obstacle avoidance," in *ICAI*, 1977, pp. 584–94.

[18] R. A. Newcombe and A. J. Davison, "Live dense reconstruction with a single moving camera," in *CVPR*, 2010, pp. 1498–1505.

[19] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *IJCV*, vol. 47, no. 1-3, 2002, pp. 7–42.

[20] S. Seitz, B. Burless, J. Diebel, D. Scharstein, and S. R., "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *CVPR*, vol. 1, 2006, pp. 519–528.

[21] J. Stückler and S. Behnke, "Combining depth and color cues for scale- and viewpoint-invariant object segmentation and recognition using random forests," in *IROS*, 2010, pp. 4566–4571.

[22] L. Wang and R. Yang, "Global stereo matching leveraged by sparse ground control points," in *CVPR*, 2011, pp. 3033–3040.

[23] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *CVPR*, 2008.