Locally Consistent ToF and Stereo Data Fusion

Carlo Dal Mutto[1], Pietro Zanuttigh[1], Stefano Mattoccia[2], and Guido Cortelazzo[1]

[1] University of Padova, Padova, Italy
[2] University of Bologna, Bologna, Italy

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 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 up-sampling algorithm based on image segmentation and bilateral filtering. 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 the one produced by both subsystems. 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 challenging computer vision problem. 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 [1] in this field has greatly improved the quality of the estimated geometry, results are yet not completely satisfactory specially when the texture information in the scene is limited. The introduction of Time-of-Flight cameras and of light-coded cameras (e.g., Microsoft Kinect) is more recent. These systems are able to robustly estimate in real time the 3D geometry of the scene but they also have some limitations like low spatial resolution, the inability to deal with low reflective 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. 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 [2], where the authors adopt a Markov Random Field (MRF) approach. A considerably wide class of methods proposed in order to solve this problem is based on the bilateral filter [3], e.g. in [4] 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 [4] can also be generalized to the case of two color cameras instead of only one. The approach of [5] is different from the other methods, because it explicitly imposes that range and color discontinuities are aligned.

Another approach is the synergic fusion of data from a ToF with two color cameras, i.e., a stereo vision system. A first approach to this problem is [6], 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 [7] 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 stereo vision algorithm. In [8] 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 and finally a confidence based strategy is used for data fusion. In [9] 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 [10] 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 [11].

In this paper a method for the fusion of data coming from a stereo system and a ToF camera is proposed. The framework is constituted by 3 different steps: in the first step, the depth data acquired by the ToF camera are up-sampled to the spatial resolution of the stereo vision images by a novel up-sampling algorithm based on image segmentation and bilateral filtering. Then in the next step (that can be performed in parallel) a dense disparity field is obtained by means of a stereo vision algorithm. Finally in the third step the up-sampled ToF depth data and the stereo vision output are synergically fused by extending the *Local Consistency* (LC) approach [12].

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.

2 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 [9]. The adopted calibration procedure firstly requires to calibrate and rectify the stereo pair. The intrinsic param-

eters 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 [9]. Once the overall 3D acquisition system is calibrated, it is possible to reproject the ToF depth measurements to the stereo pair reference frame. Note how the setup is built in order to have a similar field of view for both the systems and the algorithm is applied on the region framed by both devices. The proposed algorithm is divided into 3 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 (up-sampling 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 [12].

In the rest of this section we will describe the steps 1 and 3, while for the second step we employed a standard stereo vision method from the literature (e.g. [13]).

3 Up-sampling of ToF data

In this work the sparse disparity measurements are interpolated by a novel interpolation method that exploits both segmentation and bilateral filtering in order to obtain better results. This allows to combine the good edge preserving quality of the segmentation-based methods and the good robustness of the bilateral filter. 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 $\Lambda_l = p_j, j = 1, ..., N$ associated to the high resolution color camera. 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 $d_i = bf/z_i$ (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 Fig. 1.

The goal of the proposed interpolation method is to associate to all the points of the lattice Λ_l a disparity value $\tilde{d}_j, j = 1, ..., N$. In order to accomplish this, the color image acquired by the left camera is first segmented using the method

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).

based on mean-shift clustering proposed in [14] thus obtaining a segmentation map $S(p_j), j = 1, ..., N$ that maps each point of Λ_l to the corresponding region. 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 [3] 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 [3] 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 in the CIELab space 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:

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

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

$$\begin{aligned} \tilde{d}_s^j = & \sum_{W'_j} [f_s(p_{i,k}, p_j) I_{segm}(p_{i,k}, p_j) d_{i,k} + \\ & f_s(p_{i,k}, p_j) f_c(p_{i,k}, p_j)(1 - I_{segm}(p_{i,k}, p_j)) d_{i,k}] \end{aligned} \tag{2}$$

Note how the proposed interpolation scheme acts 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. The proposed scheme offers an attractive novel up-sampling method because it couples the precision of segmentation-based methods [5] with the edge-preserving noise reduction capability of bilateral filter weighting [15]. 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. Fig. 2 shows an example of the results of the proposed approach and compares it with [15] and [5].

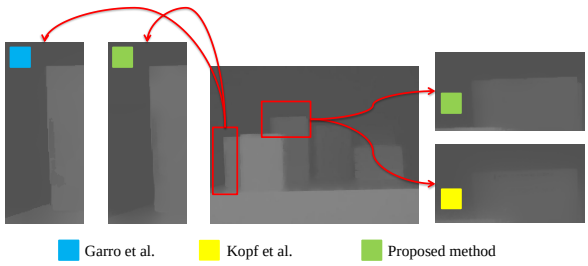

Fig. 2. Example of disparity measurements acquired by the ToF camera up-sampled to the lattice associated to the left camera. The full disparity is obtained by applying the proposed up-sampling method. In the zoomed pictures, there is a comparison of the results obtained applying the proposed method (green marker), the segmentation-based approach of [5] (blue marker) and the direct application of bilateral filtering as proposed in [15] (yellow marker).

4 Fusion of stereo and ToF disparity

After interpolating the ToF data, an additional high resolution disparity map D_s on lattice Λ_l can be inferred by means of stereo vision. Any stereo vision algorithm is potentially suited to extract the disparity map D_s, but for our experiments we adopted the *Semi Global Matching* (SGM) algorithm proposed in [13]. 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) [12] proposed for stereo matching so as to deal with the (potentially) two disparity hypotheses available with our setup.

Given a disparity field provided by a stereo algorithm, the original LC technique[3] 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 to

[3] A detailed description of the LC technique can be found in [12].

any other point within the active support. 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 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 [12], the plausibility originated by each valid range sensor and propagates this potentially multiple plausibility to neighboring points that falls within the active support. 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 39×39 neighboring points the plausibility of the two disparity hypotheses originated by both sensors in f. On the other hand, when only a single sensor provides a valid range measurement for f we propagate its plausibility to 39×39 neighboring points according to the unique valid hypothesis available. 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. Finally, we select for each point by means of a simple winner-takes-all strategy the disparity hypothesis with the highest overall plausibility.

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.

5 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 data at about 15 fps in a synchronized way, so there is no need for non-synchronized methods, such as the one proposed in [16]. The system was calibrated with the method proposed in [9], and we obtained a 3D reprojection error of about $5mm$ on the joint stereo and ToF calibration.

To test the proposed framework we acquired several different scenes. For space constraints we report here the results on three sample scenes only. Fig. 3 reports the results, note how the 3 scenes contains regions with different properties: e.g. scene a) and scene c) have a uniform background that is quite critical for stereo vision systems due to the lack of texture information (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 scenes, an accurate disparity map has been obtained by acquiring 600 images and processing them with an active space-time stereo system [17] 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 and with other state of the art methods.

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 [13]	17.79	38.10	86.36	47.42
Proposed (ToF+Stereo)	**3.76**	**6.56**	8.69	**6.34**
Kopf et al. [15]	14.98	27.69	13.19	21.95
Garro et al. [5]	13.07	27.91	12.95	18.36
Yang et al. [4]	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 [15], (*fifth row*) for the application of method [5] and (*sixth row*) for the application of method [4]. 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.

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 [15], of [5] and of [4]. 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 scene with a very limited amount of texture constitutes 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 [15], [5] and [4], 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 are available at the following url : `http://lttm.dei.unipd.it/downloads/tofstereo` .

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 $40sec$.

6 Conclusions and future work

This paper presents a novel method for the synergic fusion of 3D measurements taken from two heterogeneous 3D acquisition systems 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 ac-

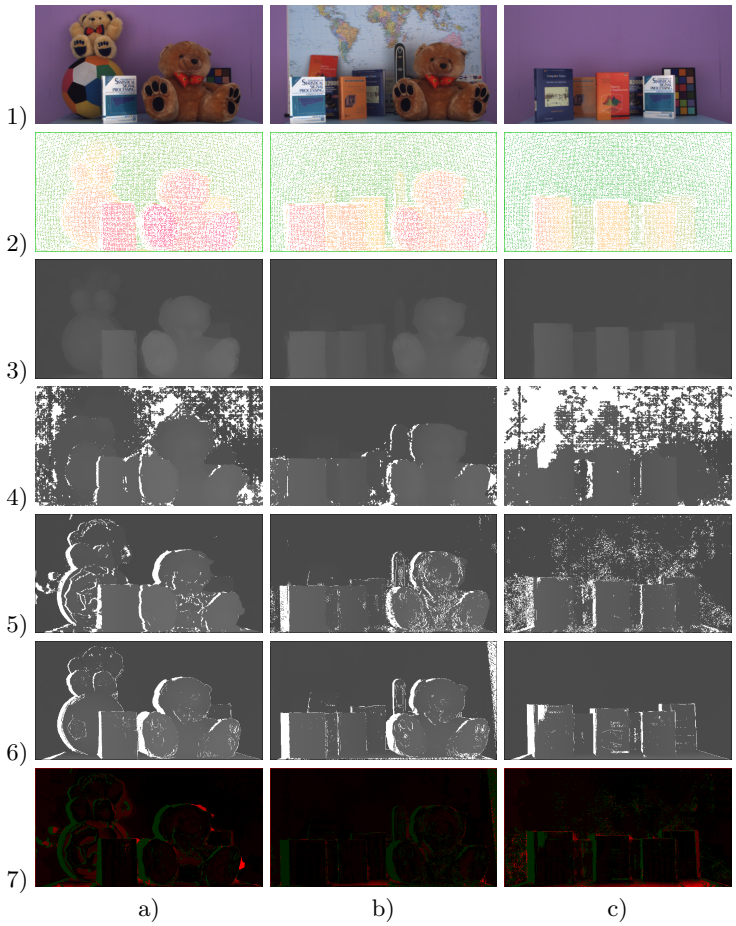

Fig. 3. Results of the proposed fusion framework. The columns correspond to the three different datasets on which the algorithm has been tested. Rows: 1) Cropped left image acquired by the left camera of the stereo pair; 2) Sparse disparity data acquired by the ToF camera and mapped on the left camera 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.

tive 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. 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).

References

1. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. Journal of Computer Vision **47** (2001) 7–42
2. Diebel, J., Thrun, S.: An application of markov random fields to range sensing. In: In Proc. of NIPS, MIT Press (2005) 291–298
3. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of the Sixth International Conference on Computer Vision. (1998)
4. Yang, Q., Yang, R., Davis, J., Nister, D.: Spatial-depth super resolution for range images. In: Proc. of CVPR. (2007) 1–8
5. Garro, V., Dal Mutto, C., Zanuttigh, P., M. Cortelazzo, G.: A novel interpolation scheme for range data with side information. In: Proc. of CVMP. (2009)
6. Kuhnert, K.D., Stommel, M.: 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) 4780 – 4785
7. 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
8. Yang, Q., Tan, K.H., Culbertson, B., Apostolopoulos, J.: Fusion of active and passive sensors for fast 3d capture. In: Proc. of MMSP. (2010)
9. Dal Mutto, C., Zanuttigh, P., Cortelazzo, G.: A probabilistic approach to ToF and stereo data fusion. In: 3DPVT, Paris, France (2010)
10. 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)
11. 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 Int. **33** (2011) 1400–1414
12. Mattoccia, S.: A locally global approach to stereo correspondence. In: Proc. of 3DIM. (2009)
13. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. on Pattern Analysis and Machine Int. (2008)
14. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. on Pattern Analysis and Machine Int. **24** (2002) 603 –619
15. Kopf, J., Cohen, M.F., Lischinski, D., Uyttendaele, M.: Joint bilateral upsampling. ACM Transactions on Graphics (Proceedings of SIGGRAPH 2007) **26** (2007)
16. Dolson, J., Baek, J., Plagemann, C., Thrun, S.: Upsampling range data in dynamic environments. In: Proceedings of CVPR. (2010) 1141–1148
17. Zhang, L., Curless, B., Seitz, S.M.: Spacetime stereo: Shape recovery for dynamic scenes. In: Proc. of CVPR. (2003) 367–374