Probabilistic ToF and Stereo Data Fusion Based on Mixed Pixels Measurement Models

Carlo Dal Mutto, *Member, IEEE,* Pietro Zanuttigh, *Member, IEEE,* and Guido Maria Cortelazzo, *Senior Member, IEEE*

Abstract—This paper proposes a method for fusing data acquired by a ToF camera and a stereo pair based on a model for depth measurement by ToF cameras which accounts also for depth discontinuity artifacts due to the mixed pixel effect. Such model is exploited within both a ML and a MAP-MRF frameworks for ToF and stereo data fusion. The proposed MAP-MRF framework is characterized by site-dependent range values, a rather important feature since it can be used both to improve the accuracy and to decrease the computational complexity of standard MAP-MRF approaches. This paper, in order to optimize the site dependent global cost function characteristic of the proposed MAP-MRF approach, also introduces an extension to Loopy Belief Propagation which can be used in other contexts. Experimental data validate the proposed ToF measurements model and the effectiveness of the proposed fusion techniques.

Index Terms—ToF, Stereo, Data Fusion, MAP-MRF, Loopy Belief Propagation, Mixed Pixels

1 Introduction

High quality depth estimation is an extremely challenging problem for which many different approaches have been proposed. Among them stereo vision systems have attracted a lot of attention but they also suffer well known reliability issues in case of lacking texture and of repeated structures. Time-of-Flight (ToF) cameras have recently introduced reliable depth acquisition to the mass market but are also currently characterized by limited spatial resolution and by noisy depth measurements. Stereo vision systems and ToF cameras operate on completely different basis, therefore a proper fusion of their data can lead to more accurate measurements than those of each single system.

This paper proposes a data fusion framework based on accurate models of the measures for the two subsystems. It begins by introducing a novel error model for ToF cameras measurements which takes into account boundary-related issues so far either ignored or only partially considered in current state-of-the-art ToF models. In order to cope with the depth discontinuities artifacts typical of ToF systems, this paper also derives new inter-pixel and intra-pixel measurements models for ToF cameras. Stereo depth measurements are modeled by extending a classical method in order to account for scene segmentation. The derived measurement models are then fused within two different probabilistic approaches, a simpler Maximum Likelihood (ML) scheme considering each pixel independent and a more refined global Maximum-A-Posteriori (MAP-MRF) approach estimating the optimal depth distribution on the basis of a scene prior enforcing piece-wise smoothness. The proposed global probabilistic framework associates different lattice locations to different depth-value ranges according to the introduced measurement models. The novel measurement and optimization model allows to consider only the meaningful depth range at each location and to sample with an higher density since the considered regions are smaller. This permits to improve the estimated depth-map accuracy and to reduce the computational complexity of the optimization procedure. On the other side the optimization of the global cost function, typical of any MAP-MRF model, can not be performed with standard approaches and for this reason an extension of the classical Loopy-Belief-Propagation (LBP) scheme prompted by the need of accounting for the pixel-dependent depth ranges is proposed in this paper. Detailed experimental results confirm the ability of the proposed method to compensate for the well-known unreliability of ToF measurements near discontinuities and of stereo measurements in textureless areas and to produce very accurate depth maps.

The paper is organized as follows: Section 2 overviews the related literature; Section 3 considers the problem framed within the more general class of depth measurements from multiple devices; Section 4 and 5 describe the proposed ToF and stereo measurements model respectively; Section 6 introduces the adopted ML probabilistic method. Section 7 presents the more refined MAP-MRF method; Section 8 explains the variation of LBP needed for the global optimization; Section 9 presents the experimental validation and Section 10 finally draws the conclusions.

- *Department of Information Engineering, University of Padova, Italy. E-mail: dalmutto,zanuttigh,corte@dei.unipd.it*

2 Related Work

The first detailed description of matricial ToF range cameras can be found in [1] while more recent books and PhD theses [2], [3], [4], [5] address ToF technology, calibration and best usage for accurate 3D measurements. A characterization of the performances of ToF cameras can be found in [6], [7], [8]: the various error sources that influence range measurements are analyzed in [6] while a qualitative analysis on how scene reflectance influences the depth measurements is given in [7]. The first ToF camera error measurements model accounting for scene properties (*i.e.,* depth discontinuity and scene reflectance) is proposed in [9]. A formal model for ToF measurement errors is proposed in [10] while a more recent confidence estimation scheme for ToF data has been presented in [11]. The issue of flying pixels associated to depth discontinuities has been discussed in [12].

The idea of combining ToF sensors with standard cameras has inspired several recent works, a complete survey can be found in [13]. A first set of works proposes the combination of a ToF camera with a single color camera. In the earliest of such attempts [14] the authors adopt a Markov Random Field (MRF) approach. A method based on bilateral filtering is proposed in [15] and extended in [16] where the input depth map is used in order to build a 3D depth probability volume (cost volume). The approach of [17] and [18] explicitly imposes the alignment between range and color discontinuities to interpolate the depth data and finally an approach based on geodesic paths from low resolution samples is proposed in [19].

The setup made by a ToF camera and a stereo pair attracted considerable attention, because the two subsystems have complementary characteristics and both of them can independently produce depth data. In [20] the depth maps acquired separately by the ToF and the stereo pair are averaged. In [7] the ToF depth data is firstly reprojected on the reference image of the stereo pair, then interpolated and finally used as initialization for a dynamic programming stereo vision algorithm. In [9] the final depth-map is recovered from the ToF and the stereo measurements by a ML local optimization. The main limitations of this probabilistic method are the resolution of the final depth-map (equal to the one of the ToF) and the lack of a global optimization step. The method proposed in [21] is based on a MAP-MRF Bayesian formulation inside which a belief propagation algorithm optimizes a global energy function. A temporal extension of this method is proposed in [22], and an automatic way to set the weights of the ToF and stereo measurements is presented in [23]. Notice that different stereo vision cost functions can be fitted inside ToF and stereo vision fusion frameworks, e.g., approaches based on explicit or implicit segmentation [24] or approaches exploiting adaptive support weights [23]. A relevant issue due to the limited resolution of available ToF sensors is that pixels close to edges can include measurements coming from surfaces at different distances (i.e, the *mixed pixels* problem). An approach for the separation of the multiple contributions in mixed pixels has been presented in [25], while the ToF and stereo fusion approach of [26] accounts for this issue. The mixed pixels problem is carefully modelled and handled in the proposed work. Another recent method [27] uses a variational approach in order to combine the results of the two devices. In [28] an extension of the Locally Consistent (LC) approach, originally developed for stereo vision measurements, is applied to stereo and ToF data fusion. Other setups were also considered in the literature, such as four color cameras in [29] and multiple ToF and color cameras in [30].

As far as the MAP-MRF solution method, let's recall that classical optimization approaches for the global energy functions are: Loopy Belief Propagation (LBP) [31], Graph Cuts (GC) [32], Iterated Conditional Modes (ICM) [33], Tree-Reweighted Message Passing (TRW) [34]. Comprehensive analysis and a comparisons of such algorithms can be found in [35] and [36]. Since usually these methods are used in problems where a global energy function is defined for a finite set of variables (sites) taking discrete values (site-wise uniform), they can not be directly applied to the optimization of the energy function obtained in this work. For this reason Section 8 proposes an extension of LBP suited to the considered optimization problem.

3 Problem Definition

The computation of an estimate $\hat{Z}$ of the actual scene depth-map Z from ToF and stereo data can be performed within a probabilistic framework as the solution of a Maximum-A-Posteriori (MAP) problem

$$\hat{Z} = \arg\max_{Z \in \mathcal{Z}} P(Z|I_1, ..., I_N) \tag{1}$$

where $P(Z|I_1, ..., I_N)$ is the posterior probability of the scene depth-map Z, given the acquired data $I_1, ..., I_N$. From Bayes rule and by assuming independent the measurement errors of the various sensors, Equation (1) can be rewritten as

$$\hat{Z} = \arg\max_{Z \in \mathcal{Z}} P(I_1|Z)...P(I_N|Z)P(Z) \tag{2}$$

where $P(I_n|Z), n = 1, ..., N$ are the likelihoods of the single device measurements given the scene depth Z and $P(Z)$ is the scene depth prior probability. The independence hypothesis, rather common in the literature [9], [21], [22], [23], seems well grounded since the two devices exploit completely different measurement principles and have different error sources. This work introduces a novel approach for both the construction of the likelihood of the single device measurements and for the maximization algorithm. With respect to likelihood construction, the proposed approach extends the formal ToF cameras model described in [1], [4], [10] in order to remove its limitations near depth discontinuities. The adoption of such a formal model for the likelihoods is a significant improvement over previous works [21], [22], [23]. With respect to maximization algorithms the adopted maximization method based on a modified version of Loopy-Belief-Propagation (LBP) exploits the measurement models of the quantities to be optimized and is more efficient than the standard LBP adopted in other approaches [21], [22], [23]. This is due to the use of a site-dependent label space that greatly reduces the number of required operations and allows to use an optimized and denser sampling of the depth values.

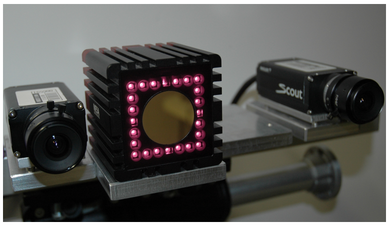


Fig. 1: Considered acquisition system made by a ToF camera T and a stereo vision system $S \triangleq \{L, R\}$.

The considered acquisition devices are a ToF camera T and a stereo vision system S made by a pair of color cameras L and R. In the setup used for the experimental results (Fig. 1) the ToF camera T is placed between the two cameras L and R but the proposed fusion method is not confined to such a geometric configuration.

Each of the 3 cameras has its own camera coordinate system (CCS) as shown in Fig. 2. Depth estimation requires to refer the acquired quantities to a unique CCS and the trinocular setup has been calibrated by the method of [9]. ToF cameras acquire an amplitude image A_T, an intensity image B_T and a depth-map Z_T (Fig. 3), all defined on lattice Λ_T associated to the ToF CCS. Such data will be denoted as $I_T \triangleq \{A_T, B_T, Z_T\}$. The data acquired by the two cameras L and R are instead synchronized pairs of color images denoted as I_L and I_R, defined on the lattices Λ_L and Λ_R. The data acquired by S are denoted as $I_S \triangleq \{I_L, I_R\}$ and use the CCS of L as reference.

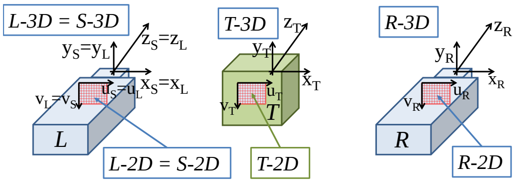


Fig. 2: CCSs (3D and 2D) associated to the various sensors of the acquisition system.

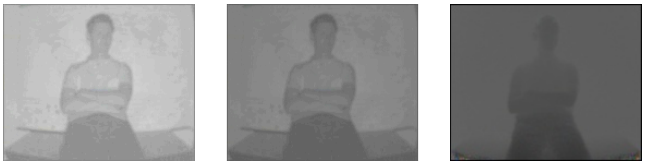


Fig. 3: Data acquired by T: A_T (left), B_T (center) and Z_T (right). (*Images A_T and B_T were processed to increase printing visibility*)

From the above notation, Equation (2) for the considered setup can be rewritten as

$$\hat{Z} = \arg\max_{Z \in \mathcal{Z}} P(I_T|Z)P(I_S|Z)P(Z) \tag{3}$$

where $P(I_T|Z)$ and $P(I_S|Z)$ are the likelihood of the ToF measurements and of the stereo measurements given the scene depth respectively, and $P(Z)$ is the scene depth prior probability. The various components of Equation (3) are analyzed and described in the following sections.

If random field $\mathcal{Z}$ is assumed pixel-wise independent with uniform distribution within the minimum and the maximum measurable depth at each pixel, Equation (3) simplifies into

$$\hat{Z} = \arg\max_{Z \in \mathcal{Z}} P(I_T|Z)P(I_S|Z) \tag{4}$$

which is a ML formulation of the fusion problem [9].

4 ToF Likelihood

As reported in [4] and [10], the distribution of the depth acquisition noise of a ToF pixel p_i can be approximated by a Gaussian with standard deviation

$$\sigma_\rho = \frac{c}{4\pi f_{mod}\sqrt{2}} \frac{\sqrt{B}}{A} \tag{5}$$

where f_{mod} is the IR frequency of the signal sent by the ToF emitters, A is the value of the amplitude image A_T at pixel p_i and B_T is the intensity image at pixel p_i. The standard deviation given by Equation (5) determines the precision (repeatability) of the distance measurement and it is directly related to the ToF operational parameters f_{mod}, A and B. The noise model of Equation (5), although well-known in the fields of ToF metrology [10], has never been used in computer vision applications yet. One of the main limitations of (5) is that it does not take into account the finite size of the ToF sensor pixels. In order to account for this issue we use Equation (5) as a starting point for the derivation of a more general likelihood of the ToF depth measurements $P(I_T|Z)$.

Let us consider a finite size pixel $p_i \in \Lambda_T$ which acquires information relative to a finite size scene area (as shown in Fig. 4). If the finite scene area is flat, then the first order Taylor approximation of the scene area with a fronto-parallel plane is realistic and the noise model of Equation (5) holds. However, if

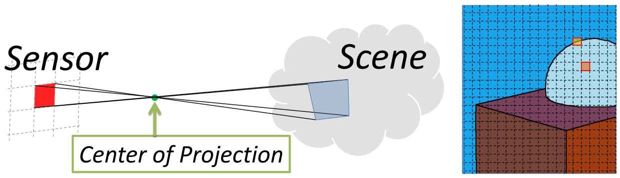


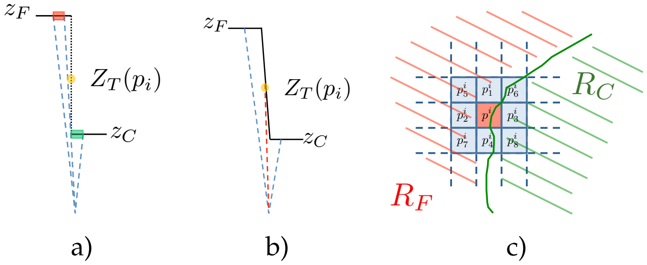
Fig. 4: Formation of a finite size sensor pixel $p_i \in \Lambda_T$ relative to a finite size scene area.

the scene area falling onto p_i corresponds to a depth discontinuity, its first order Taylor approximation is not correct, and the noise model of Equation (5) does not hold. Pixels characterized by this type of depth-estimation error are often addressed as "mixed pixels". In particular, let us consider the case of a scene area projected to p_i made by two different regions R_C at depth z_C (closest region) and R_F at depth z_F (furthest region). The depth measured at the pixel p_i in this case would be:

$$\tilde{z}_i = \alpha z_C + (1-\alpha) z_F \qquad (6)$$

where α is the percentage of scene area associated to R_C and $(1-\alpha)$ that associated to R_F. In order to obtain a likelihood of the ToF depth measurements z_i at p_i it is worth distinguishing between the following two situations shown in Fig. 5:

1) if R_C and R_F belong to two surfaces at different depth, the actual depth might be either z_C or z_F, and not in between them (Fig. 5.a); a situation we name *disconnected discontinuity*.
2) if R_C and R_F belong to the same surface the actual depth might be either close to z_C or z_F or somewhere in between them (Fig. 5.b); a situation we name *connected discontinuity*.



a) b) c)

Fig. 5: Discontinuity types: a) disconnected discontinuity; b) connected discontinuity; c) the discontinuity between R_C and R_F crosses the area associated to p_i, p_4^i, p_6^i: points $p_1^i, p_2^i, p_5^i, p_7^i$ pertain to scene region R_F while points p_3^i, p_8^i to scene region R_C.

Since it cannot be known a priori which situation occurs, the model must account for both scenarios. To this end it is worth observing that if p_i is relative to a scene area crossed by a discontinuity between R_C and R_F, some of the points p_j^i in the 8-neighbourhood $\mathcal{N}(p_i)$ of p_i measure distance z_C and some others distance z_F (Fig. 5.c). Hence the likelihood term for p_i should account for the fact that, if p_i is across a discontinuity, its actual value might either be around the depth measured by p_i itself (connected discontinuity) or around the depth measured by some of its 8-neighbours p_j^i (connected or disconnected discontinuity). The contributions of neighbouring pixels can be fused by a classical image correlation model [37], obtaining the following expression of the ToF likelihood at pixel p_i

$$\begin{aligned} P(I_T|Z) \propto & \frac{1}{\sigma_p(p_i)} \exp - \left(\frac{z - z_i}{\sigma_p(p_i)} \right)^2 \\ & + e^{-1} \sum_{j=1}^{4} \frac{1}{\sigma_p(p_j^i)} \exp - \left(\frac{z - z_j^i}{\sigma_p(p_j^i)} \right)^2 \\ & + e^{-2} \sum_{j=5}^{8} \frac{1}{\sigma_p(p_j^i)} \exp - \left(\frac{z - z_j^i}{\sigma_p(p_j^i)} \right)^2 \end{aligned} \qquad (7)$$

with

$$\sigma_p(p) = \frac{c}{4\pi f_{mod} \sqrt{2}} \frac{\sqrt{B(p)}}{A(p)}$$

where $z_j^i = z(p_j^i)$, and σ_i and σ_{ij} are the standard deviations of the depth measurements for the points p_i and p_j^i respectively, obtained from Equation (5).

Before moving forward, let us analyse what Equation (7) means and why the proposed model for the ToF likelihood is adequate. In case of no depth discontinuity, the various Gaussian contributions of Equation (7) have similar mean, therefore the ToF likelihood becomes very similar to a Gaussian with variance given by Equation (5). Hence this model, although more general, reduces to the one of Equation (5) when the assumptions for (5) hold. In presence of a depth discontinuity, Equation (7) produces a mixture-of-Gaussians model with Gaussians centered at z_C, z_F and at the measured values for the pixels crossing the discontinuity. This is likely to assign high probability to depth values around z_C and z_F (corresponding to the disconnected discontinuity) and around the measured depth z_i (corresponding to the connected discontinuity case), in agreement with the observations about the two cases of Fig. 5.

Since all terms of (7) are Gaussians, the Chebychev theorem associates nice properties based on the *useful interval* concept to ToF likelihood (7). It is a fact that, given certain depth measurements for pixel p_i and its neighbourhood, the actual depth value z^* is likely not to be very different from at least one of them. This concept can be formalized by noting that likelihood (7) is a mixture of Gaussians. For a Gaussian distribution the concept of *useful interval* ensures that the actual value of the measured quantity belongs to interval $[\mu - 3\sigma, \mu + 3\sigma]$ with probability 0.997 where μ is the mean and σ is the standard deviation of the Gaussian distribution. In the case of a mixture of Gaussians the useful interval can be defined as $[\mu_{min} - 3\sigma_{min}, \mu_{max} + 3\sigma_{max}]$, where μ_{min} and σ_{min} are the mean and the standard deviation of the Gaussian in the mixture with minimum mean value, while μ_{max} and σ_{max} are the mean and the standard deviation of the Gaussian in the mixture with maximum mean value. In the depth measurements case, μ_{min} and σ_{min} can be named μ_C and σ_C (where C stays for "close") and μ_{max} and σ_{max} can be named μ_F and σ_F (where F stays for "far"). Since all depth values outside the useful interval for pixel p_i can simply be ignored, this concept allows to prevent useless computations in the fusion algorithm as pictorially shown in Fig. 6 and explained in the following sections. Furthermore the depth values inside the *useful interval* can be sampled at high resolution. From a high-level point of view it is possible to say that the ToF likelihood model (7) accounts not only for classical ToF measurement error distribution theory but also for the matricial nature of ToF cameras sensors. The only issue not accounted by this model is the inter-reflection (multi-path) error of ToF cameras [38], [39]. However this issue is very complex to model and typically leads to artifacts around corners that can be removed in the global optimization step. Indeed the proposed model leads to accurate depth estimates as the experiments of Section 9 show.

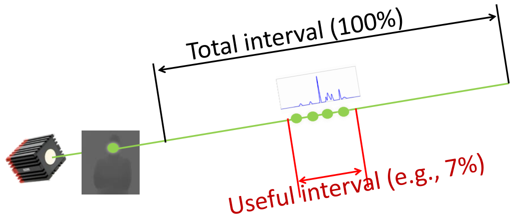


Fig. 6: The concept of useful interval allows for a reduction of the number of operations, as it will be shown in Section 9.

5 Stereo Likelihood

The literature offers a number of different approaches for modeling the likelihoods of the depth estimates z_i obtained by a calibrated and rectified stereo vision system [40], [41]. Let us denote as $p_i^L \in \Lambda_L, p_i^R \in \Lambda_R$ a pair of conjugate points (i.e., they refer to a unique 3D point P_i in the scene) with image coordinates $\mathbf{p}_i^L = [u_i^L, v_i^L]^T$ and $\mathbf{p}_i^R = [u_i^R, v_i^R]^T$. The likelihood of stereo data I_S given depth distribution Z_i can be obtained by considering multiple hypothesis $z_{i,n}, n = 1, ..., N$ for depth z_i and computing a likelihood value for each hypotheses. By taking advantage of classical stereo schemes the likelihood distribution $P(I_S|Z(P_{i,n}))$ for hypotheses $z_{i,n}, n = 1, ..., N$ can be practically computed as follows:

1) for each depth hypothesis $z_{i,n}, n = 1, ..., N$ compute the 3D coordinates of the corresponding 3D point $P_{i,n}$
2) project $P_{i,n}$ into the 2D points $p_{i,n}^L \in \Lambda_L, p_{i,n}^R \in \Lambda_R$ with image coordinates $\mathbf{p}_{i,n}^L = [u_{i,n}^L, v_{i,n}^L]^T$ and $\mathbf{p}_{i,n}^R = [u_{i,n}^R, v_{i,n}^R]^T$ respectively.
3) consider a window $W_{i,n}^L$ centered around $p_{i,n}^L$ and a window $W_{i,n}^R$ centered around $p_{i,n}^R$
4) evaluate the similarity (hence the likelihood) between $I_L(W_i^L)$ and $I_R(W_i^R)$

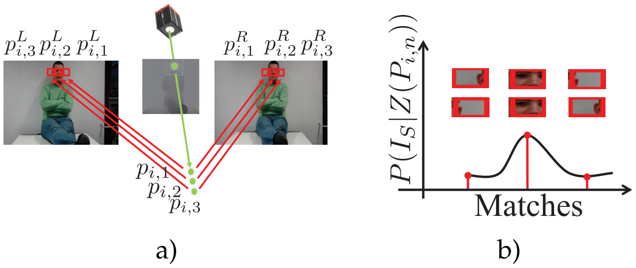

$p_{i,3}^L$ $p_{i,2}^L$ $p_{i,1}^L$
$p_{i,1}^R$ $p_{i,2}^R$ $p_{i,3}^R$
$p_{i,1}$
$p_{i,2}$
$p_{i,3}$
$P(I_S|Z(P_{i,n}))$
Matches
a)
b)

Fig. 7: Stereo likelihood computation: a) The 3D points sampled from the useful interval are re-projected onto the two stereo images; b) the stereo likelihood is computed by matching the windows centered on conjugate pairs.

The procedure above is pictorially shown in Fig. 7. The actual computation of the similarity between $I_L(W_i^L)$ and $I_R(W_i^R)$ can be obtained by the classical method of [40] (for a comprehensive presentation of cost matching and aggregation procedures refer to [42]). Recent advancements in stereo cost aggregation procedures show that image segmentation clues can improve the results [43]. In light of this we improved the cost function by accounting also for segmentation clues. Differently than [40] for the matching cost computation we adopt Truncated Absolute Difference (TAD) as in [43] instead of the Birchfield-Tomasi method [44]. Let us call S_L and S_R the segmentations of images I_L and I_R obtained by a suitable image segmentation method such as [45] and call $W_{i,n}^L$ and $W_{i,n}^R$ the rectangular windows of size $W_S = (2H_W + 1) \times (2W_W + 1)$, centered at $p_{i,n}^L$ and $p_{i,n}^R$ respectively. The likelihood of stereo measurements $P(I_S|Z(P_{i,n}))$ can be computed as

$$P(I_S|Z(P_{i,n})) = \frac{\exp - \frac{\mathcal{C}(\mathbf{p}_{i,n}^L, \mathbf{p}_{i,n}^R)}{\sigma_I^2}}{\sum_{k=1}^{N} \exp - \frac{\mathcal{C}(\mathbf{p}_{i,k}^L, \mathbf{p}_{i,k}^R)}{\sigma_I^2}} \tag{8}$$

where $\mathcal{C}(\mathbf{p}_{i,n}^L, \mathbf{p}_{i,n}^R)$ (and similarly $\mathcal{C}(\mathbf{p}_{i,k}^L, \mathbf{p}_{i,k}^R)$) is defined as

$$\begin{aligned} \mathcal{C}(\mathbf{p}_{i,n}^L, \mathbf{p}_{i,n}^R) \quad = \quad & \tfrac{1}{W_S} * \textstyle\sum_{u \in [-W_W, W_W]} \sum_{v \in [-H_W, H_W]} \\ & \{ \quad w(\mathbf{p}_{i,n}^L, [u_{i,n}^L - u, v_{i,n}^L - v]^T) * \\ & \qquad w(\mathbf{p}_{i,n}^R, [u_{i,n}^R - u, v_{i,n}^R - v]^T) * \\ & \qquad \min(I_L(\mathbf{p}_{i,n}^L) \ominus I_R(\mathbf{p}_{i,n}^R), T_h) \quad \} \end{aligned} \tag{9}$$

where T_h is the TAD threshold parameter, $\ominus$ is an operator defined as the geometric mean of the three intra-channel difference between I_L and I_R and $w(\mathbf{p},\mathbf{q})$, with $\mathbf{p}=[u_p,v_p]^T, \mathbf{q}=[u_q,v_q]^T$ is the aggregation weight of [43], namely

$$w(\mathbf{p},\mathbf{q}) \triangleq \begin{cases} 1 & if\; S(\mathbf{p}) == S(\mathbf{q}) \\ \\ I(\mathbf{p}) \ominus I(\mathbf{q}) & otherwise \end{cases} \qquad (10)$$

where S is the segmented image to which $\mathbf{p}$ and $\mathbf{q}$ belong (either S_L or S_R) and I is the acquired color image (either I_L or I_R).

6 Structure of the depth domain and the ML framework

There are two different natural choices for the lattice Λ_Z of the output depth-map and of the scene depth prior probability $P(Z)$, namely $\Lambda_Z \equiv \Lambda_S$ (e.g., [7], [15], [20], [21], [22]) or $\Lambda_Z \equiv \Lambda_T$ (e.g., [9], [46]).

Choice $\Lambda_Z \equiv \Lambda_S$ allows to adopt standard stereo likelihood expressions [40], [41] while the ToF likelihood can only be expressed in heuristic ways [21], [22]. The other advantage is that the output resolution is the one of the stereo pair which is typically the highest. Another issue is the very high computational complexity (approximations can be used to reduce it [21], [22] but with an impact on the accuracy).

Choice $\Lambda_Z \equiv \Lambda_T$ allows to exploit the previously introduced formal model for both the ToF and the stereo likelihoods, leading to a computationally lighter framework but the output resolution is the one of the ToF sensor which is typically the lowest.

In order to exploit the above presented formal models for ToF and stereo likelihoods and to obtain an high resolution output we adopt as Λ_Z a version of Λ_T interpolated by L times denoted as Λ_T^L, giving an estimated depth-map $\hat{Z}$ with resolution cardinality $L\|\Lambda_T\| \approx \|\Lambda_L\|$ (*e.g.*, $L=2,4,6$).

The choice $\Lambda_Z = \Lambda_T^L$ requires to up-sample the ToF likelihood $P(I_T|Z)$ from the lattice Λ_T to Λ_T^L. Notice that interpolation of image data would create flying pixels in edge regions that would strongly affect the computation of the ToF data term, for this reason the interpolation is performed on the probability densities and not on the depth or image data. Since the meaning of spatial-interpolation of probability densities is not the same of spatial-interpolation of images or depth-maps, we preferred to adopt a "bilinear interpolation" model, which naturally relates to standard correlation models for 2D random fields. The bilinear interpolation of the likelihood probability from Λ_T to Λ_T^L gives an up-sampled likelihood probability distribution of the measurements performed by the ToF camera denoted as $P^L(I_T|Z)$. For simplicity, the superscript L will be omitted in the sequel.

Any specific realization Z of random field $\mathcal{Z}$ is characterized by N_i possible values for each pixel: $Z(p_i) = z_i^{n_i}, n_i \in [1,...,N_i]$. If depth scene prior $P(Z)$ is considered both uniform and pixel by pixel independent, and the likelihood probabilities of the ToF $P(I_T|Z)$ and of the stereo $P(I_S|Z)$ are also pixel by pixel independent, by assuming independent the measurements of the two devices, the posterior probability distribution can be expressed as:

$$P(\mathcal{Z}(p_i)=z_i^{n_i}|I_S,I_T)=P(I_S|\mathcal{Z}(p_i)=z_i^{n_i})P(I_T|\mathcal{Z}(p_i)=z_i^{n_i}) \qquad (11)$$

The use of (11) as the argument of the optimization of Equation (2) corresponds to an ML framework similar to [9] with the difference that in [9] $\Lambda_Z = \Lambda_T$ while in this work $\Lambda_Z = \Lambda_T^L$. The ToF and stereo likelihoods of this work are also different from those of [9].

7 MAP-MRF framework

If the spatial relationships of the scene-pixels depths are taken into account, $\mathcal{Z}$ can be assumed to be a Markov Random Field (MRF) and $P(\mathcal{Z}=Z)$ can be expressed as $P(\mathcal{Z}(p_i)=z_i^{n_i}|\mathcal{Z}(p_j)=z_j^{n_j})$ where $p_j \in \mathcal{N}(p_i)$, with $\mathcal{N}(p_i)$ the neighborhood of p_i, $p_i \in \Lambda_Z$,$n_i \in [1,...,N_i]$ and $n_j \in [1,...,N_j]$.

Since $\mathcal{Z}$ is a MRF, the posterior probability distribution maintains the Markovian property and can be expressed as

$$P(\mathcal{Z}(p_i)=z_i^{n_i}|I_S,I_T)=$$
$$P(I_S|\mathcal{Z}(p_i)=z_i^{n_i})P(I_T|\mathcal{Z}(p_i)=z_i^{n_i})*$$
$$*P(\mathcal{Z}(p_i)=z_i^{n_i}|\{\mathcal{Z}(p_j)=z_j^{n_j}, j:p_j\in\mathcal{N}(p_i)\}) \qquad (12)$$

where $P(I_S|\mathcal{Z}(p_i)=z_i^{n_i})P(I_T|\mathcal{Z}(p_i)=z_i^{n_i})$ is the *data term* denoted as $P^{(data)}$ and $P(\mathcal{Z}(p_i)=z_i^{n_i}|\{\mathcal{Z}(p_j)=z_j^{n_j}, j:p_j\in\mathcal{N}(p_i)\})$ is the *smoothness term* denoted as $P^{(smooth)}$.

The data term ensures that the probability of the depth distribution given the measurements reflects the measurements themselves, and the smoothness term imposes the piecewise-smoothness of the estimated scene surface. For the smoothness term we adopted the classical truncated quadratic model [47]. The use of (12) as the argument of optimization (2) corresponds to a MAP-MRF framework. In order to summarize the proposed probabilistic framework, Fig. 8 shows a flowchart of the method showing both the ML and the MAP-MRF approaches.

8 Loopy-Belief-Propagation optimization

Classical optimization methods for global energy functions obtained from a MAP-MRF Bayesian approach are: Loopy Belief Propagation (LBP) [31], Graph Cuts (GC) [32], Iterated Conditional Modes (ICM) [33] and Tree-Reweighted Message Passing (TRW) [34]. See [35] for a comprehensive analysis and a comparison of such algorithms. These methods are usually adopted in problems where a global energy function is defined for a finite set of variables (sites)

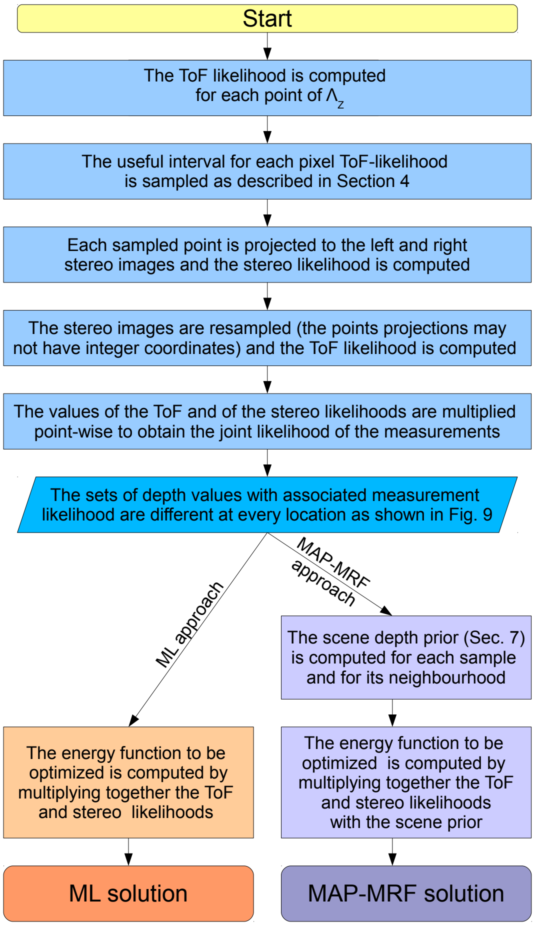

Start
The ToF likelihood is computed for each point of Λz
The useful interval for each pixel ToF-likelihood is sampled as described in Section 4
Each sampled point is projected to the left and right stereo images and the stereo likelihood is computed
The stereo images are resampled (the points projections may not have integer coordinates) and the ToF likelihood is computed
The values of the ToF and of the stereo likelihoods are multiplied point-wise to obtain the joint likelihood of the measurements
The sets of depth values with associated measurement likelihood are different at every location as shown in Fig. 9
ML approach
MAP-MRF approach
The scene depth prior (Sec. 7) is computed for each sample and for its neighbourhood
The energy function to be optimized is computed by multiplying together the ToF and stereo likelihoods
The energy function to be optimized is computed by multiplying together the ToF and stereo likelihoods with the scene prior
ML solution
MAP-MRF solution

Fig. 8: Flowchart of the proposed ML and MAP-MRF fusion frameworks.

with discrete values. The proposed energy function on the contrary is computed on a different number of samples corresponding to potentially different distances at each pixel and at each of its neighbors. This characteristic makes the considered optimization problem non-standard.

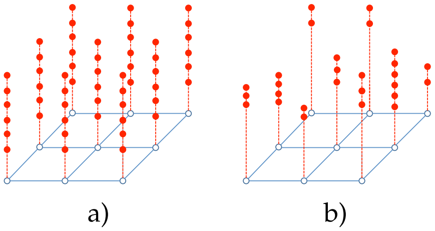


Fig. 9: a) In all the previous approaches each pixel of Λ_Z is associated to the same set of possible depth values; b) In the proposed approach each pixel of Λ_Z is associated to its own set of depth values different from the ones of the other pixels.

LBP is particularly suited to the considered situation since the message-passing structure of LBP does not impose any range-sampling condition on the messages exchanged between adjacent nodes. This fact has been exploited for other computer vision tasks (e.g., an application to stereo vision has been proposed in [48]), but has never been exploited for the considered ToF and stereo data fusion problem. Moreover, in [48] the range values are just a subset of a finite set of range values (e.g., a set of integer disparity values in $[0, 255]$), while in our case the set of range values for each pixel does not have the same limitation. Not only the optimization problem of Equation (12) can be solved by LBP, but also the effectiveness of the messages exchanged across neighboring sites can be improved by considering at each site only the meaningful range portion associated to the useful interval. In particular there are not exchanged messages associated to values not close to any measurement (in terms of range), and within a given probability the true depth value is guaranteed to belong to the range considered for message passing. Let us briefly recall that LBP optimizes a global energy function (e.g., Equation (12)) by marginalizing the posterior probabilities at each pixel on Λ_Z. The marginalization is iteratively performed by message (belief) passing between neighbor points [31], [49]. As shown in Fig. 10, the range values or depth samples associated to pixel $p_i \in \Lambda_Z$ are $\{z_i^{n_i}, n_i \in [1, N_i]\}$ and generally differ from the range values associated to $p_j \in \mathcal{N}(p_i)$ denoted as $\{z_j^{n_j}, n_j \in [1, N_j]\}$. The message that $p_j \in \mathcal{N}(p_i)$ sends to points $p_i \in \Lambda_Z$ for distance $z_i^{n_i}$ at the $(t+1)^{th}$ iteration, similarly to the classical LBP messages, is defined as:

$$m^{t+1}_{p_j \to p_i}(z_i^{n_i}) = \sum_{n_j=1}^{N_j} P^{(data)}(z_j^{n_j}) P^{(smooth)}(z_i^{n_i}, z_j^{n_j}) \prod_{l:p_l \in \mathcal{N}(p_j)-\{p_i\}} m^{t}_{p_l \to p_j}(z_j^{n_j}) \qquad (13)$$

It is worth noting how from the first sum of the RHS of (13) each depth sample $z_i^{n_i}$ receives exactly the same number of contributions from each neighboring pixel (as shown in Fig. 10). Namely each value $z_i^{n_i}, n_i \in [1, N_i]$ receives N_j contributions from the N_j range values $z_j^{n_j}$ associated to p_j. This fact is fundamental for the exploitability of LBP in the solution of the considered optimization problem. All messages are initialized at 1 before the first iteration: $m^0_{p_j \to p_i}(z_i^{n_i}) = 1, \forall p_j \in \Lambda_Z, \forall p_j \in \mathcal{N}_1(p_i), \forall n_i \in [1, ..., N_i]$ and the adopted message updating rule is synchronous. Let us recall that the goal of LBP is the marginalization of the posterior probability $\hat{P}_i(z_i^{n_i})$ of depth samples $z_i^1, ..., z_i^{N_i}$ at each site $p_i \in \Lambda_Z$. The maximization becomes a winner-takes-all algorithm [49] on the marginalized posterior probability $\hat{P}_i(z_i^{n_i})$ and the final expression for the marginal posterior probability $\hat{P}_i(z_i^{n_i})$ is:

$$\hat{P}_i(z_i^{n_i}) = \frac{1}{\mathcal{Z}} P^{(data)}(z_i^{n_i}) \prod_{j:p_j \in \mathcal{N}(p_i)} m^{\infty}_{p_j \to p_i}(z_i^{n_i}) \qquad (14)$$

where $m^{\infty}_{p_j \to p_i}(z_i^{n_i})$ is the value of the message at the last considered iteration of LBP.

A formal proof of the convergence and of the effectiveness of the modified LBP algorithm is not given as in the standard LBP case. Nevertheless the

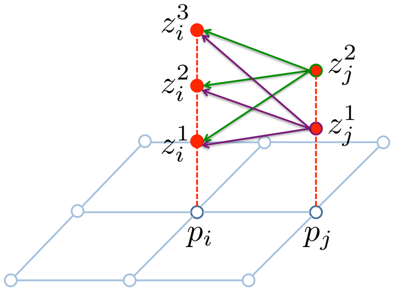


Fig. 10: Example of the structure of the messages exchanged between $p_i \in \Lambda_Z$ and $p_j \in \mathcal{N}(p_i)$ with $N_i = 3$ range (depth) samples for p_i and $N_j = 2$ for p_j. The messages relative to z_j^1 are in purple and the messages relative to z_j^2 in green. Notice how each range value $z_i^{n_i}$ receives exactly $N_j = 2$ messages.

experimental results clearly support the suitability of our modified LBP for maximizing posterior probability (12). The intuition is that only meaningful depth values affect message passing, making the messaging strategy more effective and avoiding sending noisy messages across neighboring sites. The possibility of considering different labels at different sites is a great advantage because it allows to narrow the labeling problems focus, with consequent computation reduction and accuracy increase (Fig. 11).

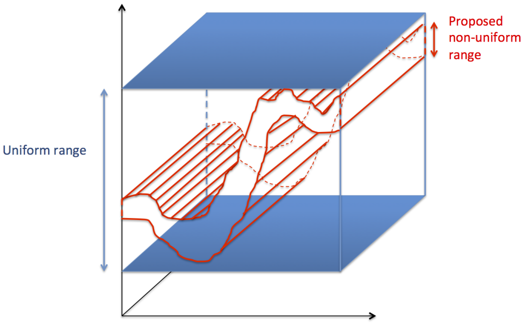


Fig. 11: Pictorial representation of the difference between classical labeling problems characterized by uniform labels at different sites and the considered labeling problem characterized by different labels at different sites.

Finally let us remark how the considered method exploits the formal ToF measurements model in order to decide which portion of the range-space is worth to consider. The ToF likelihood, the stereo likelihood and the depth prior are computed only on a sampled version of such a range-space portion. The labeling problem defined in this way focuses only on the relevant part of the range space and it can be solved both efficiently and accurately by the described optimization technique.

In order to fully understand the benefit of the proposed algorithm, let us analyze its computational complexity. The computational complexity of LBP is dominated by the computation of the messages to be passed (Equation (13)), which are performed at each site and for each iteration of the LBP algorithm. In particular, without treating the sites at the lattice border as a special case, at each site p_i in the lattice and for each range value $z_i^{n_i}$ there is a message coming from each range value $z_j^{n_j}$ of each site p_j in the neighborhood of p_j. Assuming that the considered neighborhood contains $|\mathcal{N}|$ sites and that the previous iteration messages as well as the data and the smoothness probabilities are pre-computed, a message computation is made by $|\mathcal{N}|$ multiplications and $|\mathcal{N}+1|$ memory accesses. Therefore the total number of multiplications for a single iteration of LBP (again with an approximation for the border pixels) is

$$\sum_{p_i \in \Lambda_Z} N_i \sum_{p_j \in \mathcal{N}(p_i)} N_j |\mathcal{N}| \tag{15}$$

which by assuming $N_i \approx N_j$ can be approximated as

$$|\mathcal{N}|^2 \bar{N}^2 |\Lambda_Z| \tag{16}$$

where $\bar{N}$ is the average number of range values in the lattice and $|\Lambda_Z|$ is the number of pixels in the lattice. Note that this is not an approximation if all the sites have the same number of range values and it is a very good approximation for scene areas characterized by similar ranges, *i.e.*, areas not in proximity of depth discontinuities. Empirically such sites constitute the great majority of the sites in Λ_Z. Similarly, the number of memory accesses can be approximated as $|\mathcal{N}+1|^2\bar{N}^2|\Lambda_Z|$. Therefore in L iterations of LBP the dominant term of multiplications is $L|\mathcal{N}|^2\bar{N}^2|\Lambda_Z|$ and the dominant term of memory accesses is $L|\mathcal{N}+1|^2\bar{N}^2|\Lambda_Z|$. Notice that in the implementation used for experimental results memory has been entirely pre-allocated. In order to improve the performances, it is possible to pre-compute the number of range values for each site and consequently pre-allocate the memory necessary to account for the maximum of such range values.

9 Experimental Results

In order to asses the quality of the proposed fusion frameworks for data acquired by a ToF camera and a stereo pair, we considered a setup made by two standard BASLER scA1000$^{\text{TM}}$RGB cameras $\{L, R\}$ and a MESA SR4000 ToF camera $\{T\}$. The color cameras acquire two 1032×778 RGB images $\{I_L, I_R\}$ forming a stereo pair with a baseline of $170[mm]$. The SR4000 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 $\{A_T, D_T, C_T\}$ have resolution 176×144. The ToF camera is positioned in between L and R as shown in Fig. 1 and the three cameras are hardware synchronized.

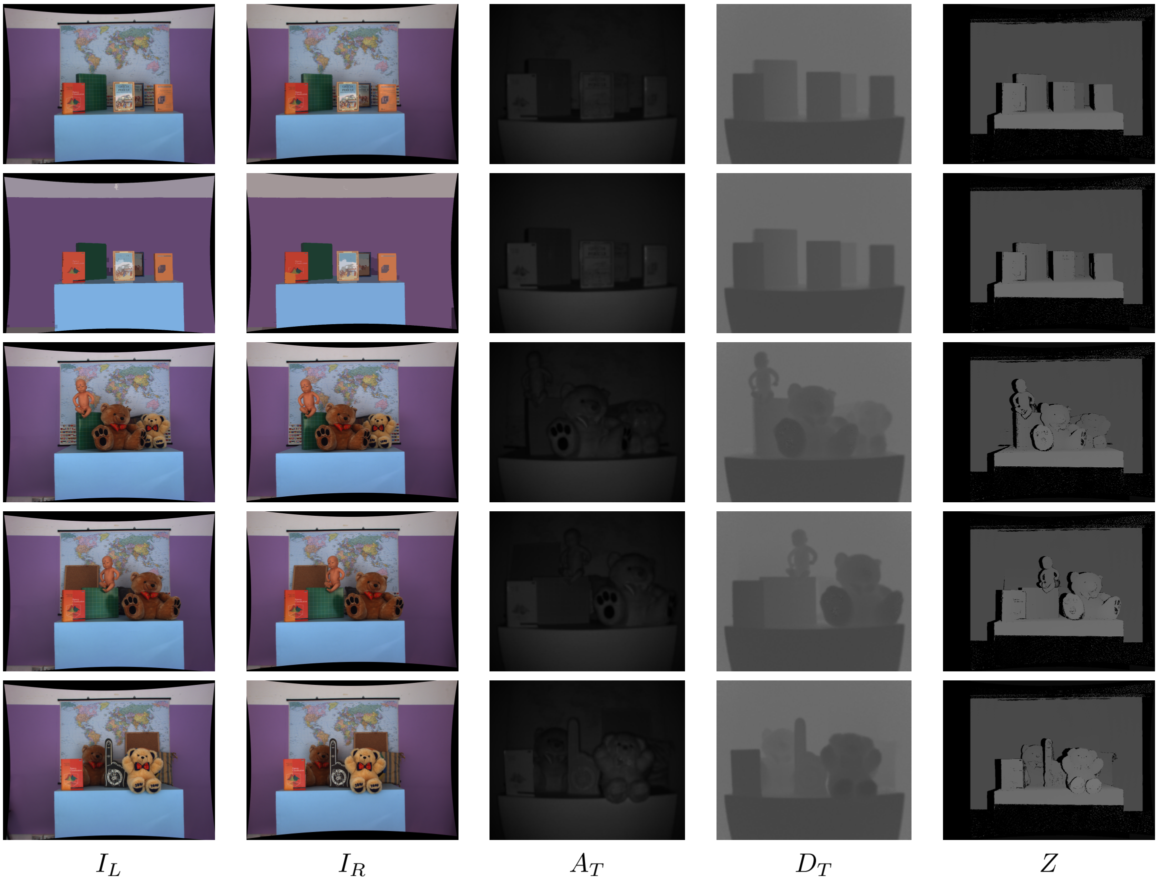


Fig. 12: Sample scenes used for the experimental results (each row correspond to a scene). First column I_L (resolution $[1032 \times 778]$), second I_R (resolution $[1032 \times 778]$), third A_T (resolution $[176 \times 144]$), fourth D_T (resolution $[176 \times 144]$) and fifth the ground truth depth-map Z (resolution $[1032 \times 778]$). Only the central portion of the acquired scene is considered for the analysis of the results.

The fusion procedures of Fig. 8 take as input the two high resolution color images I_L and I_R of the two cameras as well as the low resolution depth-map D_T and the amplitude image A_T acquired by the ToF camera T. The output of both fusion algorithms is a depth-map $\hat{Z}$ with resolution $[704 \times 576]$ (resolution can go up to $[1056 \times 864]$ or down to $[352 \times 288]$) from the point of view of the ToF camera (defined on the lattice $\Lambda_Z = \Lambda_T^L$, with $L = 4$), characterized also by high distance measurement accuracy. In order to validate the accuracy of the two fusion algorithms, we compared the quality of the estimated depth-maps $\hat{Z}$ against a ground truth acquired by a *space-time* stereo vision system [50], [51] for different scenes (shown in Fig. 12). A set of 600 frames with 600 different projected patterns are adopted for each scene. The ground truth depth-maps are estimated by integrating all the 600 images with the 600 patterns. A sub-pixel refinement and a *left-right check* were also applied. The accuracy of the depth-maps obtained by space-time stereo is of about $1-2[mm]$.

9.1 Evaluation of the ML fusion scheme

One of the major contributions of the proposed fusion method is the likelihood model adopted for ToF cameras measurements, which accounts for the matricial nature of ToF cameras, for depth discontinuities and for the near IR reflectivity of the scene. Its effectiveness can be well appreciated by comparing some examples of ToF and stereo likelihoods and of their multiplication (i.e., the joint likelihood or data term) clearly showing that the ML approach can both improve the accuracy of depth measurements far from depth discontinuities and correct erroneous measurements of the ToF camera near depth discontinuities. Let us firstly consider (Fig. 13.a) the case of a pixel p_i far from depth discontinuities from the first scene of Fig. 12. It is worth noting how the maximum of the ToF likelihood (second row) corresponds to depth $1564[mm]$, the maximum of the stereo likelihood (third row) to depth $1578[mm]$ and the maximum of the joint likelihood (i.e., the output of the ML approach, shown in the last row) to depth $1580.2[mm]$ which is also the ground truth, i.e., better than each of the single devices.

Fig. 13.b shows the case of a pixel p_i near depth

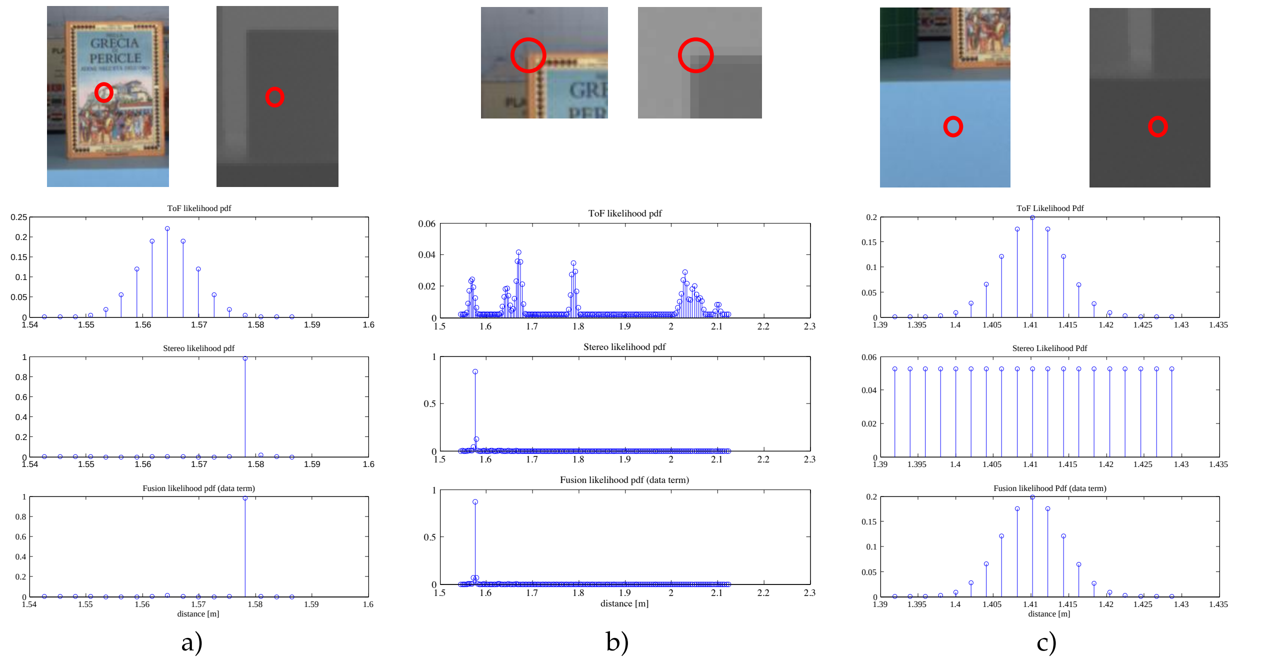


Fig. 13: Likelihood functions: Image and depth data with the considered point of the scene highlighted (first row), ToF likelihood (second row), stereo likelihood (third row) and joint likelihood (fourth row) relative to: a) a point far from scene distance discontinuities; b) a point near scene distance discontinuities; c) a point in a texture-less area.

discontinuities from the same scene: in this case, the ground truth depth is $1576[mm]$. The point is near a distance discontinuity characterized by a surface at $1584[mm]$ (where the point actually lies) and a surface at $2079[mm]$. The depth measured by the ToF is $1789[mm]$, resulting in this case rather inaccurate with an error of $213[mm]$ due to the effects explained in Section 4. The texture around the considered pixel makes accurate the stereo measurement, with the likelihood maximum at $1576[mm]$. The accuracy of the stereo likelihood is passed to the joint likelihood with maximum at $1576[mm]$ which is the output of the ML method, clearly showing how the fusion algorithm can improve the quality of ToF measurements.

The example shown in the first row of Fig. 13.c shows how the fusion algorithm can also improve the stereo measurement accuracy. Fig. 13.c refers to a point characterized by lack of texture in the color images I_L and I_R. In this situation depth measurements performed by stereo methods are unreliable, as indicated by the uniform stereo likelihood shown in the third row of Fig. 13.c. ToF measurements are not affected by the lack of texture and the ToF likelihood has a maximum at depth $1410.2[mm]$, rather close to the actual depth of $1411[mm]$. Since the stereo likelihood does not have any peak, the ToF likelihood shapes out the joint likelihood, giving it a maximum at depth $1410.2[mm]$. Fig. 13 shows that the ML algorithm gives better results than the ones obtained by the stereo vision algorithm alone. Furthermore the ToF model (7) together with the stereo model in the ML approach work robustly with respect to scene depth discontinuities and textured and textureless surfaces. Fig. 14 shows the depth error obtained by the ML approach for the first scene of Fig. 12. The error peaks of Fig. 14 correspond to the texture-less slanted surface of the table where stereo measurements become unreliable because of the lack of texture and ToF measurements become unreliable because of surface declivity, in agreement with the fact that the ML results are obtained just by picking at each pixel the depth which maximizes the joint likelihood at that pixel without any global optimization. It is also worth noting that the error peaks are about an order of magnitude greater than the average errors. The unreliability of ToF measures in case of surface

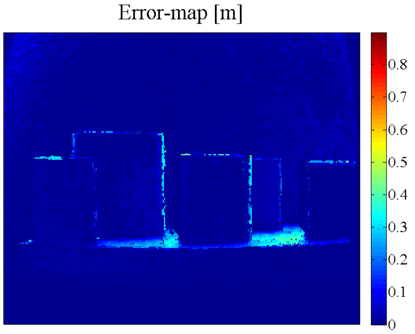


Fig. 14: Depth error-map (expressed in $[m]$) of the ML approach for the first scene of Fig. 12.

declivity can only be appreciated if one uses real ToF measurements: if in our experiments we had used non-experimental ToF data (e.g., a downsampled version of the ground-truth corrupted by additive noise) such a characteristic of real ToF measurements would not have been noticed at all.

9.2 Evaluation of the MAP fusion scheme

The proposed MAP framework based on the global optimization of (3) by LBP allows to obtain better performances with respect to the ML approach, as shown by the results in Table 1 reporting the average values of the ToF accuracy, of the stereo accuracy, of the ML accuracy and finally of the proposed MAP approach for all the 5 scenes of Fig. 12. In particular the first row of Table 1 shows that for the first scene of Fig. 12 the accuracy of the ML fusion approach is always better than the one of ToF and stereo measurements alone with an average error reduction of 13% with respect to the ToF measurements. ToF data is on the average more accurate than stereo data, nevertheless exploiting the proposed approach the information produced by the stereo system can be efficiently used to improve the ToF data accuracy even though the stereo vision system itself is less accurate. Note how the accuracy of our stereo measurements is not comparable with the accuracy of stand-alone stereo vision algorithms, because it is greatly improved by the *useful interval* supplied by ToF measurements. Table 1 also indicates that the MAP approach is more accurate than the ML one on all the considered scenes with an error reduction of 9% with respect to the ML results and a reduction of about the 20% with respect to the average ToF measurement errors. This confirms that the extra complexity introduced by the MAP model with respect to the ML scheme leads to better performances. A visual comparison of the two approaches is shown in Fig. 15 that shows a detail of the depth map of the third scene estimated by the ML and the MAP approaches. This example shows the distributed local quality improvement due to the application of the LBP which is sometimes hard to capture in terms of average error value improvement because it concerns small regions.

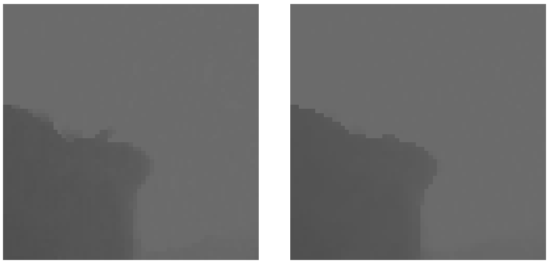


Fig. 15: Detail of the depth-map estimated by the ML (left) and the MAP (right) approaches.

Scene	ToF	Stereo	ML	MAP
1	22	30	20	**17**
2	24	35	20	**18**
3	27	30	25	**23**
4	28	29	22	**20**
5	27	31	25	**24**

TABLE 1: Accuracy (in $[mm]$) of depth information acquired by ToF, stereo, ML fusion and MAP fusion. Stereo accuracy benefits from the *useful interval* concept. The table refers to output depth-maps with resolution $[352 \times 288]$.

Fig. 16 shows the depth-maps estimated by the proposed approaches (ML and MAP) for the five scenes of Fig. 12. The resolution of the considered depth-map is $[704 \times 576]$, i.e., $16\times$ the resolution of the depth-maps acquired by the SR4000 and the estimated depth-maps are compensated for camera distortion. Note that it is possible to perform the undistortion artifacts-free directly in the three-dimensional space. Fig. 16 shows that the MAP-MRF depth estimates are much less noisy and more accurate than the ML estimates. This becomes particularly clear when comparing ML and MAP depths of the areas denoted by the arrows. In particular edge regions are more accurately represented. The accuracy of the proposed method is always comparable or better than the best one between ToF and stereo accuracies and the depth resolution is less than $1[mm]$ (it can be tuned by suitable sampling of range inside the *useful interval*).

Table 2 compares the results of the proposed work with some other approaches, i.e., the methods of [21], [15], [9] and [28]. These methods have been implemented and tested on the scenes of Fig. 12. Table 2 shows a quantitative comparison of their depth estimates and for comparison purposes it also reports the performances of the stereo vision algorithm of [52]. Table 3 shows the results of the same experiment, but only considering the error in areas close to depth discontinuities[1]. It is clear that the proposed approach outperforms the methods of [21] , [15], [9] and [28]. In particular the proposed MAP approach has the best performances on all the considered scenes, except that [28] attains similar performances on Scene 3 and [9] on Scene 5. The proposed method behaves very well also near depth discontinuities, and only [28] slightly outperforms it in Scene 1 and has similar results in Scene 2. The ML scheme has also good performances, similar to the ones of [9] which is also based on a ML optimization scheme.

Qualitative comparisons are also possible. The pro-

[1] Depth discontinuities are defined as the set of pixels in a 9×9 neighborhood of pixels for which the Laplacian of Gaussian operator applied to the ground-truth depth-map provides an output discontinuity value greater than $50[mm]$. The computed discontinuity maps are available at the URL `http://lttm.dei.unipd.it/paper_data/fusion/`.

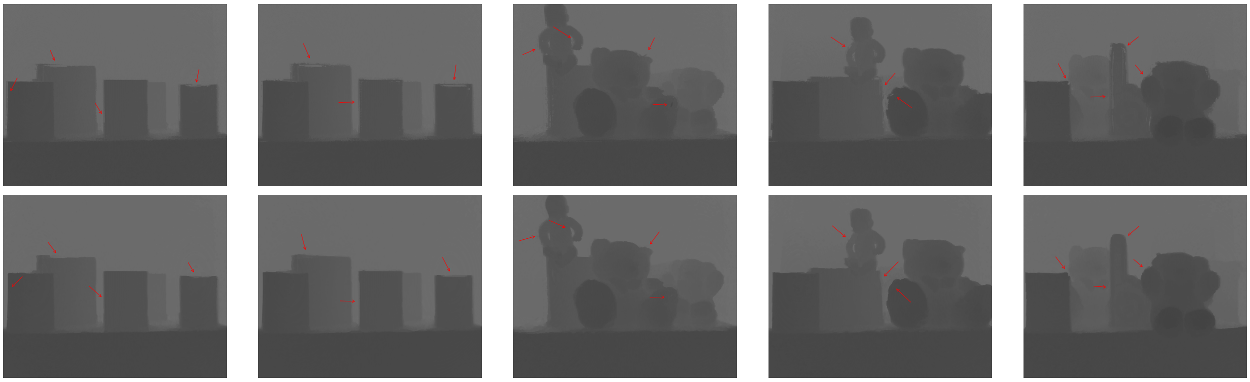


Fig. 16: ML (fist row) and MAP (second row) estimates of the depth-maps of the five scenes of Fig. 12. The arrows point to some of the most relevant artifacts of ML estimation which are removed by the MAP-MRF approach (the depth maps are available at the URL `http://lttm.dei.unipd.it/paper_data/fusion/` in order to better appreciate the details).

Scene	[9]	[15]	[28]	*[52]*	[21]	ML	MAP
1	21	26	22	*112*	25	20	**17**
2	20	20	21	*141*	24	20	**18**
3	25	33	**23**	*64*	37	25	**23**
4	22	38	24	*118*	29	22	**20**
5	**24**	38	27	*115*	33	25	**24**

TABLE 2: Comparison of the depth estimation accuracy in $[mm]$ obtained with various approaches. All the compared methods exploit both stereo and ToF measures with the exception of [52] that is a pure stereo vision method.

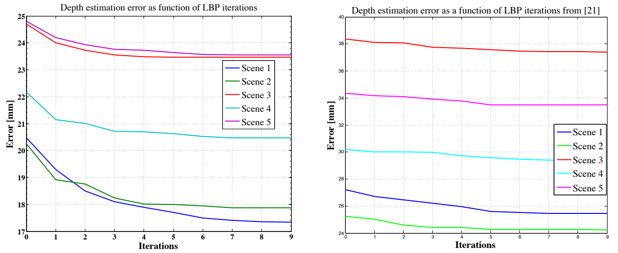
Scene	[9]	[15]	[28]	*[52]*	[21]	ML	MAP
1	66	91	**61**	*297*	75	67	**63**
2	68	96	**64**	*332*	78	69	**64**
3	64	87	66	*198*	88	65	**56**
4	65	98	67	*266*	79	66	**56**
5	64	97	70	*275*	79	68	**61**

TABLE 3: Comparison of the depth estimation accuracy in $[mm]$ obtained with various approaches in areas close to depth discontinuities[1] .

posed method provides an high resolution depth-map with respect to [9] which instead gives as output a depth map only at the low resolution of the ToF camera. Moreover, the global optimization step provides more robust scene depth estimates with respect to the methods of [9], [15], [28].

With respect to [21], that also exploits a MAP optimization the porposed approach has an error about 30% lower, due to the more refined ToF likelihood model which in our case is formally derived from the ToF camera error model and to the adoption of the *useful interval*, a provision which allows to reduce the computational complexity and to improve reconstruction results.

For what concerns the computational complexity of LBP, as shown in Section 8, the complexity of the message-passing operation is dominated by a term proportional to $\bar{N}^2$, where $\bar{N}$ is the average number of range values computed across all the sites. the useful interval on the considered scenes reduces the average range down to the 7% of the range of the full scene. The computational complexity is also proportional to the number of iterations L of LBP before convergence. The proposed method converges quite quickly to the optimal value, typically in just a few iterations. For example Fig. 17 reports the MAP estimation error of our approach and of [21] as a function of LBP iterations for all the five scenes. We can assume that convergence is reached when the depth-error reduction at the current iteration (with respect to the depth ground-truth computed with space-time stereo) is less than a tenth of the depth-error reduction at the previous iteration. With this definition, the proposed LBP algorithm converges on average in 8 iterations on the considered scenes, while standard approaches, such as [21] converges in 7 iterations. Notice also that our approach allows to obtain a faster error decrease in the first iterations. Therefore in such scenes it is possible to notice that the concept of useful interval and its exploitation in LBP reduces the computational complexity dominating term more than $100\times$ (notice that the number of iterations is similar but the amount of computation in each iteration is much smaller due to the narrower ranges as previously discussed), with respect to the usage of the same lattice, same neighborhood and same range-sampling step in a standard approach. It is worth noting that this is only an approximation of the computational complexity of LBP and that with some knowledge of the minimum and maximum scene depth it is possible to speed up the standard LBP approach. The source-code of the proposed LBP optimization algorithm is available at the URL `http://lttm.dei.unipd.it/paper_data/fusion/`.



a) b)

Fig. 17: MAP estimation error vs. LBP iterations for the 5 considered scenes: a) Our approach; b) Zhu et Al [21].

10 Conclusions

This paper proposes a new method for the fusion of data acquired by a ToF camera and a stereo pair in order to obtain high quality depth estimates. The elements of novelty of the proposed approach are several. First of all the paper introduces a novel model for ToF depth measurements which provides an accurate description of the measurement process in presence of depth discontinuities and extends standard stereo measurements model to account for advanced segmentation clues. The proposed ToF measurements model overcomes some fundamental limitations of previous models treating ToF cameras only in a pixel-wise fashion.

Depth data acquired by a ToF camera and a stereo system are combined together using both a ML approach and a global cost function derived from our premises within a MAP-MRF approach. The proposed MAP-MRF framework is characterized by different depth range values at each site of the lattice of the estimated depth-map. The presence of site-dependent ranges requires an extension of the classical LBP optimization which is another element of novelty of this paper applicable to a variety of different computer vision and 3D reconstruction problems.

The performances of the two proposed methods are assessed by experimental validation against a ground truth obtained by space-time stereo. Evaluation has been purposely performed on real data since artificially generated data cannot account for the complex interactions typical of ToF measurement.

The results clearly show the effectiveness of the proposed ToF depth measurements model in presence of depth discontinuities without any performance sacrifice with other depth configurations and the effectiveness of the proposed ML and MAP-MRF frameworks for high quality scene depth estimates

Further research will address the improvement of the measurement models for both stereo and ToF sensors. In particular more accurate models of typical stereo vision artifacts will be employed, as well as ToF models that account for artifacts introduced by multipath. The LBP scheme will be improved by considering approaches for continuous distributions modeled by Gaussian mixture models, and possibly the fusion framework will be applied to Conditional Random Fields prior models. We also envision extensions to dynamic scenes by including constraints in the time domain to the proposed MAP scheme.

References

[1] R. Lange, “3d time-of-flight distance measurement with custom solid-state image sensors in cmos/ccd-technology,” Ph.D. dissertation, University of Siegen, 2000.

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

[3] F. Remondino and D. Stoppa, Eds., *TOF Range-Imaging Cameras*. Springer, 2013.

[4] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo, *Time-of-Flight Cameras and Microsoft Kinect*. Springer, 2012.

[5] M. Schmidt, “Analysis, modeling and dynamic optimization of 3d time-of-flight imaging systems,” Ph.D. dissertation, University of Heidelberg, 2011.

[6] T. Kahlmann and H. Ingensand, “Calibration and development for increased accuracy of 3d range imaging cameras,” *Journal of Applied Geodesy*, vol. 2, pp. 1–11, 2008.

[7] 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.*, vol. 5, pp. 425–433, 2008.

[8] D. Piatti and F. Rinaudo, “Sr-4000 and camcube3.0 time of flight (tof) cameras: Tests and comparison,” *Remote Sensing*, vol. 4, no. 4, pp. 1069–1089, 2012.

[9] C. Dal Mutto, P. Zanuttigh, and G. Cortelazzo, “A probabilistic approach to tof and stereo data fusion,” in *Proceedings of 3DPVT*, Paris, France, 2010.

[10] F. Mufti and R. Mahony, “Statistical analysis of measurement processes for time-of flight cameras,” *Proceedings of SPIE the International Society for Optical Engineering*, 2009.

[11] M. Reynolds, J. Dobos, L. Peel, T. Weyrich, and G. Brostow, “Capturing time-of-flight data with confidence,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011, pp. 945–952.

[12] A. Sabov and J. Krüger, “Identification and correction of flying pixels in range camera data,” in *Proceedings of the 24th Spring Conference on Computer Graphics*. ACM, 2008, pp. 135–142.

[13] R. Nair, K. Ruhl, F. Lenzen, S. Meister, H. Schfer, C. Garbe, M. Eisemann, M. Magnor, and D. Kondermann, “A survey on time-of-flight stereo fusion,” in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Springer, 2013, vol. 8200, pp. 105–127.

[14] J. Diebel and S. Thrun, “An application of markov random fields to range sensing,” in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*. MIT Press, 2005.

[15] Q. Yang, R. Yang, J. Davis, and D. Nister, “Spatial-depth super resolution for range images,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[16] Q. Yang, N. Ahuja, R. Yang, K. Tan, J. Davis, B. Culbertson, J. Apostolopoulos, and G. Wang, “Fusion of median and bilateral filtering for range image upsampling,” *Image Processing, IEEE Transactions on*, 2013.

[17] V. Garro, C. Dal Mutto, P. Zanuttigh, and G. Cortelazzo, “A novel interpolation scheme for range data with side information,” in *Proceedings of CVMP*, 2009, pp. 52 –60.

[18] V. Garro, C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo, “Edge-preserving interpolation of depth data exploiting color information,” *Annals of Telecommunications*, pp. 1–17, 2013.

[19] J. Dolson, J. Baek, C. Plagemann, and S. Thrun, “Upsampling range data in dynamic environments,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1141–1148.

[20] K. D. Kuhnert and M. Stommel, “Fusion of stereo-camera and pmd-camera data for real-time suited precise,” 2006.

[21] J. Zhu, L. Wang, R. Yang, and J. Davis, “Fusion of time-of-flight depth and stereo for high accuracy depth maps,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[22] J. Zhu, L. Wang, J. Gao, and R. Yang, "Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 899–909, 2010.

[23] J. Zhu, L. Wang, R. Yang, J. Davis, and Z. Pan, "Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1400–1414, 2011.

[24] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo - stereo matching with slanted support windows," in *British Machine Vision Conference 2011*, 2011, pp. 1–11.

[25] J. Godbaz, M. Cree, and A. Dorrington, "Mixed pixel return separation for a full-field ranger," in *Proc. of Image and Vision Computing New Zealand*, Nov 2008, pp. 1–6.

[26] U. Hahne and M. Alexa, "Depth imaging by combining time-of-flight and on-demand stereo," *Dynamic 3D Imaging*, pp. 70–83, 2009.

[27] R. Nair, F. Lenzen, S. Meister, H. Schaefer, C. Garbe, and D. Kondermann, "High accuracy tof and stereo sensor fusion at interactive rates," in *Proceedings of 2nd Workshop on Consumer Depth Cameras for Computer Vision*, 2012.

[28] C. Dal Mutto, P. Zanuttigh, S. Mattoccia, and G. M. Cortelazzo, "Locally consistent tof and stereo data fusion," in *Proc. of 2nd Workshop on Consumer Depth Cameras for Computer Vision*, 2012.

[29] 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.

[30] Y. 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.

[31] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

[32] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1222–1239, 2001.

[33] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society*, vol. B-48, pp. 259–302, 1986.

[34] M. Wainwright, T. Jaakkola, and A. Willsky, "Map estimation via agreement on (hyper)trees: Message-passing and linear programming approaches," *IEEE Transactions on Information Theory*, vol. 51, pp. 3697–3717, 2002.

[35] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1068 –1080, 2008.

[36] M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters," in *Proc. of Int. Conf. on Computer Vision*, 2003, pp. 900–906.

[37] J. W. Woods, *Multidimensional Signal, Image, And Video Processing And Coding*. Elsevier Inc., 2006.

[38] D. Wu, A. Velten, M. OToole, B. Masia, A. Agrawal, Q. Dai, and R. Raskar, "Decomposing global light transport using time of flight imaging," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 123–138, 2014.

[39] A. Velten, T. Willwacher, O. Gupta, A. Veeraraghavan, M. G. Bawendi, and R. Raskar, "Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging," *Nature Communications*, vol. 3, p. 745, 2012.

[40] J. Sun, N. Zheng, and H. Shum, "Stereo matching using belief propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 787–800, 2003.

[41] 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*, vol. 26, pp. 359–374, 2001.

[42] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, 2001.

[43] S. Mattoccia, S. Giardino, and S. Gambini, "Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering," in *ACCV*, 2009.

[44] S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," *International Journal of Computer Vision*, vol. 35, no. 3, pp. 269–293, 1999.

[45] "Edison," coewww.rutgers.edu/riul/research/code/EDISON.

[46] Q. Yang, K. Tan, B. Culbertson, and J. Apostolopoulos, "Fusion of active and passive sensors for fast 3d capture," in *IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, 2010.

[47] M. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," *International Journal of Computer Vision*, vol. 19, pp. 57–91, 1996.

[48] L. Wang, H. Jin, and R. Yang, "Search space reduction for mrf stereo," in *Proc. of ECCV 2008*. Springer, 2008, pp. 576–588.

[49] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[50] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz, "Spacetime stereo: A unifying framework for depth from triangulation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[51] L. Zhang, B. Curless, and S. Seitz, "Spacetime stereo: shape recovery for dynamic scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[52] H. Hirschmuller, "Stereo processing by semi-global matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 328–341, 2008.

Carlo Dal Mutto Carlo Dal Mutto holds a M.Sc. Degree in Communication Engineering University of Padova (2009) as well as a Ph.D Degree in Information Engineering from the same university (2013). He has been a visiting student at Boston University (2008) and a visiting scholar at Duke University (2012). His Ph.D. research focused on acquisition and processing of color and 3D geometry data. He co-authored several papers on this topic, as well as two book chapters and a book. He holds a few industrial patents. He currently is the R&D lead at Aquifi Inc.

Pietro Zanuttigh Pietro Zanuttigh was born in 1978. He graduated in Computer Engineering at the University of Padova (Italy) in 2003 and got the Ph.D. degree from the same university. In 2007 he became an assistant professor at the University of Padova. Now he works in the multimedia technology and telecommunications group and his research activity focuses on image and 3D data acquisition, compression and processing. He is the co-auhtor of a book and of several publications on international journals and conference proceedings.

Guido M. Cortelazzo Guido M. Cortelazzo graduated in Electronic Engineering at the University of Padova (1976). He holds a Ph.D. Degree from the University of Illinois at Urbana-Champaign (1984). From 1983 to 1986 he worked with M/A-COM Linkabit. In 1986 he joined the Department of Information Engineering (DEI) of the University of Padova where he is full professor. He was active in the organization of several special sessions and meetings and a founding member of the 3DPVT (now 3DV) conference. The current professional interests of Guido M. Cortelazzo concern the automatic construction of 3D models of still and dynamic scenes and their progressive transmission. He is the author of about seventy journal papers, co-author of two books and co-editor of two special issues.