Locally Consistent ToF and Stereo Data Fusion

Anonymous ECCV submission

Paper ID ***

Abstract. Depth estimation for dynamic scenes is a challenging and relevant problem in computer vision. Although this problem can be tackled by means of ToF cameras or stereo vision systems, each of the two systems alone has its own limitations. In this paper a novel framework for the fusion of 3D data produced by a ToF camera and a stereo vision system is proposed. Initially, depth data acquired by the ToF camera are up-sampled to the spatial resolution of the stereo vision images by a novel super-resolution algorithm based on image segmentation and bilateral up-sampling. In parallel a dense disparity field is obtained by a stereo vision algorithm. Finally, the up-sampled ToF depth data and the disparity field provided by stereo vision are synergically fused by enforcing the local consistency of depth data. The depth information obtained with the proposed framework is characterized by the high resolution of the stereo vision system and by an improved accuracy with respect to either the ToF and the stereo measurements. Experimental results clearly show how the proposed method is able to outperform the compared fusion algorithms.

1 Introduction

Depth estimation for dynamic scenes is a classic computer vision problem and it has attracted a large amount of research. Many solutions have been proposed for this problem including stereo vision systems, Time-of-Flight (ToF) cameras and light-coded cameras (such as Microsoft Kinect). Concerning stereo vision systems, in spite of the fact that recent research in this field has greatly improved the quality of the estimated geometry, results are yet not completely satisfactory, specially in estimating the geometry of scenes with limited texture information (aperture problem) or with repetitive texture patterns. A detailed description of the state-of-the-art in this field can be found in [1]. The introduction of Time-of-Flight cameras and of light-coded cameras (e.g., Microsoft Kinect) is more recent. These sensors are active systems that emit an infrared illumination, and estimate 3D scene geometry by analyzing the reflection of such an illumination. The great advantage of such systems is their ability to robustly estimate in real time the 3D geometry of the scene. Their main limitations are their low spatial resolution, their inability to deal with low IR-reflective surfaces (e.g., dark or black surfaces), and the high level of noise in their measurements.

The characteristics of ToF and stereo data are somehow complementary, therefore the problem of their fusion has attracted a lot of interest in the last years. The overall goal of ToF and stereo data fusion is to combine the information of a ToF camera and a stereo system in order to obtain an improved 3D geometry that combines the best features of both subsystems, such as high resolution, high accuracy and robustness with respect to different scenes.

In this paper a method for the fusion of data coming from a stereo system and a ToF camera satisying this requirements is proposed. The proposed framework is constituted by three different steps:

- a first step, in which the depth data acquired by the ToF camera are upsampled to the spatial resolution of the stereo vision images by a novel super-resolution algorithm based on image segmentation and bilateral upsampling.
- a second step, that can be performed in parallel, in which a dense disparity field is obtained by means of a stereo vision algorithm.
- a third step in which the up-sampled ToF depth data and the stereo vision output are synergically fused by extending the *Local Consistency* (LC) approach proposed in [2].

Furthermore, even if in this paper the fusion of the data coming from a ToF camera and a stereo pair is considered, the proposed approach can be applied to other active depth sensors such as the Microsoft Kinect.

The paper is organized as follows: Section 2 reviews the state-of-the-art in the fusion of data from a ToF camera with data from a stereo pair. Section 3 reports an overall presentation of the proposed framework, Section 4 presents a novel method for the super-resolution of ToF data, Section 5 explains the choice of the considered stereo vision algorithm and Section 6 presents the proposed ToF and stereo data fusion approach based on the LC strategy. Section 7 demonstrates the effectiveness of the proposed method with experimental evidence and Section 8 finally draws the conclusions.

2 Related Work

In spite of the fact that ToF cameras reached the market only a few years ago, the fusion of information provided by ToF range cameras and by standard color cameras has already received considerable attention. The first attempt to combine a low resolution ToF range camera with a high resolution color camera in order to provide an high resolution depth map is presented in [3], where the authors adopt a Markov Random Field (MRF) approach, by defining a graphical model of the acquired frame that is minimized via conjugate gradient. The fundamental hypothesis of this method, as well as of most of the other methods that couple a ToF range camera with a color camera is that discontinuities in the range data and in the color images tend to align.

A considerably wide class of methods proposed in order to solve this problem is based on the bilateral filter [4], e.g. in [5] an approach based on bilateral filtering is proposed where the input depth map is used in order to build a 3D volume of depth probability (cost volume). The method of [5] can also be generalized to the case of two color cameras instead of only one. In order to reduce the computational burden of the iterative bilateral filtering on the cost volume, a hierarchical version of the bilateral filtering method is proposed in [6]. In [7], an interesting temporal extension of [5] to the case of non-synchronized distance measurements is proposed in order to treat the more general case in which range measurements and color information are not synchronized.

The approach of [8] is different from the other methods, because it explicitly imposes that range and color discontinuities are aligned, while in the other methods this hypothesis is accounted only implicitly. In [8] the high resolution color image is firstly segmented, and then an interpolation of the low resolution depth map samples is performed exploiting only the samples that lie in the same segment of the interpolated location.

Another possible approach is the synergic fusion of data from a ToF with a couple of color cameras, i.e., a stereo vision system. A first approach to this problem is [9], in which the depth map acquired by the ToF and the depth map acquired by the stereo pair are separately obtained and averaged. Another approach was proposed in [10] where the depth map acquired by the ToF is reprojected on the reference image of the stereo pair, it is then interpolated and finally used as initialization for the application of a dynamic programming stereo vision algorithm. A different method was proposed in [6]. After the upsampling of the depth map acquired by the ToF by a hierarchical application of bilateral filtering, the authors apply a plane-sweeping stereo algorithm to the acquisition volume defined with respect to the ToF reference frame. Finally the depth information acquired from the ToF and from the stereo are fused together by means of a confidence based strategy. In [11] a different method is proposed, based on a probabilistic formulation. The final depth map is recovered from the one acquired by the ToF and the one estimated with the stereo vision system by performing a MAP local optimization in order to increase the accuracy of the depth measurements. The method proposed in [12] is instead based on a global MAP-MRF framework solved by means of belief propagation. An extension of this method that takes into account also the reliability of the data acquired by the two systems has been proposed in [13].

3 Proposed Method

As previously stated, the considered acquisition system is composed of a ToF range camera and a stereo system. The two acquisition systems are jointly calibrated by means of the method proposed in [11], which provides an accurate estimation of the relative positions of the ToF sensor with respect to the stereo cameras. Other possible calibration methods such as [6, 12, 10], were not considered since they are characterized by higher calibration errors. The adopted calibration procedure firstly requires to calibrate and rectify the stereo pair. The intrinsic parameters of the ToF sensor are then estimated and finally the extrinsic calibration parameters between the two systems are estimated by the closed-form technique adopted in [11]. The calibration procedure computes the intrinsic parameters matrix of the rectified stereo cameras K_s, the intrinsic parameters matrix of the ToF camera K_t, the baseline of the rectified stereo pair b and the matrix that describes the relative rototranslation between the ToF camera reference frame and the stereo pair reference frame (associated with the left camera) $T_{t \rightarrow s}$. Once the overall 3D acquisition system is calibrated, it is possible to reproject the ToF depth measurements to the stereo pair reference frame.

The proposed algorithm is divided into three different steps:

1. Computation of a high resolution depth-map from the ToF data by reprojection of the low resolution depth measurements acquired by the ToF camera into the lattice associated with the left camera and interpolation of the visible points only (super-resolution step).
2. Computation of a high resolution depth-map by applying a stereo vision algorithm on the rectified images acquired by the stereo pair.
3. Locally consistent fusion of depth measurements obtained by the stereo vision algorithm and the up-sampled version of the data obtained by the ToF sensor by means of an extended version of the LC technique [2].

4 Super-resolution of ToF data

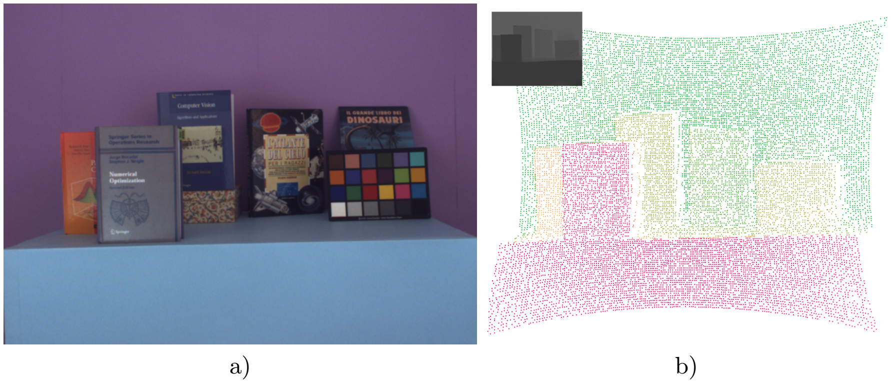
Different approaches have been proposed for the super-resolution of ToF data by exploiting auxiliary color images. In some previous works [14] bilateral filtering has been used to exploits color information in order to assist the depth interpolation. Other methods [8] exploit instead color data segmentation, which leads to high quality results, but they are prone to segmentation artifacts. In this work the sparse disparity measurements are interpolated by a novel interpolation method that exploits both these ideas in order to obtain better results. This allows to combine the good edge preserving quality of the segmentation-based method proposed in [8] and the good error reduction properties of the bilateral filter. The proposed super-resolution method is one of the major contributions of this paper. It is an improvement with respect to the state-of-the-art, since it allows to obtain high quality results like the method of [8] and at the same time it has the robustness of the approaches based on the bilateral filter like [14]. Over-smoothing effects typical of simply bilateral filter-basted methods and segmentation artifacts sensitivity of [8] are undesired effects that the proposed method allows to overcome.

The first step of the proposed method consists in the reprojection of the low resolution depth measurements acquired by the ToF camera into the lattice associated with the left camera and the interpolation of the visible points only, in order to obtain an high resolution depth map. In order to accomplish this step, all the 3D points $P_i^T, i = 1, ..., n$ acquired by the ToF camera are first projected onto the left camera lattice Λ_l (excluding the ones that are not visible from the left camera point of view) thus obtaining a set of samples $p_i, i = 1, ..., n$ over the left camera lattice. Note how the n samples acquired by the ToF camera cover only a small subset of the N samples of the lattice $\Lambda_l = p_j, i = 1, ..., N$

associated to the high resolution color camera. For example in our setup the ToF resolution is $n = 176 \times 144$, while the high resolution cameras used for the experimental results have a resolution of $N = 1032 \times 778$ pixels and n is just 3% of N. The data acquired by the ToF camera allow to associate to each non-occluded acquired sample p_i a depth value $z_i, i = 1, ..., n$ that can be mapped to a disparity value $d_i, i = 1, ..., n$ by the well known relationship between depth and disparity:

$$d_i = \frac{bf}{z_i} \tag{1}$$

where b is the baseline and f is the focal length of the rectified stereo system. This procedure makes available a set of sparse disparity measurements on the lattice associated to the left camera of the stereo pair, as shown in the example of Fig. 1.



a) b)

Fig. 1. Example of sparse disparity measurements: a) cropped color image framing the acquired scene; b) disparity data acquired by the ToF camera reprojected on the lattice associated to the left camera (the depth map acquired by the ToF camera is shown in the upper left corner at its original size).

The goal of the proposed interpolation method is to associate to all the points of the lattice Λ_l a disparity value $\tilde{d}_j, j = 1, ..., N$. Such a value is simply $d_i, i = 1, ..., n$ for the fraction of points of Λ_l that have an associated disparity value from the reprojected ToF depth measurements but it must be computed by interpolation for all the other points of Λ_l, i.e., the majority of the points in Λ_l. In order to accomplish this, the color image acquired by the left camera is firstly segmented using the method based on mean-shift clustering proposed in [15] thus obtaining a segmentation map $S(p_j), j = 1, ..., N$ that maps each point of Λ_l to the corresponding region.

In order to achieve a low noise level in the upsampled depth map and at the same time preserve edges a novel interpolation scheme is here introduced. In the following step a window W_j of size $w \times w$ centered on each of the p_j

samples that does not have a disparity value already available is considered for the computation of the estimated disparity value $\tilde{d}_j$. The samples that already have a disparity value from the ToF measures will instead just take that value. The set of points inside the window can be denoted with $p_{j,k}, k = 1, ..., w^2$ and finally $W'_j \subset W_j$ is the set of the points $p_{i,k} \in W_j$ with an associated disparity value d_i. In standard bilateral filtering [4] the interpolated disparity of point p_j is computed as the weighted average of the disparity values in W'_j where the weights are computed by exploiting both a weighting function in the spatial domain and one in the range domain. In the proposed approach we employ a standard 2D Gaussian function as in [4] for the spatial domain weighting function $f_s(p_{i,k}, p_j)$. The range domain function $f_c(p_{i,k}, p_j)$ is also a Gaussian function but it is not computed on the depth itself, but instead we computed it on the color difference between the two samples. Furthermore, in order to exploit segmentation information to improve the performance of the bilateral filter, also a third indicator function $I_{segm}(p_{i,k}, p_j)$ defined as:

$$I_{segm}(p_{i,k}, p_j) = \begin{cases} 1 \text{ if } S(p_{i,k}) = S(p_j) \\ 0 \text{ if } S(p_{i,k}) \neq S(p_j) \end{cases} \tag{2}$$

is introduced. The interpolated depth values are finally computed as:

$$\begin{aligned} \tilde{d}^j_s = & \sum_{W'_j} [f_s(p_{i,k}, p_j) I_{segm}(p_{i,k}, p_j) d_{i,k} + \\ & f_s(p_{i,k}, p_j) f_c(p_{i,k}, p_j)(1 - I_{segm}(p_{i,k}, p_j)) d_{i,k}] \end{aligned} \tag{3}$$

Note how the proposed interpolation scheme act as a standard low-pass interpolation filter inside each segmented region while samples that are outside the region are weighted on the basis of both the spatial and range weighting functions thus getting a lower weight, specially if their color is also different from the one of the considered sample. The output of the interpolation method is a disparity map $D_{t,s}$ defined on the lattice Λ_l. For clarity's sake let us emphasize that in our case about the 3% of the values of $D_{t,s}$ is obtained by reprojection of depth measurements by the ToF camera, while the remaining 97% is obtained by the proposed interpolation scheme. The proposed scheme offers an attractive novel super-resolution method because it couples the precision of segmentation-based methods [8] with the edge-preserving noise reduction capability of bilateral filter weighting [14]. Moreover, since the proposed method does not only take into account the samples inside the regions, this approach is also robust with respect to segmentation artifacts, i.e., the main drawback of [8]. The image shown in Fig. 2 is obtained by applying the proposed method on the sparse disparity map of Fig. 1. It is interesting to notice the improvements with respect to the results obtained with method of [14] based on bilateral up-sampling (that produces over-smoothing) and with respect to the one of [8], that is prone to segmentation artifacts.

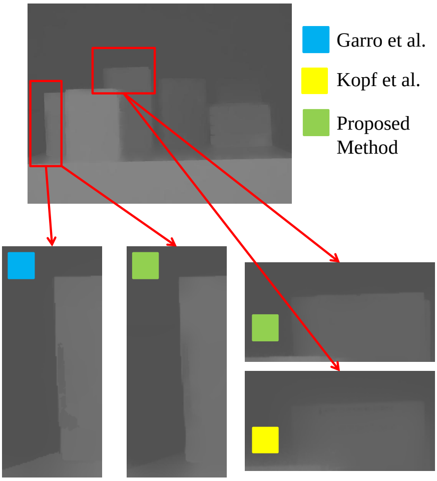


Fig. 2. Example of disparity measurements acquired by the ToF camera up-sampled to the lattice associated to the left camera. The full disparity is obtained by applying the proposed super-resolution method. In the zoomed pictures, there is a comparison of the results obtained applying the proposed method (green marker), the results of the application of the segmentation-based approach of [8] (blue marker) and the results of the direct application of bilateral filtering as proposed in [14] (yellow marker).

5 Stereo vision disparity estimation

Since our setup includes two calibrated color cameras, an additional high resolution disparity map D_s on lattice Λ_l can be inferred by means of stereo vision. The data fusion algorithm presented in the next section is independent of the choice of the stereo vision algorithm, therefore any stereo vision algorithm is potentially suited to extract the disparity map D_s. For our experiments we adopted the *Semi Global Matching* (SGM) algorithm proposed in [16]. This method infers disparity by performing multiple *scanline optimizations* (SO) on 1D domains so as to obtain a good trade-off between accuracy and computational complexity. SGM minimizes along each of the 8 or 16 scanlines an energy term made of a point-wise matching cost and of a regularization term that enforces smoothness within the 1D domain (i.e. each scanline). Each SO can be computed efficiently in polynomial time. Summing up, for each point and for each disparity hypothesis, the energy provided by each independent SO allows to avoid the well known streaking effect arising with a single SO. A detailed description of the SGM algorithms can be found in [16]; for our experiments we used the implementation available in the OpenCV library.

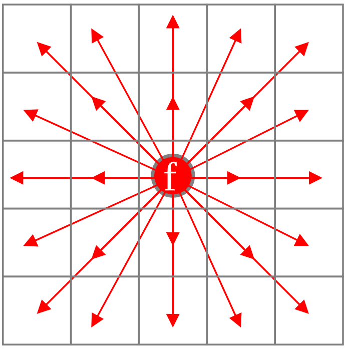


Fig. 3. Propagation of the *plausibility* for the disparity hypothesis $d(f)$ to the neighboring points.

6 Fusion of stereo and ToF disparity

Given the depth maps provided by an active ToF camera and a passive stereo vision system we aim at combining the potentially multiple range hypotheses available for each point by means of a technique that enables to obtain a locally consistent depth field. Our method extends the *Locally Consistent* technique (LC) [2] proposed for stereo matching so as to deal with the (potentially) twice disparity hypotheses available with our setup.

Given a dense disparity field provided by a stereo matching algorithm, the original LC technique enabled to improve the overall accuracy by propagating, within a patch referred to as *active support* centered on each point f of the initial disparity field, the *plausibility* of the same disparity assignment made for the central point f to each point g within the parch. As depicted in Figure 3, for an active support of size 5×5 pixels centered in f, the plausibility $P_{f \rightharpoonup g}(d(f))$ propagated from f to each point g within the active support is computed according to the image content [2]. Specifically, the cues deployed by LC to propagate the plausibility within the active support centered in f at a given disparity hypothesis $d(f)$ are the the color intensity of each point in the reference and the target image with respect to the corresponding central point of the active support, the matching cost for the assumed disparity hypothesis and a prior constraint related to the Euclidean distance of the examined point with respect to the center f of the active support. Therefore, after propagating this information, the *overall plausibility* of each disparity hypothesis is given by the amount of plausibility for the same disparity hypothesis received from neighboring points.

In this paper, we extend the LC approach in order to deal with the multiple input range fields provided by the active and the passive range measurement available in our setup. It is worth noting that, in this circumstance, for each point of the input image we can have 0 (both sensors don't have a potentially valid range measurement), 1 (only one of the two sensors provides a potentially valid range measurement) or 2 disparity hypotheses (both sensors provide a potentially, yet not necessarily equal, valid range measurement). Our method, for each point of the reference image with at least one range measurement computes, within an active support of size 39 × 39 and with the same strategy proposed in [2], the plausibility originated by each valid range sensor and propagates this potentially multiple plausibility to neighboring points that falls within the active support. It is worth noting that with 39 × 39 active support each valid disparity assignment provided by the ToF camera or the stereo sensor is propagated to 1521 neighboring points. Therefore, with this strategy, in the optimal case (i.e. when both range measurements for the examined point f are available) we are able to propagate within 1521 neighboring points the plausibility of the two disparity hypotheses originated by both sensors in f. When only a single sensor provides a valid range measurement for f we propagate its plausibility to 1521 neighboring points according to the unique valid hypothesis. Finally, when the point f under examination has not a valid range measurement we do not propagate any plausibility at all towards neighboring points. Nevertheless, it is worth observing that in this latter case, as well as in the other two former scenarios, one point receives several plausibilities from neighboring points if there are neighboring points (i.e. valid range measurements provided by ToF or stereo vision) within the size of the active support that propagated the plausibility of their disparity hypotheses. In most cases the depicted scenario is verified in practice. Once accumulated the overall plausibility for each point incoming from neighboring points according to the described strategy, for each point and for each hypothesis, we cross-check and normalize the overall plausibility similarly to [2]. Finally, we select for each point by means of a simple winner-takes-all strategy the disparity hypothesis with the highest overall plausibility.

It is worth to note that 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 (and partially filtered-out, in our experiments, enforcing the left-right consistency check), our approach propagates only plausibility originated by the ToF camera. Conversely, in regions where the ToF camera does not provide reliable information (e.g. dark objects) we propagate the plausibility of the disparity hypotheses provided by the stereo sensor. Of course, in regions with both range measurements we propagate the plausibility originated by both sensors.

7 Experimental Results

In order to evaluate the performance of the proposed algorithm we used an acquisition system made by a Mesa SwissRanger SR4000 ToF range camera with a resolution of 176 × 144 pixels and by two Basler scA1000 video cameras (with a resolution of 1032 × 778 pixels) synchronized in hardware with the ToF camera. Such a system can collect about 15 fps in a synchronized way, so there is no need for non-synchronized methods, such as the one proposed in [7]. The

Disparity map	MSE Scene a)	MSE Scene b)	MSE Scene c)	Average MSE
Proposed (ToF Interp.)	7.60	10.98	**7.08**	8.56
SGM stereo [16]	17.79	38.10	86.36	47.42
Proposed (ToF+Stereo)	**3.76**	**6.56**	8.69	**6.34**
Kopf et al. [14]	14.98	27.69	13.19	21.95
Garro et al. [8]	13.07	27.91	12.95	18.36
Yang et al. [5]	15.18	28.12	15.72	19.67

Table 1. MSE with respect to the ground truth: (*first row*) for the interpolated disparity map from the ToF depth measurements, (*second row*) for the disparity map calculated with the SGM stereo vision algorithm, (*third row*) for the final disparity map calculated after the data fusion, (*fourth row*) for the application of method [14], (*fifth row*) for the application of method [8] and (*sixth row*) for the application of method [5]. All the MSEs calculated for scene a), scene b) and scene c) are reported in the first three columns of the table. In the last column, the average MSE on the three scenes is reported. The MSE has been calculated only on non-occluded pixels for which a ground-truth disparity value is available.

system was calibrated with the method proposed in [11], thus obtaining a 3D reprojection error of about $5mm$ on the joint stereo and ToF calibration.

The proposed framework for the data fusion has been tested on different scenes acquired with the acquisition system described before. The results on three samples scenes are reported in Fig. 4. Note how scene a) and scene c) have a uniform background that is quite critical for stereo vision systems due to the lack of the textures (and in fact in row 4 many missing areas are visible) while scene b) has a texture pattern also on the background. For each of the acquired scene, an accurate disparity map has been acquired with an active space-time stereo system [17] integrating 600 images that has been considered as the ground-truth. The estimated disparity map with the interpolated data from the ToF measurements, the disparity map estimated with the SGM stereo vision algorithm and the disparity map obtained at the end of the proposed data fusion algorithm have been compared with the ground-truth disparity map.

The average *mean-squared-errors* (MSE) have been calculated for each of the three estimated disparity maps on each scene, and the results are reported in Table 1. In the table the proposed framework is also compared with the state-of-the-art methods of [14], of [8] and of [5]. In the last column of the table the average MSE of the estimated disparity maps on the three different scenes is also reported. From the MSE values on the three different scenes, it is immediate to notice how the proposed framework is capable of providing more accurate results than the interpolated ToF data and the stereo measurements. The results are also significantly better than the compared state-of-the-art methods on all the considered scenes. While concerning scene a) and b) it is immediately clear how the proposed method provides the best results, in scene c) it is the interpolation of the ToF measurements with the proposed method that provides the minimum MSE. This is due to the fact that this planar texture-less scenes constitute a simple case for the ToF depth measurements and a difficult case for stereo algorithms. This fact is reflected also on the high MSE value of the stereo vision system alone. However, as soon as a more complex scene geometry is considered (e.g., the puppet in scene a)) the results of the proposed fusion framework are superior to the single application of the interpolation algorithm on the ToF disparity measurements. In presence of more texture information (e.g., scene b)) the contribution of the stereo is relevant, and the final results of the data fusion algorithm halves the MSE if compared with the application of the interpolation algorithm on ToF data alone. Note also how the proposed method not only provides a lower MSE than the approaches of [14], [8] and [5], but also the improvement is very large in scenes a) and b) where both the stereo system and the ToF camera provides accurate information. This is a clear hint of the fact that the fusion algorithm is able to combine efficiently the two information sources. More detailed results are available in the additional material. All the datasets used in this paper will be made available on our website (currently removed for blind review).

The current implementation is not fully optimized and takes about 50 seconds. Nevertheless each component of the overall proposed method is well suited for a real-time GPU implementation. The current bottleneck is the *local consistency* data fusion step, that takes about 40*sec*.

8 Conclusions and future work

This paper presents a novel method for the synergic fusion of 3D measurements taken from two heterogeneous 3D acquisition systems (one active, the ToF, and one passive, the stereo system) in order to combine the advantages of both systems. There are two main contributions introduced in this paper. The first is a novel super-resolution method used as interpolation technique to up-sample the active sensor data that is able to combine precision near discontinuities, robustness against segmentation artifacts and edge preserving noise reduction. The second is the adoption of the *local consistency* framework in the context of heterogeneous sensors data fusion, i.e. an active sensor and a stereo vision system. The interpolation technique for the up-sampling of the active sensor data is “per se”a novel super resolution method capable to provide an high resolution depth map, very precise and robust with respect to errors in the depth measurements of both the active sensor and the stereo pair. The results obtained by the application of the proposed overall framework are always better than the results of the application of the compared methods. The proposed method is able to accurately estimate high resolution depth fields.

Furthermore the proposed framework is suited for a real-time GPU implementation, that is currently under development. Even though the method in this work is exemplified on an acquisition system made by a stereo pair and a ToF camera, we are considering its extension to different scenarios, e.g., to the case of a stereo pair and a structured light camera (e.g. Microsoft Kinect).

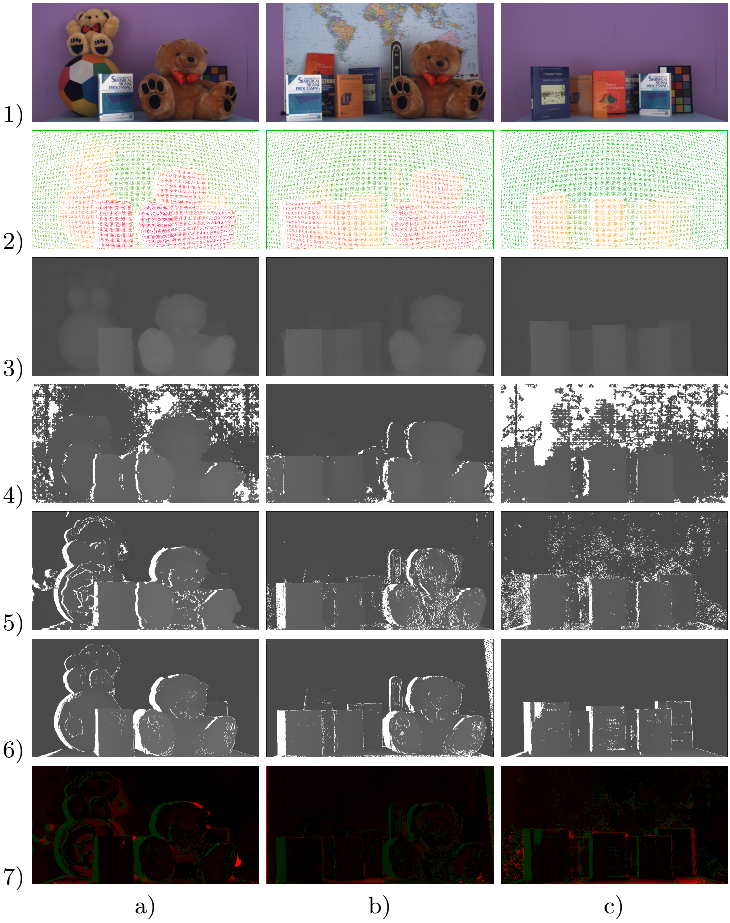


Fig. 4. Results of the proposed fusion framework. The columns correspond to the three different datasets on which the algorithm has been tested. Rows: 1) Cropped left image acquired by the left camera of the stereo pair; 2) Sparse disparity data acquired by the ToF camera and mapped on the left camera lattice (cropped); 3) Interpolated disparity map acquired by the ToF camera with the proposed interpolation framework (cropped); 4)Disparity map calculated with the SGM stereo vision algorithm (cropped); 5) Proposed locally consistent disparity map calculated from both ToF and stereo data (cropped); 6) Ground truth disparity map (cropped); 7) Difference between the final disparity map of row 5 and the ground truth (cropped). All the images have been cropped in order to account only for the pixels for which the ground truth disparity values are present. Green pixels in the last row correspond to points that have been ignored because occluded or because a ground truth disparity value is not available. In order to make the errors visible, the magnitude of the disparity errors (shown in red) have been multiplied by 10 in the images of the last row.

References

1. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision **47** (2001) 7–42
2. Mattoccia, S.: A locally global approach to stereo correspondence. In: Proc. of 3DIM. (2009)
3. Diebel, J., Thrun, S.: An application of markov random fields to range sensing. In: In Proc. of NIPS, MIT Press (2005) 291–298
4. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of the Sixth International Conference on Computer Vision. (1998)
5. Yang, Q., Yang, R., Davis, J., Nister, D.: Spatial-depth super resolution for range images. In: Proc. of CVPR. Volume 0., Los Alamitos, CA, USA, IEEE Computer Society (2007) 1–8
6. Yang, Q., Tan, K.H., Culbertson, B., Apostolopoulos, J.: Fusion of active and passive sensors for fast 3d capture. In: Proc. of MMSP. (2010)
7. Dolson, J., Baek, J., Plagemann, C., Thrun, S.: Upsampling range data in dynamic environments. In: Proceedings of CVPR, Los Alamitos, CA, USA (2010) 1141–1148
8. Garro, V., Dal Mutto, C., Zanuttigh, P., M. Cortelazzo, G.: A novel interpolation scheme for range data with side information. In: Proc. of CVMP, London, UK (2009)
9. Kuhnert, K.D., Stommel, M.: Fusion of stereo-camera and pmd-camera data for real-time suited precise (2006)
10. Gudmundsson, S.A., Aanaes, H., Larsen, R.: Fusion of stereo vision and time of flight imaging for improved 3d estimation. Int. J. Intell. Syst. Technol. Appl. **5** (2008) 425–433
11. Dal Mutto, C., Zanuttigh, P., Cortelazzo, G.: A probabilistic approach to tof and stereo data fusion. In: 3DPVT, Paris, France (2010)
12. Zhu, J., Wang, L., Yang, R., Davis, J.: Fusion of time-of-flight depth and stereo for high accuracy depth maps. In: Proc. of CVPR. (2008)
13. Zhu, J., Wang, L., Yang, R., Davis, J.E., Pan, Z.: Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps. IEEE Trans. on Pattern Analysis and Machine Intelligence **33** (2011) 1400–1414
14. Kopf, J., Cohen, M.F., Lischinski, D., Uyttendaele, M.: Joint bilateral upsampling. ACM Transactions on Graphics (Proceedings of SIGGRAPH 2007) **26** (2007)
15. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. Pattern Analysis and Machine Intelligence, IEEE Trans. on **24** (2002) 603 –619
16. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. on Pattern Analysis and Machine Intelligence (2008)
17. Zhang, L., Curless, B., Seitz, S.M.: Spacetime stereo: Shape recovery for dynamic scenes. In: Proc. of CVPR. (2003) 367–374
18. Microsoft: Kinect. http://www.xbox.com/en-US/kinect (2012)