Reliable Fusion of ToF and Stereo Depth Driven by Confidence Measures

Giulio Marin, Pietro Zanuttigh
Department of Information Engineering
University of Padova, Italy
{maringiu,zanuttigh}@dei.unipd.it

Stefano Mattoccia
Computer Science and Engineering
University of Bologna, Italy
stefano.mattoccia@unibo.it

***Abstract*—In this paper we propose a framework for the fusion of depth data produced by a Time-of-Flight (ToF) camera and stereo vision system. Initially, depth data acquired by the ToF camera are upsampled by an ad-hoc algorithm based on image segmentation and bilateral filtering. In parallel a dense disparity map is obtained using the Semi-Global Matching stereo algorithm. Reliable confidence measures are extracted for both the ToF and stereo depth data. In particular, ToF confidence also accounts for the mixed-pixel effect and the stereo confidence accounts for the relationship between the pointwise matching costs and the cost obtained by the semi-global optimization. Finally, the two depth maps are synergically fused by enforcing the local consistency of depth data accounting for the confidence of the two data sources at each location. Experimental results clearly show that the proposed method produces accurate high resolution depth maps and outperforms the compared fusion algorithms.**

***Index Terms*—Stereo vision, Time-of-Flight, data fusion, confidence metrics.**

I. Introduction

Depth estimation is a challenging computer vision problem for which many different solutions have been proposed. Among them, passive stereo vision systems are widely used since they only require a pair of standard cameras and can provide a high resolution depth estimation in real-time. However, even if recent research in this field has greatly improved the quality of the estimated geometry [1], results are still not completely reliable and strongly depend on scene characteristics. Active devices like ToF cameras and light-coded cameras (e.g., Microsoft Kinect), are able to robustly estimate in real time the 3D geometry of a scene but they are also limited by a low spatial resolution and a high level of noise in their measurements, especially for low reflective surfaces. Since the characteristics of ToF cameras and stereo data are complementary, the problem of their fusion has attracted considerable interest in the last few years.

An effective fusion scheme requires two fundamental building blocks: the first is an estimation of dense confidence measures for each device and the second is an efficient fusion algorithm that estimates the depth values from the data of the two sensors and their confidence values. In this paper we address these requirements by introducing accurate models for the estimation of the confidence measures for ToF and stereo data depending on the scene characteristics at each location, and then extending the Local Consistency (LC) fusion framework of [2] to account for the confidence measures associated with the acquired data. First, the depth data acquired by the ToF camera are upsampled to the spatial resolution of the stereo vision images by an efficient upsampling algorithm based on image segmentation and bilateral filtering. A reliable confidence map for the ToF depth data is computed according to different clues including the mixed pixel effect caused by the finite size of ToF sensor pixels. Second, a dense disparity map is obtained by a global (or semi-global) stereo vision algorithm, and the confidence measure of the estimated depth data is computed considering both the raw block matching cost and the globally optimized cost function. Finally, the upsampled ToF depth data and the stereo vision disparity map are fused together. The proposed fusion algorithm extends the LC method [3] by taking into account the confidence measures of the data produced by the two devices and providing a dense disparity map with subpixel precision. Both the confidence measures and the subpixel disparity estimation represent novel contributions not present in the previous versions of the LC framework [2], [3], and to the best of our knowledge, the combination of local and global cost functions is new and not used by any other confidence measure proposed in the literature.

II. Related Work

Matricial ToF range cameras have been the subject of several recent studies, e.g., [4], [5], [6], [7], [8], [9]. In particular, [8] focuses on the various error sources that influence range measurements while [9] presents a qualitative analysis of the influence of scene reflectance on the acquired data.

Stereo vision systems have also been the subject of a significant amount of research, and a recent review on this topic can be found in [1]. The accuracy of stereo vision depth estimation strongly depends on the framed scene's characteristics and the algorithm used to compute the depth map, and a critical issue is the estimation of the confidence associated with the data. Various metrics have been proposed for this task and a complete review can be found in [10].

These two subsystems have complementary characteristics, and the idea of combining ToF sensors with standard cameras has been used in several recent works. A complete survey of this field can be found in [6], [11]. Some work focused on the combination of a ToF camera with a single color camera [12], [13], [14], [15], [16], [17]. An approach based on bilateral filtering is proposed in [13] and extended in [14]. The approach of [16] instead exploits an edge-preserving scheme to interpolate the depth data produced by the ToF sensor. The recent approach of [15] also accounts for the confidence measure of ToF data. The combination of a ToF camera and a stereo camera is more interesting, because in this case both subsystems can produce depth data [9], [18], [19], [20]. A method based on a probabilistic formulation is presented in [21], where the final depth-map is recovered by performing a ML local optimization in order to increase the accuracy of the depth measurements from the ToF and stereo vision system. This approach has been extended in [22] with a more refined measurement model which also accounts for the mixed pixel effect and a global optimization scheme based on a MAP-MRF framework. The method proposed in [23], [24] is also based on a MAP-MRF Bayesian formulation, and a belief propagation based algorithm is used in order to optimize a global energy function. An automatic way to set the weights of the ToF and stereo measurements is presented in [25]. Another recent method [26] uses a variational approach to combine the two devices. The approach of [2], instead, uses a locally consistent framework [3] to combine the measurements of the ToF sensor with the data acquired by the color cameras, but the two contributions are equally weighted in the fusion process. This critical issue has been solved in this paper by extending the LC framework. Finally the approach of [27] computes the depth information by hierarchically solving a set of local energy minimization problems.

III. Proposed Method

We consider an acquisition system made of a ToF camera and a stereo vision system. The goal of the proposed method is to provide a dense confidence map for each depth map computed by the two sensors, then use this information to fuse the two depth maps into a more accurate description of the 3D scene. The approach assumes that the two acquisition systems have been jointly calibrated, e.g., using the approach of [21]. In this method, the stereo pair is rectified and calibrated using a standard approach [28], then the intrinsic parameters of the ToF sensor are estimated. Finally, the extrinsic calibration parameters between the two systems are estimated with a closed-form technique. The proposed algorithm is divided into three different steps:

1) The low resolution depth measurements of the ToF camera are reprojected into the lattice associated with the left camera and a high resolution depth-map is computed by interpolating the ToF data. The confidence map of ToF depth data is estimated using the method described in Section IV.
2) A high resolution depth map is computed by applying a stereo vision algorithm on the images acquired by the stereo pair. The confidence map for stereo depth data is estimated as described in Section V.
3) The depth measurements obtained by the upsampled ToF data and the stereo vision algorithm are fused together by means of an extended version of the LC technique [3] using the confidence measures from the previous steps.

IV. ToF Depth and Confidence Estimation

A. High Resolution Depth Estimation from ToF Data

Since stereo data typically have higher resolutions than those of ToF cameras, the projection of ToF data on the lattice associated with the left color camera produces a set of sparse depth measurements that need to be interpolated. In order to obtain an accurate high resolution map, especially in proximity of edges, we exploit the method of [2], combining cross bilateral filtering with the help of segmentation. First, all the 3D points acquired by the ToF camera are projected onto the left camera lattice Λ_l, obtaining a set of samples $p_i, i = 1, ..., N$ that does not include samples that are occluded from the left camera point of view. The color image acquired by the left camera is then segmented using mean-shift clustering [29], obtaining a segmentation map used to guide an extended bilateral filter developed for the interpolation of the p_i samples. The output of the interpolation method is a disparity map defined on the left camera lattice Λ_l. Since the fusion algorithm works in the disparity space, the interpolated depth map is converted into a disparity map with the well known relationship $d = bf/z$, where d and z are disparity and depth values, b is the baseline of the stereo system and f is the focal length of the rectified stereo camera.

B. Confidence Estimation of ToF Depth Data

As reported in many studies on matricial ToF technology [6], the reliability of the ToF measurements is affected by several issues, e.g., the reflectivity of the acquired surface, the measured distance, multi-path issues or mixed pixels in proximity of edges, and thus is very different for each different sample. A proper fusion algorithm requires a reliable confidence measure for each pixel. In this paper we propose a novel model for the confidence estimation of ToF measurements, using both radiometric and geometric properties of the scene. As described in the rest of this section, our model is based on two main clues that can be separately captured by two metrics. The first one, P_{AI}, considers the relationship between amplitude and intensity of the ToF signal, while the second one, P_{LV}, accounts for the local depth variance. The two confidence maps P_{AI} and P_{LV} consider independent geometric and photometric properties of the scene, therefore, the overall ToF confidence map P_T is obtained by multiplying the two confidence maps together

$$P_T = P_{AI} P_{LV}. \quad (1)$$

1) Confidence from amplitude and intensity values: ToF cameras provide both the amplitude and the intensity of the received signal for each pixel. The amplitude of the received signal depends on various aspects, but the two most relevant are the reflectivity characteristics of the acquired surfaces and the distance of the scene samples from the camera. Intensity also depends on these two aspects, but is additionally affected by the ambient illumination in the wavelength range of the camera. A confidence measure directly using the distance of objects in the scene could be considered, but distance strongly affects the amplitude, and thus the proposed measure already implicitly takes the distance into account. The received amplitude strongly affects the accuracy of the measures and a higher amplitude leads to a better signal-to-noise ratio and thus to more accurate measurements [5]. As reported in [6], [22], the distribution of the ToF pixel noise can be approximated by a Gaussian with standard deviation

$$\sigma_z = \frac{c}{4\pi f_{mod}} \frac{1}{SNR} = \frac{c}{4\pi f_{mod}} \frac{\sqrt{I/2}}{A} \qquad (2)$$

where f_{mod} is the IR frequency of the signal sent by the ToF emitters, A is the amplitude value at the considered pixel, I is the intensity value at the same location and c is the speed of light. Note that since the data fusion is performed on the upsampled disparity map, the confidence maps must be of the same resolution, but amplitude and intensity images are at the same low resolution of the ToF depth map. In order to solve this issue, each pixel $\mathbf{p}_L$ in the left color image is first back-projected to the 3D world and then projected to the corresponding pixel coordinates in the ToF lattice $\mathbf{p}_L^{TOF}$.

From (2) it can be observed that when amplitude A increases, precision improves, since the standard deviation decreases, while when intensity I increases, the precision decreases. Intensity I depends on two factors: the received signal amplitude A and the background illumination. An increase in the amplitude leads to an overall precision improvement given the squared root dependence with respect to I in (2), while in the second case precision decreases since A is not affected.

Before mapping σ_z to the confidence values, it is important to notice that the proposed fusion scheme works on the disparity domain, while the measurement standard deviation (2) refers to depth measurements. For a given distance z, if a certain depth error Δ_z around z is considered, the corresponding disparity error Δ_d also depends on the distance z, due to the inverse proportionality between depth and disparity. If σ_z is the standard deviation of the depth error, the corresponding standard deviation σ_d of the disparity measurement can be computed as:

$$2\sigma_d = |d_1 - d_2| = \frac{bf}{z - \sigma_z} - \frac{bf}{z + \sigma_z} = \\ = bf \frac{2\sigma_z}{z^2 - \sigma_z^2} \quad \Rightarrow \quad \sigma_d = bf \ \frac{\sigma_z}{z^2 - \sigma_z^2} \qquad (3)$$

where b is the baseline of the stereo system and f is the focal length of the camera. Equation (3) provides the corresponding standard deviation of the noise in the disparity space for a given depth value. The standard deviation of the measurements in the disparity space is also affected by the mean value of the measurement itself, unlike the standard deviation of the depth measurement.

In order to map the standard deviation of the disparity measurements to the confidence values, we define two thresholds computed experimentally over multiple measurements. The first is $\sigma_{min} = 0.5$, corresponding to the standard deviation of a bright object at the minimum measurable distance of 0.5 m, while the second is $\sigma_{max} = 3$, corresponding to the case of a dark object at the maximum measurable distance of 5 m with the SR4000 sensor used in the experimental results dataset. If a different sensor is employed, the two thresholds can be updated by considering these two boundary conditions. Then, we assume that values smaller than σ_{min} correspond to the maximum confidence value, i.e., $P_{AI} = 1$, values bigger than σ_{max} have $P_{AI} = 0$ while values in the interval $[\sigma_{min}, \sigma_{max}]$ are linearly mapped to the confidence range $[0, 1]$, i.e.:

$$P_{AI} = \begin{cases} 1 & if\, \sigma_d \leq \sigma_{min} \\ \frac{\sigma_{max} - \sigma_d}{\sigma_{max} - \sigma_{min}} & if\, \sigma_{min} < \sigma_d < \sigma_{max} \\ 0 & if\, \sigma_d \geq \sigma_{max} \end{cases} \qquad (4)$$

2) Confidence from local variance: One of the main limitations of (2) is that it does not take into account the effect of the finite size of ToF sensor pixels, i.e., the mixed pixel effect [22]. In order to account for this issue we introduce another term in the proposed confidence model. When the scene area associated with a pixel includes two regions at different depths, e.g. close to discontinuities, the resulting estimated depth measure is a convex combination of the two depth values. For this reason, it is reasonable to associate a low confidence to these regions. The mixed pixel effect leads to convex combinations of depth values but this is not true for the multipath effect. These considerations do not affect the design of the ToF confidence since the LV metric just assumes that pixels in depth discontinuities are less reliable. If pixel p_i^{TOF} in the low resolution lattice of the ToF camera is associated with a scene area crossed by a discontinuity, some of the pixels p_j^{TOF} in the 8-neighborhood $\mathcal{N}(p_i^{TOF})$ of p_i^{TOF} belong to points at a closer distance, and some other pixels to points at a farther distance. Following this intuition the mean absolute difference of the points in $\mathcal{N}(p_i^{TOF})$ has been used to compute the second confidence term, i.e.:

$$D_l^{TOF} = \frac{1}{\left|\mathcal{N}(p_i^{TOF})\right|} \sum_{j \in \mathcal{N}(p_i^{TOF})} |z_i - z_j| \qquad (5)$$

where $\left|\mathcal{N}(p_i^{TOF})\right|$ is the cardinality of the considered neighborhood, in this case equal to 8, and z_i and z_j are the depth values associated with pixels p_i^{TOF} and p_j^{TOF},

respectively. We use the mean absolute difference instead of the variance to avoid assigning very high values to edge regions due to the quadratic dependence of the variance with respect to the local differences. For this term we used the depth values and not the disparity ones because the same depth difference would lead to different effects on the confidence depending if close or far points are considered. This computation is performed for every pixel with a valid depth value. Notice that some p_j^{TOF} considered in an 8-connected patch may not have a valid value. In order to obtain a reliable map, a constant value $K_d = T_h$ has been used in the summation (5) in place of $|z_i - z_j|$ for the pixels p_j^{TOF} without a valid depth value. To obtain the confidence information D_l on the left camera lattice, samples p_i on this lattice are projected on the ToF camera lattice and the corresponding confidence value is selected after a bilinear interpolation.

Points with high local variance are associated with discontinuities, therefore, low confidence should be assigned to them. Where the local variance is close to zero, the confidence should be higher. In order to compute the confidence term we normalize D_l to the $[0, 1]$ interval by defining a maximum valid absolute difference $T_h = 0.3$ corresponding to 30 cm and assigning higher likelihood values to the regions with lower local variability:

$$P_{LV} = \begin{cases} 1 - \frac{D_l}{T_h} & if \quad D_l < T_h \\ 0 & if \quad D_l \geq T_h \end{cases} \tag{6}$$

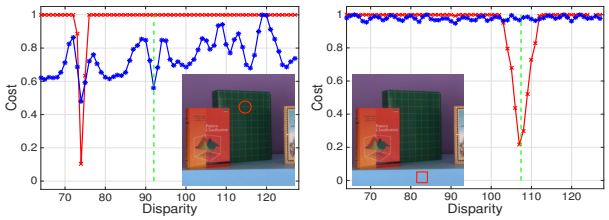
V. Stereo Disparity and Confidence Estimation

A. Disparity Computation from Stereo Data

The considered setup includes two calibrated color cameras, therefore an additional high resolution disparity map D_s can be inferred by stereo vision. The data fusion algorithm presented in the next section is independent of the choice of the stereo vision algorithm, however, for our experiments we used the Semi-Global Matching (SGM) algorithm [30]. The goal of this algorithm is to perform a 1D disparity optimization on multiple paths. Such an optimization minimizes on each path an energy term made of point-wise or aggregated matching costs C^l and a regularization term. We used the pointwise Birchfield-Tomasi metric over color data and 8 paths for the optimization, with window size of 7×7, $P_1 = 20$ and $P_2 = 100$. The energy terms are summed up obtaining a global cost function C^g that usually presents a very sharp peak at the minimum cost's location. In the rest of the section we analyze how the relationship between local cost C^l and global cost C^g can provide an effective confidence measure.

B. Confidence Estimation of Stereo Vision Data

The reliability of the disparity map is affected by the content of the acquired images, in particular by the texture of the scene. Uniform regions are usually the most challenging since it is difficult to estimate corresponding image points reliably. Global (or semi-global) methods tackle this problem by propagating neighbor values enforcing a smoothness



a) b)

Fig. 1: Comparison of local (*blue*) and global (*red*) costs: a) Cost functions of a repetitive pattern; b) Cost functions of a uniform region. The green line represent the ground truth disparity value.

constraint at the cost of a higher uncertainty in the disparity assignments. The globally optimized cost function typically has a very sharp peak, often resulting from the enforced smoothness constraint, corresponding to the propagated value even in areas where the data are not reliable. Current stereo vision confidence estimation approaches analyzing the cost function [10] do not account for the impact of global optimizations performed by most recent stereo vision methods. We believe that an optimal confidence metric can only be obtained by analyzing both cost functions. In the proposed approach this issue is handled by introducing a novel confidence measure considering both the local cost function C^l and the globally optimized one C^g.

In our analysis, at each pixel location for each disparity hypothesis d, we consider the point-wise local cost $C^l(d)$ and the global cost from the SGM algorithm $C^g(d)$, both scaled to the interval $[0, 1]$. Ideally the cost function should have a very well-defined minimum corresponding to the correct depth value but, as expected, in many practical situation this is not the case. Fig. 1 shows two points in the scene where the confidence should be low. In Fig. 1 a) the region surrounding the selected point has a periodic pattern and in Fig. 1 b) the region surrounding the selected point has a uniform color. However, the global cost function has a sharp peak and conventional confidence measures based only on global cost analysis would assign a high confidence to these pixels.

The terminology used to denote the points of interest on the cost functions is the following: the minimum cost for a pixel is denoted by C_1 and the corresponding disparity value by d_1, i.e.: $C_1 = C(d_1) = \min_d C(d)$, where disparity d has subpixel resolution. The second smallest cost value which occurs at disparity d_2 is C_2. For the selection of C_2, disparity values that are too close to d_1 (i.e., $|d_2 - d_1| \leq 1$) are excluded to avoid suboptimal local minima too close to d_1.

The proposed stereo confidence metric P_S is the combination of multiple clues, depending both on the properties of the local cost function and on the relationship between local and global costs. In particular it is defined as the

product of three factors:

$$P_S = \frac{\Delta C^l}{C_1^l}\left(1 - \frac{\min\{\Delta d^l, \gamma\}}{\gamma}\right)\left(1 - \frac{\min\{\Delta d^{lg}, \gamma\}}{\gamma}\right) \quad (7)$$

where $\Delta C^l = C_2^l - C_1^l$ is the difference between the second and first minimum local cost, $\Delta d^l = |d_2^l - d_1^l|$ is the corresponding absolute difference between the second and first minimum local cost locations, $\Delta d^{lg} = |d_1^l - d_1^g|$ is the absolute difference between the local and global minimum cost locations and γ is a normalization factor. The first term accounts for the robustness of the match, both the cost difference and the value of the minimum cost are important, as the presence of a single strong minimum with an associated small cost are usually sufficient conditions for a good match. However, in the case of multiple strong matches, the first term still provides a high score, e.g., in regions of the scene with a periodic pattern (Fig. 1b). The second term is a truncated measure of the distance between the first two cost peaks. It discriminates potentially bad matches due to the presence of multiple local minima. If the two minimum values are close enough, the associated confidence measure should provide a high value since the global optimization is likely to propagate the correct value and to provide a good disparity estimation. So far only the local cost has been considered so the last term accounts for the relationship between the local and global cost functions, scaling the overall confidence measure depending on the level of agreement between the local and global minimum locations. If the two minimum locations coincide, there is a very high likelihood that the estimated disparity value is correct, while on the other hand, if they are too far apart the global optimization may have produced incorrect disparity estimations, e.g. due to the propagation of disparity values in textureless regions. The constant γ controls the weight of the two terms and sets the maximum distance of the two minimum locations, after which the estimated value is considered unreliable. In our experiments we set $\gamma = 10$. Finally, if a local algorithm is used to estimate the disparity map, the same confidence measure can be used by considering only the first two terms.

Although the proposed metric is not as good as top performing stereo metrics evaluated in [10] in terms of AUC (e.g., PKRN), it performs better when used in our fusion framework. Indeed our goal is to propose a good confidence metric for the stereo system in the context of data fusion, where low confidence should be assigned to pixels belonging to textureless surfaces propagated by the global optimization, since ToF data are more reliable there. This feature is well captured by the proposed metric, but not by conventional stereo confidence metrics.

VI. Fusion of Stereo and ToF Disparity

Given the disparity maps and the confidence information for the ToF camera and the stereo vision system, the final step combines the multiple depth hypotheses available for each point by means of a technique that guarantees a locally consistent disparity map. Our method extends the LC technique [3], originally proposed for stereo matching, in order to deal with the two disparity hypotheses provided by our setup and modifies the original formulation to take advantage of the confidence measures to weight the contributions of the two sensors.

In the original LC method, given a disparity map provided by a stereo algorithm, the overall accuracy is improved by propagating, within an active support centered on each point f of the initial disparity map, the plausibility $\mathcal{P}_{f,g}(d)$ of the same disparity assignment made for the central point by other points g within the active support. Specifically, the clues deployed by LC to propagate the plausibility of disparity hypothesis d are the color and spatial consistency of the considered pixels:

$$\mathcal{P}_{f,g}(d) = e^{-\frac{\Delta_{f,g}}{\gamma_s}} \cdot e^{-\frac{\Delta^{\psi}_{f,g}}{\gamma_c}} \cdot e^{-\frac{\Delta^{\psi}_{f',g'}}{\gamma_c}} \cdot e^{-\frac{\Delta^{\omega}_{g,g'}}{\gamma_t}} \quad (8)$$

where f, g and f', g' refer to points in the left and right image respectively, Δ accounts for spatial proximity, Δ^{ψ} and Δ^{ω} encode color similarity, and γ_s, γ_c and γ_t control the behavior of the distribution (see [3] for a detailed description). For the experimental results these parameters have been set to $\gamma_s = 8$, $\gamma_c = \gamma_t = 4$. The overall plausibility $\Omega_f(d)$ of each disparity hypothesis is given by the aggregated plausibility for the same disparity hypothesis propagated from neighboring points according to

$$\Omega_f(d) = \sum_{g \in \mathcal{A}} \mathcal{P}_{f,g}(d). \quad (9)$$

For each point the plausibility originated by each valid depth measure is computed and these multiple plausibilities are propagated to neighboring points that fall within the active support. Finally, the overall plausibility accumulated for each point is cross-checked by comparing the plausibility stored in the left and right views and the output depth value for each point is selected by means of a winner-takes-all strategy. The LC approach has been extended in [2] to allow the fusion of two different disparity maps. In this case, for each point of the input image there can be 0, 1 or 2 disparity hypotheses, depending on which sensor provides a valid measurement. Although [2] produces reasonable results, it has the fundamental limitation that gives exactly the same relevance to the information from the two sources without taking into account their reliability.

In this paper we propose an extension to this approach in order to account for the reliability of the measurements of ToF and stereo described in Sections IV-B and V-B. In order to exploit these additional clues, we extend the model of [2] by multiplying the plausibility for an additional factor that depends on the reliability of the considered depth acquisition system, computed for each sensor in the

considered point, as follows:

$$\Omega'_f(d) = \sum_{g \in \mathcal{A}} \Big(P_T(g)\mathcal{P}_{f,g,T}(d) \; + \; P_S(g)\mathcal{P}_{f,g,S}(d) \Big) \quad (10)$$

where $P_T(g)$ and $P_S(g)$ are the confidence maps for ToF and stereo data respectively, $\mathcal{P}_{f,g,T}(d)$ is the plausibility for ToF data and $\mathcal{P}_{f,g,S}(d)$ for stereo data.

The proposed fusion approach implicitly addresses the complementary nature of the two sensors. In fact, in uniformly textured regions, where the stereo range sensing is quite inaccurate, the algorithm should propagate mostly the plausibility originated by the ToF camera. Conversely, in regions where the ToF camera is less reliable (e.g. dark objects), the propagation of plausibility concerned with the stereo disparity hypothesis should be more influential. Without the two confidence terms of (10), all the clues are propagated with the same weight, as in [2]. In this case an erroneous disparity hypothesis from a sensor could negatively impact the overall result. Therefore, the introduction of reliability measures allows us to automatically discriminate between the two disparity hypotheses provided by the two sensors and thus improve the fusion results.

The adoption of the proposed model for the new plausibility is also supported by the nature of the confidence maps, that can be interpreted as the probability that the corresponding disparity measure is correct. A confidence of 0 means that the disparity value is not reliable and in this case such hypothesis should not be propagated. The opposite case is when the confidence is 1, meaning a high likelihood that the associated disparity is correct. All the intermediate values will contribute as weighting factors. This definition is also coherent when a disparity value is not available, for example due to occlusions: the associated confidence is 0 and propagation does not occur at all. An interesting observation on the effectiveness of this framework is that Eq. (10) can be extended to deal with more than two input disparity maps, simply adding other plausibility terms for the new disparity clues and an associated confidence measures. Other families of sensors can be included as well, by simply devising proper confidence measures.

Both ToF and stereo disparity maps are computed at subpixel resolution, but the original LC algorithm [3] only produces integer disparities, therefore we propose an additional extension in order to handle subpixel precision. We consider a number of disparity bins equals to the number of disparities to be evaluated multiplied by the inverse of the desired subpixel resolution (i.e., we multiply by 2 to if the resolution is 0.5). Then, at every step the algorithm propagates the plausibility of a certain disparity by contributing to the closest bin. With this strategy, the computation time remains the same as in the original approach [3], [31] and only the final winner-takes-all step is slightly affected.

VII. Experimental Results

To evaluate the performance of the proposed algorithm, we used the dataset provided in [22], that at the time of this writing is the largest available collection of real world ToF and stereo data with ground truth. This dataset contains 5 different scenes acquired by a trinocular setup made of a Mesa SR4000 ToF range camera and two Basler video cameras. The ToF sensor has a resolution of 176×144 pixels while the color cameras one is 1032×778 pixels, which is also the output resolution of the proposed method. Calibration and ground truth information are also provided with the dataset. The different scenes contain objects with different characteristics that allow one to evaluate the proposed algorithm in challenging situations, including depth discontinuities, materials with different reflectivity and objects with both textured and un-textured surfaces. Scene 1 and 2 present piecewise smooth surfaces, ideal for the implicit assumption of stereo matching, but also reflective materials and textureless regions. Scene 3, 4 and 5 are more complex and also include curved and fuzzy surfaces.

The disparity maps of ToF and stereo vision system have been computed as described in Section IV and Section V. Fig. 2 shows the estimated disparity maps and the relative error maps. The interpolated data from the ToF measurements are shown in column 3, while the corresponding error map is in column 4. Columns 5 and 6 show the stereo disparity map estimated by SGM and the relative error map. The final disparity map and its relative error produced with the proposed fusion framework are shown in columns 7 and 8. ToF depth data clearly shows poor performance in proximity of the edges. Stereo vision has sharper edges but several artifacts are present and there are more regions without a valid disparity value (depicted in dark blue). The fusion algorithm reliably fuse the information coming from the two sensors providing a disparity map that in all the scenes has higher accuracy than each of the two systems considered independently.

Fig. 3 shows the confidence maps that are used in the fusion process: the first row shows the left color camera, the second row shows the ToF confidence map, and the third row shows the stereo one. Starting from the ToF confidence, the amplitude and intensity related term tends to assign lower confidence to the upper part of the table that is almost parallel to the emitted rays. Therefore the amplitude of the received signal is low, thus reducing the precision. This term also assigns a smaller confidence to farther regions, reflecting another well known issue of ToF data. ToF confidence is low for dark objects but measurement accuracy depends on the reflectivity of the surface at ToF IR wavelengths and the reflectivity can be different for objects looking similar to the human eye (i.e., the black plastic finger in scene 5 reflects more IR light than the bear's feet). In addition, the four corners of the image also have lower confidence, in agreement with the lower quality of

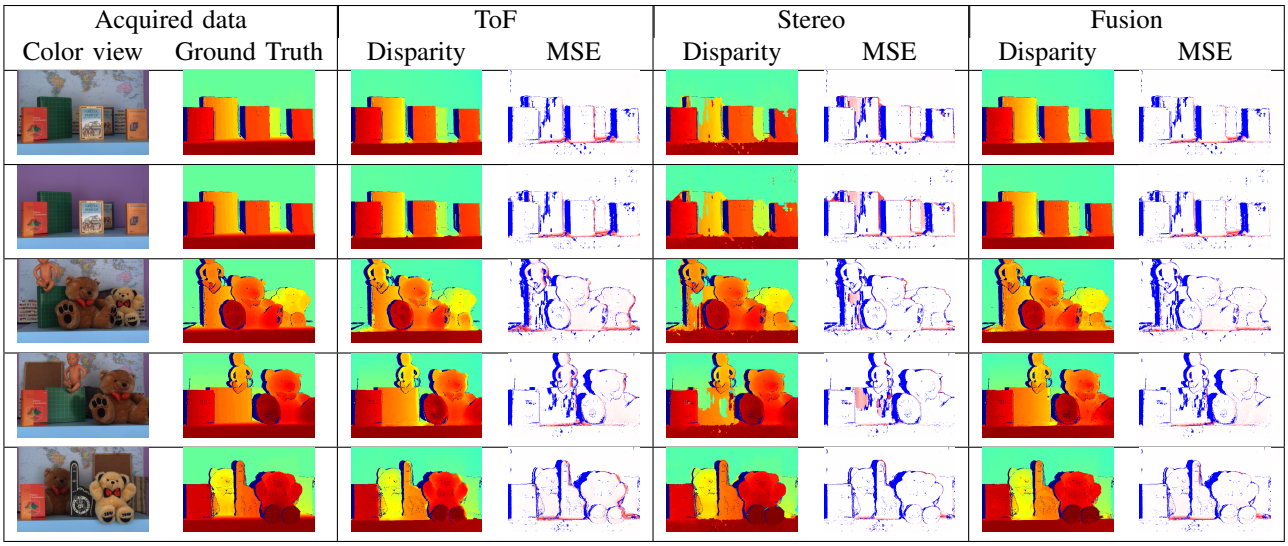


Fig. 2: Results of the proposed fusion framework. Each row corresponds to one of the 5 different scenes. Dark blue pixels correspond to points that have been ignored because of occlusions or because a ground truth disparity value is not available. The intensity of red pixels is proportional to the MSE. (*Best viewed in color*).

the signal in those regions, affected by higher distortion and attenuation. Local variance instead, as expected, contributes by assigning a lower confidence value to points near depth discontinuities.

Stereo confidence has on average a lower value, consistently with the fact that stereo data is less accurate (see Table I) but locally reflects the texture of the scene, providing high values in correspondence of high frequency content, and low values in regions with uniform texture (the blue table) or periodic pattern (e.g., the green book). Scene 2 compared to scene 1 clearly shows the effect that textured and untextured regions have in the confidence map. The map in the first scene is able to provide enough texture to consider reliable the depth measurements in that region. In the orange book on the left side, stereo confidence assigns high values only to the edges and to the logo in the cover, correctly penalizing regions with uniform texture. The teddy bear in scene 3, 4 and 5 has more texture than the table or the books and the relative confidence value is higher overall. The proposed stereo metric has been developed targeting the fusion of data from the two sensors, and low confidence is associated to textureless regions on purpose, even if the estimated depth is correct.

Table I compares the proposed approach with other state-of-the-art methods for which we got an implementation from the authors or we were able to re-implement. Since the output of the fusion process is a disparity map, we computed the error in the disparity space and considered the mean squared error (MSE) as the metric. For a fair comparison, we computed the error on the same set of valid pixels for all the methods, where a pixel is considered valid if it has a valid disparity value in all the compared maps and in the ground truth data. We also consider the ideal case obtained by selecting for each pixel the ToF or stereo disparity closer to the ground truth.

The average MSE has been calculated considering all the five scenes, and the results are reported in Table I. The disparity map of the proposed framework is compared with the estimates of ToF and stereo system alone and with the state-of-the-art methods of [2], [13], [23] and [22]. For the methods of [2] and [22] we obtained the results from the authors. The method of [22] has been computed from the ToF viewpoint at a different resolution, therefore we reprojected the data on the left camera viewpoint to compare it with other methods. We re-implemented the methods of [13] and [23] following the description in the papers. From the MSE values on the five different scenes, it is noticeable how the proposed framework provides more accurate results than the interpolated ToF data and the stereo measurements alone. Even if stereo data have typically lower accuracy the proposed method is still able to improve the results of the ToF interpolation, especially by leveraging on the more accurate edge localization of stereo data. The proposed approach also obtains a lower average MSE than all the compared methods. The average error is about 24% lower than [2], which is the best among the compared schemes. Conventional stereo confidence metrics of [10] produce an higher MSE if compared with our stereo metric, e.g., by using PKRN as confidence in the fusion framework the average MSE is 7.9. Our method has better performance than that of the compared schemes for all scenes except the very simple scene 2, in particular notice how it has a larger margin on the most complex scenes. This implies that our approach captures small details and complex structures while many of the compared approaches rely on low pass filtering and smoothing techniques which work well on simple planar surfaces but cannot handle more complex

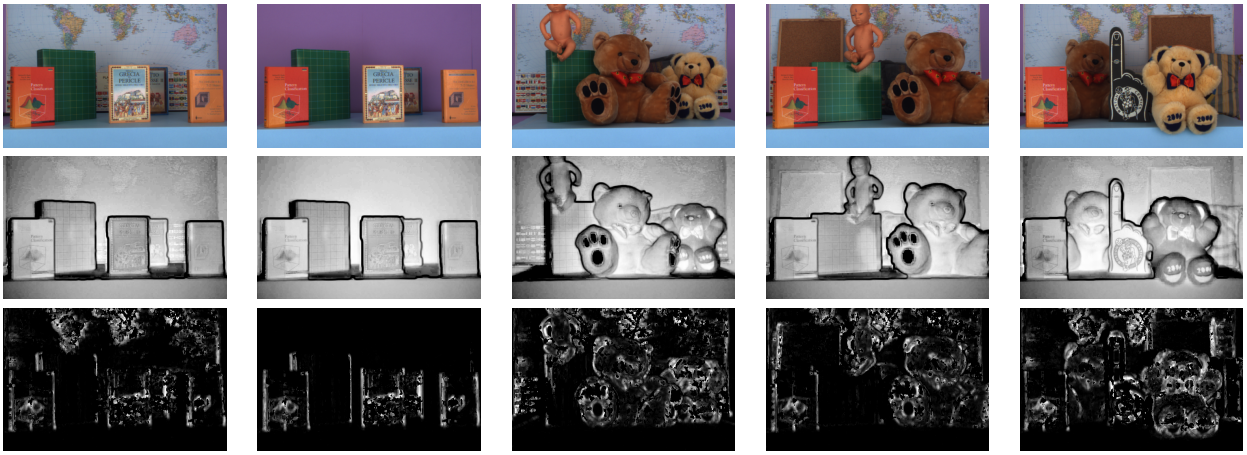


Fig. 3: Confidence maps for ToF and stereo disparity. Brighter areas correspond to higher confidence values, while darker pixels are less confident.

Scene	1	2	3	4	5	Avg.
ToF Int.	9.83	10.33	14.43	8.68	15.12	11.67
Stereo	19.17	27.83	18.06	25.52	11.49	20.42
Fusion	**7.40**	9.33	**6.92**	**6.30**	**8.39**	**7.67**
[2]	7.43	**9.27**	12.60	7.99	13.01	10.06
[13]	8.49	9.92	11.44	9.88	15.19	10.98
[23]	9.04	10.04	13.04	9.52	14.03	11.13
[22]	10.98	13.19	9.83	13.93	13.10	12.21
Ideal	2.50	2.60	3.22	2.42	3.16	2.78

TABLE I: MSE in disparity units with respect to the ground truth, computed only on non-occluded pixels for which a disparity value is available in all the methods.

situations. Enlarged figures and a more detailed analysis are available at http://lttm.dei.unipd.it/paper_data/eccv16.

VIII. Conclusions and Future Work

This paper presents a scheme for fusing ToF and stereo data exploiting the complementary characteristics of the two systems. Novel confidence models have been proposed for both ToF and stereo data. In particular, ToF reliability also considers the artifacts on edges caused by the finite size of the sensor pixels. Stereo confidence combines both local and global costs to consider the effects of global optimization steps. Finally, an extended version of the LC framework, including the reliability data and sub-pixel disparity estimation, has been proposed. Experimental results show how the proposed confidence metrics can be used to properly combine the outputs of the two sensors by giving more relevance to either of the two systems in regions where its depth estimation is more reliable, producing results that outperform state-of-the-art approaches.

Further research will be devoted to improve the proposed confidence metrics and to include other stereo vision algorithms in the proposed framework. In addition to passive stereo and ToF, also other depth camera's technologies will be properly modeled and considered in the fusion framework. Finally the use of a MAP-MRF formulation will be also considered in the data fusion.

Acknowledgments. Thanks to Arrigo Guizzo for some preliminary work.

References

[1] B. Tippetts, D. Lee, K. Lillywhite, and J. Archibald, "Review of stereo vision algorithms and their suitability for resource-limited systems," *Journal of Real-Time Image Processing*, pp. 1–21, 2013.

[2] C. Dal Mutto, P. Zanuttigh, S. Mattoccia, and G. Cortelazzo, "Locally consistent tof and stereo data fusion," in *Workshop on Consumer Depth Cameras for Computer Vision (ECCV Workshop)*. Springer, 2012, pp. 598–607.

[3] S. Mattoccia, "A locally global approach to stereo correspondence," in *Proc. of 3D Digital Imaging and Modeling (3DIM)*, October 2009.

[4] M. Hansard, S. Lee, O. Choi, and R. Horaud, *Time-of-Flight Cameras: Principles, Methods and Applications*, ser. SpringerBriefs in Computer Science. Springer, 2013.

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

[6] P. Zanuttigh, G. Marin, C. Dal Mutto, F. Dominio, L. Minto, and G. M. Cortelazzo, *Time-of-Flight and Structured Light Depth Cameras: Technology and Applications*, 1st ed. Springer

International Publishing, 2016. [Online]. Available: http://www.springer.com/book/9783319309712

[7] 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.

[8] 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.

[9] 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.

[10] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2121–2133, 2012.

[11] R. Nair, K. Ruhl, F. Lenzen, S. Meister, H. Schäfer, 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*, ser. Lecture Notes in Computer Science, M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb, Eds. Springer Berlin Heidelberg, 2013, vol. 8200, pp. 105–127.

[12] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *In Proc. of NIPS*. MIT Press, 2005, pp. 291–298.

[13] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[14] 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.

[15] S. Schwarz, M. Sjostrom, and R. Olsson, "Time-of-flight sensor fusion with depth measurement reliability weighting," in *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2014*, 2014, pp. 1–4.

[16] 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*, 2009.

[17] J. Dolson, J. Baek, C. Plagemann, and S. Thrun, "Upsampling range data in dynamic environments," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1141–1148.

[18] K.-D. Kuhnert and M. Stommel, "Fusion of stereo-camera and pmd-camera data for real-time suited precise 3d environment reconstruction," in *Proc. of Int. Conf. on Intelligent Robots and Systems*, 2006, pp. 4780 – 4785.

[19] 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.

[20] 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 3D Digital Imaging and Modeling (3DIM)*, October 2009.

[21] C. Dal Mutto, P. Zanuttigh, and G. Cortelazzo, "A probabilistic approach to ToF and stereo data fusion," in *Proc. of 3DPVT*, Paris, France, 2010.

[22] ——, "Probabilistic tof and stereo data fusion based on mixed pixels measurement models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2260–2272, 2015.

[23] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[24] 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.

[25] J. Zhu, L. Wang, R. Yang, J. E. 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, no. 7, pp. 1400–1414, 2011.

[26] 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 European Conference on Computer Vision Workshops (ECCVW)*, 2012.

[27] G. Evangelidis, M. Hansard, and R. Horaud, "Fusion of Range and Stereo Data for High-Resolution Scene-Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2178 – 2192, 2015.

[28] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1330–1334, 1998.

[29] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603 –619, 2002.

[30] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[31] S. Mattoccia, "Fast locally consistent dense stereo on multicore," in *6th IEEE Embedded Computer Vision Workshop (CVPR Workshop)*, June 2010.