A Probabilistic Approach to ToF and Stereo Data Fusion

Carlo Dal Mutto Pietro Zanuttigh Guido M. Cortelazzo

Department of Information Engineering
University of Padova
Via Gradenigo 6/B, Padova, Italy
{dalmutto,pietro.zanuttigh,corte}@dei.unipd.it

Abstract

Current 3D video applications require the availability of high quality depth information. Depth information can be acquired real-time by stereo vision systems and ToF cameras. Both solutions present critical issues, that can be overcome by their combined use. In this paper, a heterogeneous acquisition system is considered, made of two high resolution standard cameras (stereo pair) and one ToF camera. The stereo system and the ToF camera must be properly calibrated together in order to operate jointly. Therefore this work introduces first a generalized multi-camera calibration technique which does not exploit only the luminance (color) information, but also the depth information extracted by the ToF camera. A probabilistic ad hoc fusion algorithm is then derived in order to obtain high quality depth information from the information of both the ToF camera and the stereo-pair. Experimental results show that the proposed calibration algorithm leads to a very accurate calibration suitable for the fusion algorithm, that, in turn, allows for precise extraction of the depth information.

1. Introduction

The extraction of depth information suitable for the creation of 3D video content is a very challenging issue. Various systems [5, 23] have been proposed in order to solve this task, each one with pros and cons, and the research on this topic is still active. Traditionally this problem has been tackled by means of stereo vision systems, that exploit the information coming from two or more standard cameras [2, 23, 25]. Stereo vision systems have been greatly improved in the last years and obtained interesting results, however they cannot handle all scene situations (aperture problem). Moreover the most advanced stereo vision systems are characterized by very time-consuming algorithms, not suited for real-time operation. Hence, stereo vision systems do not provide completely satisfactory solutions for the extraction of depth information from generic scenes. Other traditional systems proposed in order to solve such problems are active methods such as structured light or laser scanners. Such methods can obtain better results than passive stereo vision systems, but generally require long acquisition times, therefore they are not suitable for the acquisition of dynamic scenes.

New depth acquisition systems, such as Time of Flight (ToF) range cameras (*e.g.*, Mesa Imaging SwissRanger$^{\text{TM}}$[19], CanestaVision$^{\text{TM}}$[16] chips and similar) have recently reached the market. Such devices compute depth by sending an infrared signal (*e.g.*, the SwissRanger$^{\text{TM}}$emits a radiation with illumination wavelength $850nm$) and measuring the phase shift of the reflected light signals. ToF cameras are quite compact, can extract depth information in real-time and are not very sensitive to scene peculiarities. However, they have a limited resolution (*e.g.*, the SwissRanger$^{\text{TM}}$produces a depth image with resolution 176x144), limited accuracy, and they are very sensitive to the background illumination at the ToF illumination wavelength. Contrary to other active systems, ToF cameras are suitable for the acquisition of dynamic scenes.

Interesting results can be achieved by using an heterogeneous acquisition system, coupling a stereo vision system with some ToF cameras. An acquisition system composed by a ToF camera and a stereo pair is proposed in [27]. The two subsystems are coupled by an empirical application of the belief propagation algorithm, originally proposed in [25]. In [10] the two systems are combined by converting the ToF depth measurement into disparity, and then using it as an initialization for a hierarchical stereo matching algorithm. In [8, 14, 22] information from the different sensors is combined by fusing data on 3D probabilistic occupancy grids, exploiting also silhouette cues. Finally [1] presents a fusion algorithm for the estimation of patchlets exploiting information from a ToF camera and a stereo pair.

This paper proposes a novel method to obtain accurate depth maps from data acquired by a trinocular heterogeneous acquisition system made by a ToF camera T and two standard videocameras $\{L, R\}$, forming a stereo pair $S \triangleq \{L, R\}$. Probabilistic models are derived for both the ToF camera and the stereo pair, then a Bayesian approach is exploited for the fusion of the probability models for T and S, and finally the joint probability is maximized applying a local method. The ToF camera T acquires:

- An amplitude image A_T, which describes the pixel by pixel reflectance of the framed scene at the illumination wavelength;
- A depth image D_T, which gives the pixel by pixel depth information about the framed scene;
- A confidence map C_T (in the case of the SwissRanger$^{\text{TM}}$), which qualifies the pixel by pixel precision of the depth measurements. High values of C_T indicate high precision in the depth measurement, while low values indicate low precision.

The two standard cameras acquire RGB images $\{I_L, I_R\}$. The acquisition scheme is shown in Figure 1. In order to correctly use the information from $\{T, L, R\}$, it is firstly necessary to properly calibrate the proposed system. Once the system is calibrated, a fusion algorithm, that combines the data acquired by the three cameras $\{T, L, R\}$, must be developed. The goal of the fusion algorithm is the estimation of the depth distribution Z of the scene portion framed by all the 3 cameras.

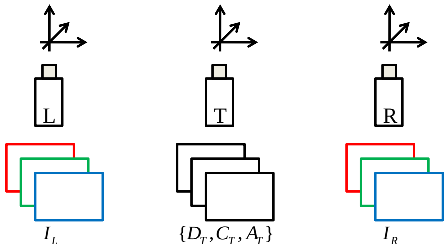

Figure 1. Acquisition system

The paper is organized in the following way: Section 2 presents a novel ad hoc calibration method needed to perform a very accurate system calibration; Section 3 derives a proper fusion algorithm needed to exploit the information coming from all the 3 cameras; Section 4 presents some experimental results about its performance.

2. Calibration

System calibration as well known is subdivided into 2 different tasks: calibration of the intrinsic parameters and calibration of the extrinsic parameters.

2.1. Intrinsic Parameters

Concerning the camera projection properties, the classical Heikkila model [13] is considered for all the 3 cameras. The estimation of the intrinsic parameters and the compensation of radial and tangential distortions can be performed by standard calibration algorithms [26]. All the acquired data $\{A_T, D_T, C_T, I_L, I_R\}$ will be considered in the rest of the paper free from radial and tangential distortion. The calibration of T requires also to compensate for the systematic error in the depth measurement [20], which can be performed by a polynomial correction functional approach [24]. The depth image D_T will be considered in the rest of the paper free from systematic errors in the depth measurement.

2.2. Extrinsic Parameters

A camera reference frame is associated with each of the 3 cameras. The world reference frame is assumed to coincide with the reference frame of T. All the depths reported next, will be considered with references to the T camera reference frame.

The calibration of the extrinsic parameters is the estimation of the relative rototranslations between the 3 camera reference frames $\{T, L, R\}$. For high resolution cameras, the extrinsic parameters' calibration is a standard operation [11, 26], which can be performed with high precision. For low resolution ToF cameras, the problem is quite new, and there are no consolidated procedures giving the necessary precision [9, 21]. Two main factors limiting the precision are:

- the low resolution of the ToF cameras (*e.g.*, the SwissRanger$^{\text{TM}}$produces a depth image with resolution 176×144);
- standard calibration methods just consider reflectance information and not depth information, which is the main information acquired by the ToF cameras.

For the reasons above, an ad hoc calibration algorithm is introduced next, leading to accurate calibration results.

The ToF camera T and the stereo pair S, can both convey 3D information. By a standard planar calibration checkerboard, it is possible to identify a set of n corners P^i: $\mathcal{P} = \{P^i, i = 1, ..., n\}$. Such points project into the 2D camera image points $\mathbf{p}_T^i, \mathbf{p}_L^i, \mathbf{p}_R^i, i = 1, ..., n$; where:

- $\mathbf{p}_T^i \in A_T, D_T, C_T$ and has coordinates (u_T^i, v_T^i);
- $\mathbf{p}_L^i \in I_L$ and has coordinates (u_L^i, v_L^i);
- $\mathbf{p}_R^i \in I_R$ and has coordinates (u_R^i, v_R^i).

As for the coordinates of corners $\mathbf{p}_L^i$ and $\mathbf{p}_R^i$ in the standard camera images, the coordinates of corners $\mathbf{p}_T^i$ can be obtained by the application of a standard corner detector on the amplitude images A_T. For the stereo pair S, a standard calibration of the extrinsic parameters (including stereo rectification) [11, 26] is performed using $\{(\mathbf{p}^i_L, \mathbf{p}^i_R), i = 1, ..., n\}$ as input. The relative rototranslation M_{RL} of the R reference frame with respect to the L reference frame is then obtained. The 3D coordinates with respect to the L reference frame $\{\mathbf{P}^i_S, i = 1, ..., n\}$, of the corners $\{P^i, i = 1, ..., n\}$ are computed from $\{(\mathbf{p}^i_L, \mathbf{p}^i_R), i = 1, ..., n\}$ by triangulation [12].

The 3D coordinates with respect to the T reference frame$\{\mathbf{P}^i_T, i = 1, ..., n\}$, of the corners $\{P^i, i = 1, ..., n\}$ are computed from the values of D_T in $\{\mathbf{p}^i_T, i = 1, ..., n\}$, by inverting the standard pinhole camera equation. For this reason, the 3D coordinates $\{\mathbf{P}^i_T, i = 1, ..., n\}$ are called back-projected coordinates.

Given the set of points $\mathcal{P}$, with coordinates $\{\mathbf{P}^i_S, i = 1, ..., n\}$ with respect to the L reference frame, and coordinates $\{\mathbf{P}^i_T, i = 1, ..., n\}$ with respect to the T reference frame, the estimation of the rototranslation M_L of the L reference frame with respect to the T reference frame is an absolute orientation problem, that can be solved by applying Horn's algorithm [15] and RANSAC [4]. Horn's algorithm gives a closed-form solution to the relative orientation problem, minimizing the sum of the Euclidean distance errors between all corresponding points:

$$\arg\min_{[M_L]} \sum_{i=1}^{n} ||\mathbf{P}^i_T - M_L \cdot \mathbf{P}^i_S||_2. \tag{1}$$

Horn's algorithm allows to obtain the best rototranslation M_L because the closed-form solution prevents falling into local minima, a common issue of gradient-based methods (classically adopted in the minimization task). Moreover, Horn's algorithm fully exploits the ability of the stereo pair and of the ToF camera to perform stand-alone 3D reconstructions. Horn's algorithm is used inside a RANSAC estimation scheme in order to limit the errors due to outliers caused by the ToF depth measurement noise or by badly detected corners.

Finally the relative rototranslation M_R, of the R reference frame with respect to the T reference frame is computed as composition of rototranslations M_L and M_{RL}.

3. Fusion Algorithm

As already said in Section 1, the goal of the fusion algorithm is the estimation of the depth distribution Z of the portion of the scene framed by all the 3 cameras $\{T, L, R\}$ by combining information coming from the ToF camera T and the stereo pair S. The output of the fusion algorithm is an estimate $\hat{Z}$ of the depth distribution Z. Both Z and $\hat{Z}$ are expressed with respect to the T reference frame. $\hat{Z}$ is estimated from the information given by $\{D_T, C_T, I_L, I_R\}$. All these 4 "images", and the depth distribution Z can be modeled as random fields:

- D_T is the random field of the depth measured by T, defined over the lattice of the undistorted images produced by T;
- C_T is the random field of the confidence in the depth measurement performed by T, defined over the lattice of the undistorted images produced by T;
- I_L is the random field of the color images acquired by L, defined over the lattice of the undistorted and rectified images produced by L;
- I_R is the random field of the color images acquired by R, defined over the lattice of the undistorted and rectified images produced by R;
- Z is the random field of the depth distribution of the portion of the scene framed by all the 3 cameras $\{T, L, R\}$, defined over the subset $\mathcal{Z}$ of the lattice of the undistorted images produced by T that is framed also by the 2 cameras $\{L, R\}$ (*i.e.* the final depth estimate $\hat{Z}$ has the same resolution as D_T).

Estimate $\hat{Z}$ is also defined on $\mathcal{Z}$. For simplicity, the random fields can be grouped as $I_T = \{D_T, C_T\}$ and $I_S = \{I_L, I_R\}$. From Bayes rule, the joint posterior probability of Z given $\{I_T, I_S\}$ can be expressed as:

$$P[Z|I_T, I_S] = \frac{P[I_T, I_S|Z]P[Z]}{P[I_T, I_S]}. \tag{2}$$

The goal of the fusion algorithm is to obtain the argument $\hat{Z}$ maximizing eq. (2), i.e,

$$\hat{Z} = \arg\max_{Z} P[Z|I_T, I_S] = \arg\max_{Z} \frac{P[I_T, I_S|Z]P[Z]}{P[I_T, I_S]}. \tag{3}$$

Since the denominator of the RHS does not have any reference to Z, maximizing the RHS of eq. (3) or maximizing:

$$\hat{Z} = \arg\max_{Z} \frac{P[I_T, I_S|Z]P[Z]}{C}, \tag{4}$$

for an arbitrary C independent from Z lead to the same value $\hat{Z}$. Value C can be taken equal to $P[I_T]P[I_S]$. Since Z is uniform (there is no reason for a non-uniform distribution of the depth of a point from a camera), $P[Z]$ is a constant value, therefore eq. (4) can be written as:

$$\hat{Z} = \arg\max_{Z} \frac{P[I_T, I_S|Z]P[Z]P[Z]}{P[I_T]P[I_S]}. \tag{5}$$

Finding a model for conditional probability $P[I_T, I_S|Z]$ is a hard task. A convenient possibility is to assume $\{I_T|Z\}$

and $\{I_S|Z\}$ independent, in this way the RHS of eq. (5) can be approximated as:

$$\arg\max_{Z} \frac{P[I_T|Z]P[Z]}{P[I_T]} \frac{P[I_S|Z]P[Z]}{P[I_S]}. \tag{6}$$

Hence the final expression for $\hat{Z}$ is:

$$\hat{Z} = \arg\max_{Z} P[Z|I_T, I_S] \approx \arg\max_{Z} P[Z|I_T]P[Z|I_S]. \tag{7}$$

3.1. ToF camera model

Posterior probability $P[Z|I_T]$ of the depth distribution given the measurements of the ToF camera T can be expanded as $P[Z|D_T, C_T]$. For each pixel $\mathbf{p} \in \mathcal{Z}$ the random field Z can be considered as juxtaposition of independent per-pixel measurements $Z(\mathbf{p})$. This is equivalent to assume the independent ray model for the ToF camera T. According to this model, for each pixel $\mathbf{p} \in \mathcal{Z}$, its depth $Z(\mathbf{p})$ is measured independently from the depths of its neighboring pixels. Such a model is not always accurate, especially in the presence of depth discontinuities. As reported in [20], each pixel in the ToF measurement process also receives an energy contribution from its neighbors (an effect typically called scattering) which near discontinuities may have quite different values. Hence a better model will be used later in this section in order to account for scattering near discontinuities. For each $\mathbf{p} \in \mathcal{Z}$, one can consider conditional probability $P[Z(\mathbf{p})|D_T, C_T]$ as typical of probability measurement processes where D_T is the process of the measurement and C_T is the process describing the measurement precision. At each pixel $\mathbf{p} \in \mathcal{Z}$, the error in the depth measurement of a ToF camera can be mainly described as the sum of the following error components [20]:

1. a thermal noise component, with Gaussian distribution;
2. a quantization error component;
3. a photon shot noise component, with Poisson distribution;
4. a scattering generated noise component especially in presence of depth discontinuities.

The main error components are the thermal noise, and, in presence of depth discontinuities, the scattering generated noise. The other error components can be neglected or approximated as a part of the thermal noise. The thermal noise is characterized by a normal distribution with zero mean and variance σ_t^2. An estimate of σ_t^2 can be obtained from the confidence map C_T evaluated at the considered pixel p. High values of the confidence map C_T mean low values of σ_t^2 and vice-versa. The scattering generated noise does not have a predictable distribution. Following the model introduced in [3], it is assumed to be characterized by a normal distribution with zero mean and variance σ_s^2 that is the variance of the measured depths in the second order neighborhood of the considered pixel $\mathbf{p}$.

As stated before, near discontinuities the thermal noise component can be neglected, while far from discontinuities it is the scattering generated noise component that can be neglected. For the above reasons, posterior probability $P[Z(\mathbf{p})|I_T]$ can be expressed as:

$$P[Z(\mathbf{p})|I_T] \sim \mathcal{N}(d, \sigma_w^2); \tag{8}$$

where d is the value of D_T at the pixel $\mathbf{p}$, and $\sigma_w = \max(\sigma_t, \sigma_s)$. For practical purposes, in order to reduce the time complexity of the algorithm, one may crop the distribution within interval $\mathcal{P}_s = [d - 3\sigma_w, d + 3\sigma_w]$, called the "practical support " of the ToF probability distribution, since $P[Z(\mathbf{p}) \in \mathcal{P}_s] = 0.997$.

3.2. Stereo pair model

While a distribution model for the ToF camera posterior probability can be proposed on the basis of well known physical quantities, for the stereo pair only a heuristic , nevertheless reasonable, probability model for $P[Z|I_L, I_R]$ can be obtained. As in the case of the ToF, the random field Z can be considered as the juxtaposition of per-pixel measurements $Z(\mathbf{p})$ for each $\mathbf{p} \in \mathcal{Z}$. For each $\mathbf{p} \in \mathcal{Z}$, the posterior probability $P[Z(\mathbf{p})|I_L, I_R]$ relates probability of depth values $Z(\mathbf{p})$ with the similarity between I_L and I_R in the neighborhood of the point projections into the same images. In order to derive very precisely the proposed model, one must first of all identify the occluded pixels in both I_L and I_R. The occluded pixels can be approximately estimated by first rendering the 3D reconstruction obtained from D_T from the view points of the images $\{I_L, I_R\}$ and then by comparing the depth of each $\mathbf{p} \in \mathcal{Z}$ with the z-buffer of the rendered scene in order to check if it is visible from the considered image. This approximation holds especially in the case of a scene surface not to close to the acquisition system and in the case of a setup with T positioned in between L and R. This approximation is well suited especially in the case of a scene surface not to close to the acquisition system.

For each $\mathbf{p} \in \mathcal{Z}$ occluded in I_L or I_R, the stereo posterior probability can be assumed uniformly distributed in the practical support of $P[Z(\mathbf{p})|I_T]$, *i.e.*,

$$P[Z(\mathbf{p})|I_L, I_R] \sim \mathcal{U}([d - 3\sigma_w, d + 3\sigma_w]). \tag{9}$$

For each non-occluded pixel $\mathbf{p} \in \mathcal{Z}$, with coordinates $[u, v]$, the posterior probability $P[Z(\mathbf{p})|I_L, I_R]$ can be computed as a normalized matching cost of the stereo pair, by the following steps:

1. The interval $[d - 3\sigma_w, d + 3\sigma_w]$ is sampled with the precision of the stereo pair, and then interpolated k_{int} times in order to have a sub-pixel precision, obtaining a set of values $z_i(\mathbf{p}), i = 1, ..., m$. The experimental results of this paper have been obtained with $k_{int} = 8$;
2. All the points with coordinates $[u, v]$ and depths $z_i(\mathbf{p})]^T, i = 1, ..., m$, are back-projected and re-projected into the images I_L and I_R at locations $\mathbf{p}_{L,i} = [u_{L,i}, v_{L,i}], i = 1, ..., m$ and $\mathbf{p}_{R,i} = [u_{R,i}, v_{R,i}], i = 1, ..., m$;
3. For each of the projected couple of points, a cost function $\mathcal{C}_i(p)$ is computed as the TAD ("Truncated Sum of Absolute Differences") inside an aggregation support $\mathcal{S}_i(p)$, as for standard stereo pairs [23]:

   $$\mathcal{C}_i(\mathbf{p}) = min \left\{ \sum_{\mathbf{m}_L, \mathbf{m}_R \in \mathcal{S}_i(p)} |I_L(\mathbf{m}_L) \ominus I_R(\mathbf{m}_R)|, T_h \right\} \tag{10}$$

   where T_h is the truncation threshold and:

   $$|I_L(\mathbf{m}_L) \ominus I_R(\mathbf{m}_R)| \triangleq \sum_{c=r,g,b} |I_{L,c}(\mathbf{m}_L) - I_{R,c}(\mathbf{m}_R)|, \tag{11}$$

   in which $\{r, g, b\}$ are the 3 color components of the images $\{I_L, I_R\}$. The aggregation support $\mathcal{S}_i(\mathbf{p})$ is computed for each pixel with a multiple windows approach [6] in order to obtain very good results especially for points very close to discontinuities;
4. Finally, for each depth $z_i(\mathbf{p}), i = 1, ..., m$, the probability $P[Z = z_i(\mathbf{p})|I_L, I_R]$ is computed, as proposed in [25], as:

   $$P[Z(\mathbf{p}) = z_i(\mathbf{p})|I_L, I_T] = \frac{e^{-\frac{\mathcal{C}_i(\mathbf{p})}{\sigma_I}}}{\sum_{i=1}^{m} e^{-\frac{\mathcal{C}_i(\mathbf{p})}{\sigma_I}}}, \tag{12}$$

   where σ_I is the noise standard deviation in images $\{I_L, I_R\}$.

3.3. Full model

The full probability model of the depth distribution inside the practical support $[d - 3\sigma_w, d + 3\sigma_w]$ can be finally obtained by combining (7), (8) and (9) for occluded points:

$$P[Z(\mathbf{p}) = z_i(\mathbf{p})|I_T, I_S] \propto \frac{\exp\left[-\frac{(z_i(\mathbf{p})-d)^2}{2\sigma_w^2}\right]}{\sqrt{2\pi\sigma_w^2}}, \tag{13}$$

and combining (7), (8) and (12) for non-occluded points:

$$P[Z(\mathbf{p}) = z_i(\mathbf{p})|I_T, I_S] = \frac{\exp\left[-\frac{(z_i(\mathbf{p})-d)^2}{2\sigma_w^2}\right]}{\sqrt{2\pi\sigma_w^2}} \frac{e^{-\frac{\mathcal{C}_i(\mathbf{p})}{\sigma_I}}}{\sum_{i=1}^{m} e^{-\frac{\mathcal{C}_i(\mathbf{p})}{\sigma_I}}} \tag{14}$$

The full model is the set of output distributions of the fusion algorithm applied to each single pixel. It does not impose any explicit global model or constraint to the depth distribution Z. Hence, for each $\mathbf{p} \in \mathcal{Z}$, the best estimated depth $\hat{Z}(\mathbf{p})$ is selected as:

$$\hat{Z}(\mathbf{p}) = \arg\max_{z_i(\mathbf{p})} P[Z(p) = z_i(\mathbf{p})|I_T, I_S], i = 1, ..., m, \tag{15}$$

where $P[Z(\mathbf{p}) = z_i(\mathbf{p})|I_T, I_S]$ is given by (13) for occluded points and by (14) for non-occluded points, as summarized by the following pseudo-code:

```
Input: $\{D_T, C_T, I_L, I_R\}$
Output: $\hat{Z} = \{\hat{Z}(\mathbf{p}), \mathbf{p} \in \mathcal{Z}\}$, estimation of the depth
        distribution Z
foreach $p \in \mathcal{Z}$ do
    $\sigma_t$ = function of $C_T(\mathbf{p})$;
    $\sigma_s$ = std. deviation of $D_T$ in the 2-neighbor. of $\mathbf{p}$;
    $\sigma_w = \max(\sigma_t, \sigma_s)$;
    $z_i(\mathbf{p})$ samples in $[d - 3\sigma_w, d + 3\sigma_w], i = 1, ..., m$;
    for i=1 to I do
        calculate the ToF prob. as in eq. (8);
        if (i framed by all $\{T, L, R\}$ && p non-occl.)
        then
            calculate the Stereo prob. as in eq. (9);
            calculate the joint prob. as in eq. (13);
        else
            calculate the Stereo prob. as in eq. (12);
            calculate the joint prob. as in eq. (14);
        end
        select $\hat{Z}(\mathbf{p})$ as the $z_i(\mathbf{p}), i = 1, ..., m$, that
        maximizes the joint prob.;
    end
end
```

Algorithm 1: Fusion algorithm pseudo-code

4. Experimental Results

In order to analyze the performance of the fusion algorithm, some experiments on both synthetic and real data have been performed.

4.1. Synthetic Scenes

Synthetic data offer the advantage of evaluating the performance of the fusion algorithm against a ground truth. A set of 8 scenes, generated in the Autodesk 3ds Max^{TM}framework, (available at the url: http://lttm.dei.unipd.it/downloads/3DPVT10/) is considered in order to analyze the performances on different situations. The scenes are framed by a synthetic acquisition system, that simulates the real one.

The synthetic acquisition system is made by two synthetic standard RGB cameras $\{L, R\}$, with $8mm$ optics and horizontal field of view of 33.4^o, that acquire two RGB images $\{I_L, I_R\}$, with a resolution of 1032×778, and form a stereo pair S with a baseline of $20cm$; and by a synthetic ToF camera T, with focal of $10mm$ and horizontal field of view of 43.6^o, that acquires a 16-bit depth map D_T (depth image) with near-plane set to 0, far-plane set to $5m$ and resolution 176×144. The synthetic ToF camera is positioned in between the synthetic standard cameras L and R. An example of the acquired data is shown in Figure 2.

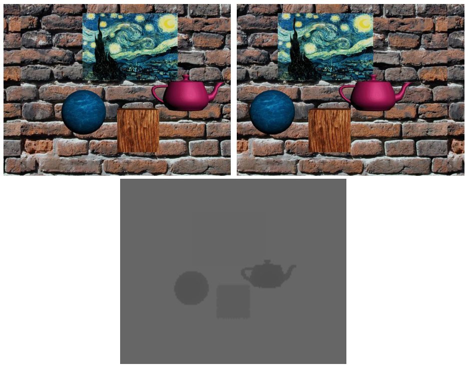

Figure 2. Images $\{I_L, I_R, D_T\}$, acquired by the acquisition system

In order to properly analyze the fusion algorithm performance, a Gaussian noise component with zero mean and variable standard deviation is added to all the three images $\{D_T, I_L, I_R\}$. With respect to the ToF camera T, the Gaussian error is the synthetic version of the depth measurement error. In the synthetic case, there is no confidence map C_T, so σ_w is assumed equal to the standard deviation of the added noise. The reconstruction error, after the fusion algorithm application, is computed as the mean squared error of the estimated depth distribution $\hat{Z}$ with respect to the real depth distribution. The reconstruction error performances of the fusion algorithm as a function of the image noise is compared against that of the ToF alone and of the stereo system. Figure 3 shows quantitative results as a function of the noise in the ToF data while image noise has a fixed standard deviation; Figure 4 refers instead to the opposite case. Qualitative results are shown in Figure 5.

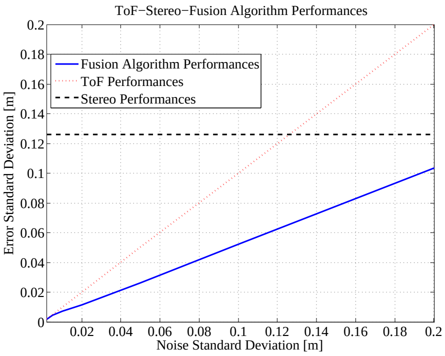

Figure 3. Reconstruction error of the fusion algorithm compared to the one of the ToF and of the stereo system. The abscissa represent the standard deviation of the noise in D_T while the noise in $\{I_L, I_R\}$ has standard deviation $\sigma_I = 0.05m$.

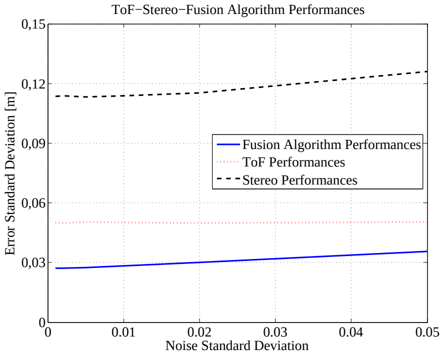

Figure 4. Reconstruction error of the fusion algorithm compared to the one of the ToF and of the stereo system. The abscissa represent the standard deviation of the noise in $\{I_L, I_R\}$ while noise in D_T has standard deviation $\sigma_w = 0.05m$.

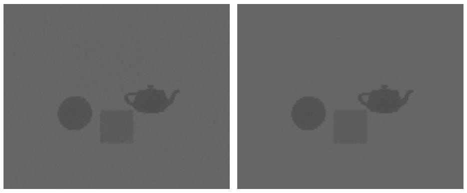

Figure 5. On the left: noisy depth image D_T (noise standard deviation 0.02). On the right: estimated depth distribution $\hat{Z}$ after application of the fusion algorithm.

4.2. Middlebury Scenes

An interesting performance evaluation of the fusion algorithm effectiveness can also be obtained by analyzing its application on data coming from the classical Middlebury repository [18]. The considered acquisition system is made by full-resolution views 1 and 5 of the Middlebury framework. In order to simulate the acquisition of an actual ToF camera, the depth map is obtained by downsampling by a factor of 10 the disparity image relative to view 1 and by converting it into a depth map Z by applying to each pixel the transformation:

$$z = \frac{bf}{d}, \tag{16}$$

where d is the per pixel value of the disparity image, b is the baseline of the stereo pair formed by the cameras acquiring views 1 and 5, and f is the cameras' focal. The depth map Z obtained in this way is considered as the ground-truth. In order to properly analyze the fusion algorithm performance, a Gaussian noise component with 0 mean and variable standard deviation is added to the depth Z. As in the case of synthetic scenes, there is no confidence map C_T, so σ_w is assumed to be equal to the standard deviation of the added noise.

The reconstruction error after the fusion algorithm application is computed as mean squared error of the estimated depth distribution $\hat{Z}$ with respect to the depth distribution Z of the Middlebury groundtruth.

The performance of the fusion algorithm is qualitatively compared against that of the ToF alone in Fig. 6.

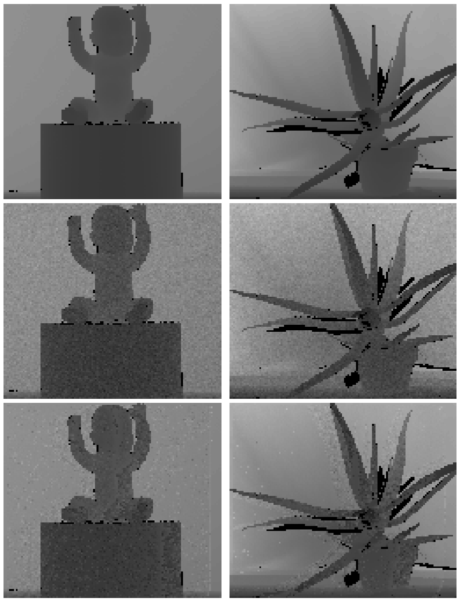

Figure 6. On the top, noise-free depth map. In the middle, noisy depth map before the application of the fusion algorithm (noise standard deviation $\sigma_w = 0.05m$). On the bottom, depth map after the application of the fusion algorithm.

4.3. Real Scenes

A qualitative evaluation of the performance of the fusion algorithm can be obtained by analyzing its application on real scenes' data.

The acquisition system is made by a ToF camera and two standard BASLER scA1000$^{\text{TM}}$RGB cameras $\{L, R\}$, with $8mm$ optics and horizontal field of view of 33.4^o, that acquire RGB images $\{I_L, I_R\}$ with resolution 1032×778. The standard cameras $\{L, R\}$ form a stereo pair S with a baseline of approximately $20cm$. The Mesa Imaging SwissRanger SR4000$^{\text{TM}}$ToF camera $\{T\}$, with a $10mm$ optics and horizontal field of view of 43.6^o acquires a 16-bit depth image D_T, with values in $[0, 5m]$, a 16-bit amplitude image A_T, and a confidence map C_T with integer values in $[0, 8]$. Data $\{A_T, D_T, C_T\}$ are framed with resolution 176×144. The ToF camera is positioned in between the standard cameras L and R.

The calibration step is performed via 70 acquisitions of a checkerboard with 28 suitable internal corners, and checkerside of $11cm$.

The compensation of radial and tangential distortions of images $\{A_T, D_T, C_T, I_L, I_R\}$, and the stereo rectification of images $\{I_L, I_R\}$ are performed by the OpenCV Library [17]. The systematic error in the depth measurement performed by T is compensated by a method of the Swissranger Library [19]. The final error obtained after the proposed calibration step, computed as average of the Euclidean distance values between the 3D coordinates of all the corresponding corners of the checkerboard estimated with the ToF ($\mathbf{P}_T^i$) and by the stereo system ($M_L * \mathbf{P}_S^i$):

$$\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{P}_T^i - M_L * \mathbf{P}_S^i||_2, \tag{17}$$

is of $0.7cm$. This is far better than the error obtained by standard calibration techniques, that is of about $2.5cm$, and also better than the one obtained by the method proposed in [7], that is of about $1.3cm$.

Some experiments, based on repeated depth measurements in situations with different values of C_T, were performed in order to correctly associate the confidence values of C_T in $[0, 8]$, to the relative standard deviation σ_w of the error in the depth measurement performed by T, obtaining the approximated values reported in Table 1. Some visual results are reported in Figure 7 and Figure 8.

C_T	0	1	2	3	4	5	6	7	8
$\sigma_w[cm]$	50	20	15	10	8	4	2	1	0.5

Table 1. Relationship between C_T and σ_w.

Figure 7. Images $\{I_L, I_R\}$ acquired by the stereo pair S.

5. Conclusions

This work proposes a novel probabilistic approach to ToF and stereo data fusion. The preliminary system calibration task has been satisfactory solved, exploiting a closed-form solution by means of the Horn's algorithm that fully exploits the features of the considered acquisition system.

After an adequate answer to system calibration, the work

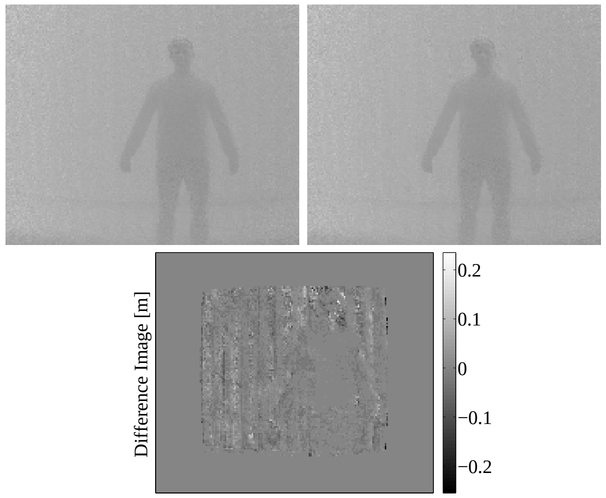

Figure 8. On top-left, raw depth map D_T acquired by the ToF before fusion algorithm application; on top-right, depth map $\hat{Z}$ obtained after the fusion algorithm application; on the bottom, difference between D_T and $\hat{Z}$.

derives a fusion algorithm in a probabilistic framework, underlining all the considered assumptions.

Experimental results on data coming from real and synthetic scenes show the effectiveness of the proposed algorithm for ToF and stereo data fusion.

The generation of high resolution depth information and the introduction of a more complex probability model, in order to apply a global optimization step, will be the subject of future work.

References

[1] C. Beder, B. Barzak, and R. Koch. A combined approach for estimating patchlets afrom pmd depth images and stereo intensity images. In *Proc. of DAGM Conf.*, 2007. 1

[2] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:359–374, 2001. 1

[3] N. Brusco, P. Zanuttigh, D. Taubman, and G. M. Cortelazzo. Distortion-sensitive synthesis of texture and geometry in interactive 3d visualization. In *Proc. of 3DPVT Conf.*, 2006. 4

[4] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Readings in Computer Vision: issues, problems, principles and paradigms*, 1:726–740, 1987. 3

[5] A. Frick, F. Kellner, B. Bartczak, and R. Koch. Generation of 3d-tv ldv-content with time-of-flight camera. In *Proc. of 3DTV Conf.*, 2009. 1

[6] A. Fusiello, V. Roberto, and E. Trucco. Symmetric stereo with multiple windowing. *International Journal of Pattern Recognition and Artificial Intelligence*, 14:1053–1066, 2000. 5

[7] V. Garro, C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. A novel interpolation scheme for range data with side information. In *Proc. of CVMP Conf.*, London, UK, November 2009. 7

[8] L. Guan, J.-S. Franco, and M. Pollefeys. 3d object reconstruction with heterogeneous sensor data. In *Proc. 3DPVT Conf.*, 2008. 1

[9] L. Guan and M. Pollefeys. A unified approach to calibrate a network of camcorders and tof cameras. In *Proc. of M2SFA208 Conf.*, 2008. 2

[10] S. A. Gudmundsson, H. Aanaes, and R. Larsen. Fusion of stereo vision and time of flight imaging for improved 3d estimation. *Int. J. Intell. Syst. Technol. Appl.*, 5:425–433, 2008. 1

[11] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 2, 3

[12] R. I. Hartley and P. Sturm. Triangulation. In *Proc. of ARPA Image Understanding Workshop*, 1994. 3

[13] J. Heikkila and O. Silven. A four-step camera calibration procedure with implicit image correction. In *Proc. of CVPR Conf.*, 1997. 2

[14] C. Hernindez Esteban, G. Vogiatzis, and R. Cipolla. Probabilistic visibility for multi-view stereo. In *Proc. of CVPR Conf.*, 2007. 1

[15] B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, 4:629–642, 1987. 3

[16] http://canesta.com/. 1

[17] http://opencv.willowgarage.com/wiki/. 7

[18] http://vision.middlebury.edu/stereo/. 6

[19] http://www.mesa imaging.ch. 1, 7

[20] T. Kahlmann and H. Ingensand. Calibration and development for increased accuracy of 3d range imaging cameras. *Journal of Applied Geodesy*, 2:1–11, 2008. 2, 4

[21] T. M. Kim, D. Chan, C. Theobald, and S. Thrun. Design and calibration of a multi-view tof sensor fusion system. In *Proc. of CVPR Conf.*, 2008. 2

[22] Y. M. Kim, C. Theobald, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun. Multi-view image and tof sensor fusion for dense 3d reconstruction. In *Proc. of 3DIM Conf.*, 2009. 1

[23] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, 2001. 1, 5

[24] I. Schiller, C. Beder, and R. Koch. Calibration of a pmd-camera using a planar calibration pattern together with a multi-camera setup. In *Proc.of ISPRS Conf.*, 2008. 2

[25] J. Sun, N. Zheng, and H. Shum. Stereo matching using belief propagation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25:787–800, 2003. 1, 5

[26] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1330–1334, 1998. 2, 3

[27] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *Proc. of CVPR Conf.*, 2008. 1