Hybrid Kinect Depth Map Refinement For Transparent Objects

Gorkem Saygili Laurens van der Maaten Emile A. Hendriks
Computer Vision Lab, Delft University of Technology, The Netherlands
Email: {g.saygili, l.j.p.vandermaaten, e.a.hendriks}@tudelft.nl

***Abstract*—Depth sensors such as Kinect fail to find the depth of transparent objects which makes 3D reconstruction of such objects a challenge. The refinement algorithms for Kinect depth maps either do not address transparency or they only provide sparse depth on such objects which is inadequate for dense 3D reconstruction. In order to solve this problem, we propose a fully-connected CRF based hybrid refinement algorithm. We incorporate stereo cues from cross-modal stereo between IR and RGB cameras of the Kinect and Kinect's depth map. Our algorithm does not require any additional cameras and still provides dense depth estimations of transparent objects and specular surfaces with high accuracy.**

I. Introduction

The Kinect provides real-time and high resolution depth maps that are adequate for many tracking and object recognition applications [1]. However, the lack of its depth measurements on specular, absorbing and transparent surfaces affect many other tasks such as 3D reconstruction and virtual-view rendering. Such surfaces are common in everyday household objects of which depth can not be measured by the Kinect [2]. The recently introduced Kinect 2, available in 2014, provide high accuracy depth maps based on its time-of-flight (ToF) depth sensor. ToF sensors can also not measure depth on transparent objects [3]. In order to increase the accuracy of Kinect on challenging surfaces, many algorithms have been proposed. Most of those studies are based on different types of bilateral filters which smooths the depth image using the guidance of the color image and fills the unknown depth locations [5]–[9]. Using bilateral filtering for the inpainting of the Kinect depth maps can correct the missing depth values on specular and absorbing surfaces as long as there are sufficient depth measurements around the unknown locations. However, these algorithms fail to recover the depth of transparent objects since there is no depth information on transparent surfaces. Chiu *et al.* [2] proposed using cross-modal stereo between IR and RGB cameras of the Kinect to obtain depth cues for the transparent objects in Kinect depth maps. Their algorithm can find sparse depth estimations but is inadequate for dense 3D reconstructions. The main reason for sparsity is the structural difference of the IR and RGB images. In their later work [4], they achieve better results by learning a mapping between color channels of RGB image and IR image. However the resulting depth maps are still sparse and not adequate for dense representation of the transparent objects. None of these works considered pairwise inference between pixels such as global energy minimization to increase the accuracy on the challenging surfaces. There are other algorithms [10], [11] that use additional RGB cameras to develop a hybrid solution for recovering unknown depth values of Kinect depth maps. Using additional RGB cameras and sensors are not practical for many applications therefore we prefer to build a hybrid setup that is based on cameras and sensors of the Kinect only.

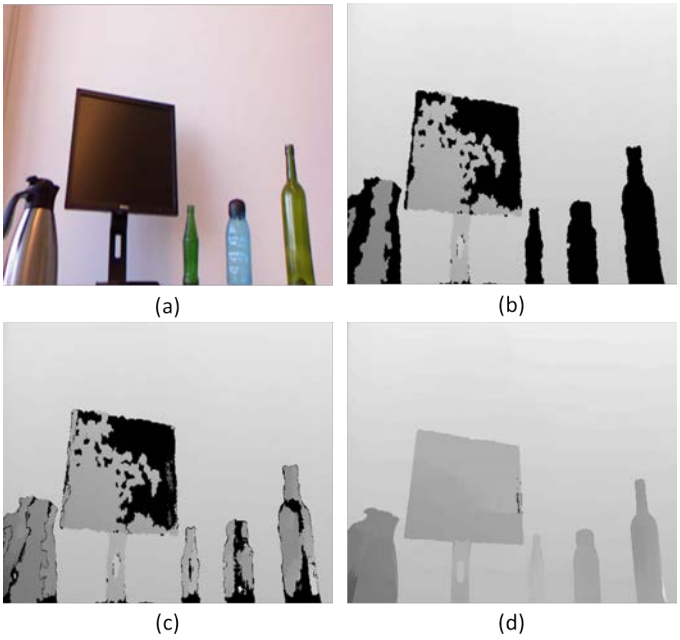

Fig. 1. Depth refinement results; (a) color image, (b) original depth, (c) cross-modal stereo result [4], (d) proposed algorithm result.

In this work, we propose a fully-connected CRF-based solution which is using cross-modal stereo and Kinect's depth measurements for dense depth recovery of transparent objects. The cross-modal stereo is a simple block matching approach that is applied on filtered IR and RGB images as in [4]. The fully-connected CRF model combines the information of stereo matching and Kinect's depth measurements with smoothness prior to recover unknown depth of the Kinect's depth map. Our algorithm can recover the depth of transparent objects as well as the depth of specular and absorbing surfaces. The resulting depth map can be used for accurate 3D reconstruction of challenging surfaces.

Our approach comprises two different stages that will be discussed in Section 2. Quantitative comparisons will be done in Section 3 and we draw our conclusions in Section 4.

II. MRF-Based Hybrid Depth Map Refinement

Our algorithm consists of two main steps: (1) cross-modal stereo between rectified IR and RGB images of Kinect, (2)

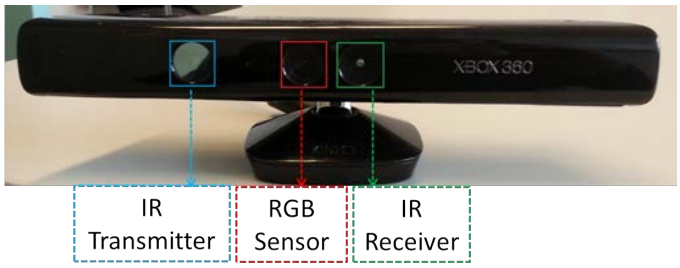

Fig. 2. Kinect: the distance between IR transmitter and receiver is 7.5 cm, the distance between RGB and IR receiver is 2.5 cm approximately.

CRF-based energy formulation and minimization. The first step produces stereo depth cues on transparent surfaces. The accuracy of stereo depth is not enough for dense depth estimations on these surfaces as mentioned in [2], [4] and shown in Fig. 3. The second step produces dense depth map by fusing the stereo cues from the first step, Kinect's depth measurements and spatial cues in a fully-connected CRF model. These steps are described below.

A. Cross-Modal Stereo

The Kinect provides three views: RGB view, depth view and IR view. IR and depth views are provided by the same camera which is not aligned with the RGB camera as depicted in Fig. 2. Similar to Chiu *et al.* [4], we first rectify IR and RGB views of the Kinect. Then we do cross-modal stereo matching between IR and RGB images. Rather than their linear filtering for increasing the similarity between IR and RGB, we incorporated rank transform [12] to calculate the cost for stereo. Rank transform is shown to be one of the most robust measures for stereo matching in terms of radiometric differences between stereo pairs [13] which increases the accuracy of stereo matching between IR and RGB cameras of the Kinect as depicted in Fig. 3 c-d. The erroneous estimations in Fig. 3.c are suppressed with rank transformation therefore the resulting stereo estimations are more accurate on challenging surfaces such as transparent objects.

Let $I(x, y)$ and $RT(x, y)$ denote the intensity and rank transform value of a pixel at (x, y) inside a local neighborhood $N(x, y)$, the Rank transform-based cost function at disparity d, $C_{Rank}(x, y, d)$, can be calculated as:

$$RT(x,y) = \left|\forall(x',y') \in N(x,y); I(x',y') < I(x,y)\right|,$$
$$C_{Rank}(x,y,d) = |RT(x,y) - RT'(x-d,y)|. \quad (1)$$

The resulting cost function is aggregated over a local patch to suppress the noise in the cost space:

$$C_{ste}(x,y,d) = \sum_{\forall(x',y')\in N(x,y)} C_{Rank}(x',y',d). \quad (2)$$

Even though Rank transform is robust against radiometric differences between two different sensors, the cost space comprises erroneous measures similar to ordinary stereo matching between RGB images. In order to suppress such errors, we incorporate a stereo confidence metric that is known as uniqueness ratio. Let $c_1(x, y)$ and $c_2(x, y)$ denote the minimum cost and second minimum cost for pixel at (x, y), respectively. In order to find the matches with high confidence, the ratio of the cost values should satisfy Eq. 3:

$$\mu_c(x,y) = \begin{cases} 1, & \frac{c_2(x,y)-c_1(x,y)}{c_1(x,y)} \geq \tau_u \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where τ_u is the uniqueness threshold. As depicted in Fig.3.c and d, our stereo result has fewer erroneous depth estimations compared to the result of [2]. However, our result is also sparse on transparent surfaces as as indicated with red boxes in Fig. 3. Additionally, the stereo estimations are not precise at the depth discontinuities which are the pixels around the red boundaries. In the second step of our algorithm, we proposed a fully-connected CRF based global energy minimization for fusing the stereo and Kinect depth estimations with piecewise smoothness prior to extend our sparse estimations into a dense depth representation of the scene with transparent objects.

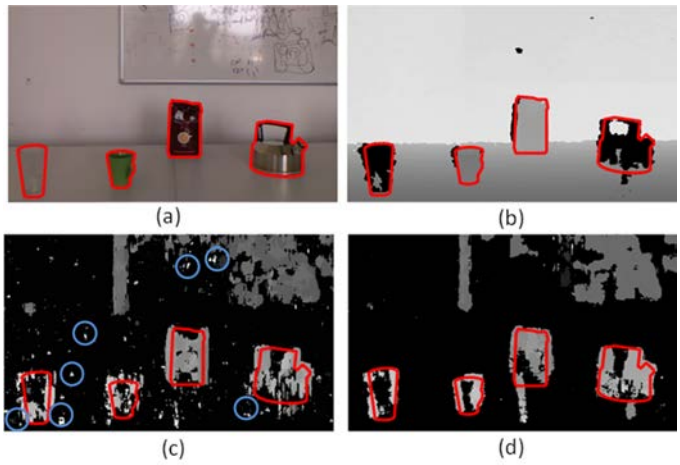

Fig. 3. Stereo matching results; (a) color image, (b) original depth, (c) stereo result using the filter proposed in [4], (d) stereo result using rank transform. Some of the erroneous estimations of [4] are indicated with blue.

B. Fully-Connected CRF Energy Model

Similar to multi-class image segmentation, estimating disparity of every pixels in an image can be formulated as maximum a posteriori (MAP) inference in a CRF and solved using highly efficient approximate inference algorithm.

In this paper, we formulated the energy of the CRF such that we fuse cross-modal stereo's and Kinect's estimations and incorporate global smoothness priors in a fully-connected model. A Fully-connected graph structure is preferred over 4 or 8-connected local grid structure since local inference usually over-smooths the edges (depth-discontinuities). Fig. 4 depicts the results for the 4-connected MRF [14] and our fully-connected CRF model where the borders of transparent objects are indicated in red. Both algorithms can produce dense depth estimations of the transparent objects. However, since cross-modal stereo and Kinect estimations are lacking precision at the discontinuities, 4-connected structure is not adequate to have accurate estimations at the transparent object borders. In contrast, fully-connected CRF enhance the quality using additional information from far-away pixels.

A fully-connected structure is computationally expensive compared to a locally connected structure. Recently,

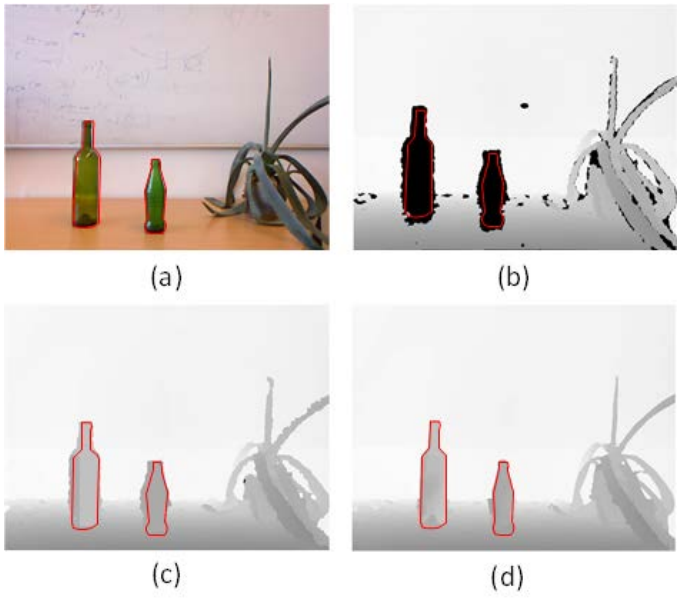

Fig. 4. The accuracy near depth discontinuities (encircled by red) of 4-connected MRF and fully-connected CRF: (a) Color image, (b) raw depth, (c) 4-connected MRF [14], (d) fully-connected CRF.

Krähenbühl *et al.* [15] proposed to use a linear combination of Gaussian kernels to approximate pairwise interactions in a fully-connected CRF model. The proposed algorithm provides accurate results with faster convergence compared to ordinary inference models.

The general energy function to minimize is composed of a unary, E_u, and a pairwise, E_p, terms. Let $\boldsymbol{x_i}$ denote the label for the pixel (x_i, y_i), the energy function of the CRF is defined as:

$$E(\boldsymbol{x}) = \sum_{\forall i} E_u(\boldsymbol{x}_i) + \sum_{\forall i<j} E_p(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{4}$$

The unary term is composed of stereo and Kinect estimations:

$$C_s(x,y,d) = \begin{cases} C_{ste}(x,y,d), & \mu_c(x,y) = 1 \\ \tau_{ste} * C_{ste}(x,y,d), & \text{otherwise} \end{cases} \tag{5}$$

$$C_k(x,y,d) = \begin{cases} 0, & |D(x,y) - d| < 1 \\ \tau_{kin}, & \text{otherwise}, \end{cases} \tag{6}$$

$$E_u(\boldsymbol{x}_i) = C_s(x_i, y_i, d_i) + C_k(x_i, y_i, d_i), \tag{7}$$

where $D(x,y)$ is the disparity measurement of the Kinect. The maximum disparity between IR and RGB cameras of the Kinect is 16 pixels because of the short distance between sensors as depicted in Fig. 2. In order to decrease quantization errors, the disparity cost for 16 disparities are interpolated to 256 using lowpass interpolation as depicted in Fig. 5. The cost of stereo estimations that have sufficient confidence defined by Eq. 3, are incorporated directly as unary energy. If the confidence of stereo is low, the stereo cost is penalized with τ_{ste}. The locations where Kinect has depth measurement, we convert and quantize the depth of Kinect to disparity of stereo for the new range of 256. The estimations that are not close to Kinect measurements are penalized with τ_{kin}.

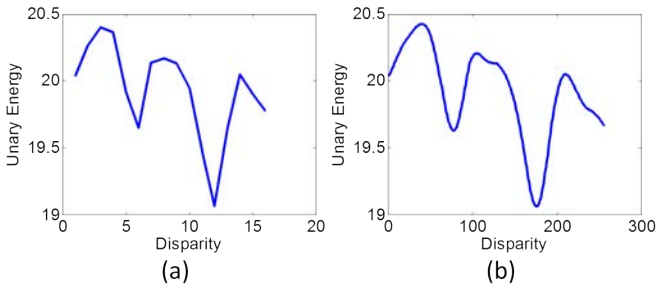

Fig. 5. Interpolation example; (a) the disparity range of 16 and unary energy before interpolation , (b) the disparity range of 256 and unary energy after interpolation

Similar to the model in [15], we use a Potts model that incorporates color similarity and spatial distance in our pairwise connections. Let p_i and I_i denote the spatial location and color of the ith pixel respectively:

$$\mu_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} 1, & \boldsymbol{x}_i \neq \boldsymbol{x}_j \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

$$E_{p1} = \exp\left(-\frac{|p_i - p_j|^2}{2\theta_s^2} - \frac{|I_i - I_j|^2}{2\theta_c^2}\right), \tag{9}$$

$$E_{p2} = \exp\left(-\frac{|p_i - p_j|^2}{2\theta_a^2}\right), \tag{10}$$

$$E_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mu_p(\boldsymbol{x}_i, \boldsymbol{x}_j)(w_1 E_{p1} + w_2 E_{p2}), \tag{11}$$

where the similarities are controlled by θ_s, θ_c, w_1 and w_2 respectively. $\mu_p(\boldsymbol{x_i}, \boldsymbol{x_j})$ denotes the Potts term for the energy function differences in labels between pairs that penalizes. E_{p1} in Eq. 9 denotes a bilateral function for the pairwise term in which color similarity and spatial distance between the pixels are considered with exponential terms for the inference. E_{p2} in Eq. 10 is the Gaussian smoothing prior that penalizes closer pairs of pixels that share different labels more strongly. w_1 and w_2 are the weighting parameters for E_{p1} and E_{p2}, respectively.

III. Experiments

We conduct several experiments to measure the performance of our algorithm. In our first experiment, we created six scenes with transparent objects that are not visible in Kinect's depth maps. The RGB and raw depth images of all of the created scenes are shown in Fig. 6.a-b. In each scene, there are various challenging objects with transparent, absorbing, specular surfaces and there are occlusion problems that occur because of the distance between IR transmitter and the receiver of Kinect as shown in Fig. 2.

As the first experiment, we present 3D reconstruction reconstruction performances of cross-modal stereo [2] and our algorithm in Fig. 7 a and b, respectively. The quality of the depth map has a direct influence on the quality of 3D reconstructions. Since the results of cross-modal stereo [2] contains erroneous and unknown depth values, the 3D reconstruction of the scene includes wrong voxels in 3D space as depicted in Fig. 7. In contrast, highly accurate dense depth estimations of such challenging objects with our algorithm increases the

TABLE I. PERCENTAGE OF ERRONEOUS DISPARITY VALUES OF PROPOSED ALGORITHM WITH TOP PERFORMING ALGORITHMS.

	Tsukuba			**Venus**			**Teddy**			**Cones**		
	nonocc	*all*	*disc*	*nonocc*	*all*	*disc*	*nonocc*	*all*	*disc*	*nonocc*	*all*	*disc*
Base	29.6	31.2	39.9	24.2	25.5	36	23.8	31.7	41.5	12.6	22.6	31.2
Stereo + Inpainting	11.5	13.2	26.5	7.39	8.75	26.7	13.3	21.8	30.0	7.01	16.2	19.1
Stereo + MRF	6.08	7.65	22.7	1.85	3.15	19.6	**10.1**	**16.9**	27.0	**6.49**	14.2	18.4
Proposed	**2.98**	**3.28**	**13.5**	**1.35**	**1.78**	**6.8**	10.6	17.2	**25.2**	7.27	**13**	**16.5**

Fig. 6. Refined depth maps; (a) color images, (b) original depth images, (c) our cross modal stereo (CMS) results with Rank transform, (d) CMS + inpainting results, (e) CMS + MRF results, (f) our algorithm results.

accuracy of 3D reconstruction significantly as depicted in Fig.7.b.

An alternative solution to obtain dense estimations on transparent objects might be combining cross-modal stereo with a bilateral filtering-based inpainting algorithm [5] or using similar energy functions as ours in 4-connected MRF graph structure. In our next experiment, we compare the performance of our algorithm with cross-modal stereo fused with MRF [14] and inpainting [5] as depicted in Fig. 6 f-d respectively. Different from inpainting, our algorithm and MRF can correct errors of the Kinect and cross-modal stereo rather than only estimating the unknown pixels. Furthermore, our algorithm provides the most precise object boundaries as depicted in challenging object boundaries for every images in Fig. 6.f. Both bilateral filtering and MRF failed to preserve sharp depth discontinuities because of their vulnerability to cross-modal stereo's and Kinect's imprecision at those locations as already discussed in Section 2.

Many of the Kinect inpainting algorithms have been tested using ground truth depth map and stereo images of Middleburry dataset [3], [6], [16]. In order to test the performance of our algorithm, we also used Middleburry dataset and compare the performance of our algorithm with inpainting and MRF approaches. To have a fair comparison, we only use stereo

Fig. 7. 3D Point cloud reconstructions for; (a) cross-modal stereo [2](b) our algorithm.

and spatial cues in all of the three algorithms. We discard erroneous disparities using Eq. 3 for the inpainting algorithm. The resulting accuracies of the algorithms are given in Table I. We checked the appearance of errors in different locations of the image such as non-occlusion (nonocc), all pixels (all) and locations close to disparity discontinuities (disc). The best results for different image regions are depicted in bold. In almost all of the challenging regions in all dataset images, our algorithm outperforms other algorithms with significant difference. Bilateral filtering with cross-modal stereo is the worst performer since bilateral filter has local inference and it does not incorporate the informative stereo cues from the matching cost of Eq. 1. The 4-connected local grid structure with MRF performs better than bilateral filtering but it is worse than our algorithm because of its insufficient local inference compared to fully-connected CRF. The results near depth discontinuity regions indicated with *disc* shows the significant improvement that is gained with our fully-connected CRF model over other algorithms. Fig. 9 show the results on Middleburry dataset. The accuracy of our algorithm is much higher than the accuracy of 4-connected MRF and inpainting especially at disparity discontinuities and homogeneous regions.

As the final experiment, we covered transparent objects in one of our images so that we observe the ground truth depth from Kinect measurements. We calculate the relative depth error as percentage by using Eq. 12. This experiment aims to calculate the accuracy inside the object rather than the performance at the borders. The results for cross-modal stereo (stereo), and our cross-modal stereo with bilateral filtering (Inp), MRF and our algorithm are given in Table II. Our algorithm outperforms other algorithms with approximately 5 percent error. The reason of having this error is mainly because of short distance between IR and RGB sensors of Kinect. Since the distance is short, cross-modal stereo fail to measure accurate depth estimations on low-textured regions especially for far-away objects.

The main reason for the sparsity of cross-modal stereo between IR and RGB sensors of the Kinect was mentioned to be the structural difference between IR and RGB data. However, the dense estimations can be obtained by incorporating spatial inference with global CRF as we showed in our experiments. However, we observe that the distance between the IR and RGB sensors of the Kinect limits the depth estimation range significantly.

In stereo vision, the distance between the cameras (baseline) affects the range of the disparity estimation. The cameras should be close to increase the overlap between the views in order to match sufficient amount of pixels for dense estimation. In contrast, the cameras should also be far from each other in order to estimate the depth of far-away objects. As an example the baseline between the eyes of a human (interocular distance) is about 6.3 cm. The baseline between Kinect sensors is around 2.5 cm. The main limitation of the cross-modal stereo is the insufficient distance between IR and RGB sensors of the Kinect for estimating the depth of long distance objects. As depicted in Fig. 2, the distance between IR receiver and RGB camera of the Kinect is about 2.5 cm which is not enough to estimate the depth of far-away objects. Fig. 8 depicts the range limits that we observed in our experiments. Since IR and RGB sensors are close, the maximum observable depth with cross-modal stereo is limited to 1.62 meters approximately.

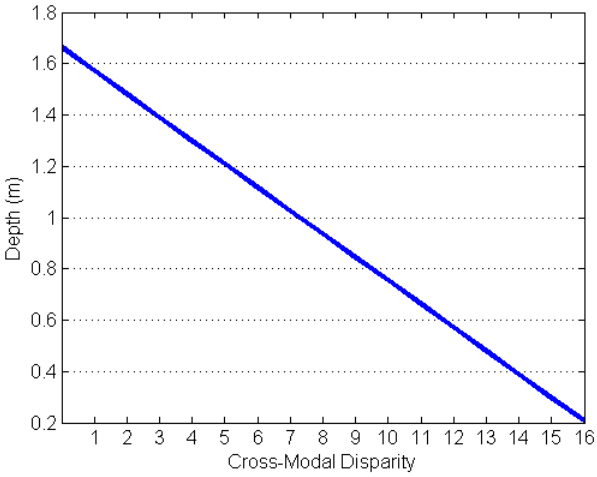

Fig. 8. Cross-modal stereo range. The maximum observable depth is bounded between $0 - 1.62$ meters.

TABLE II. RELATIVE DEPTH ESTIMATION RESULTS (PERCENT)

	Stereo [2]	Inp	MRF	Our Algorithm
Transparent	75.3	7.6	8.0	5.9

$$E_r = \sum_{\forall (x,y) \in R} \frac{|d_p(x,y) - d(x,y)|}{|R| d(x,y)}, \tag{12}$$

IV. CONCLUSION

In this paper, we proposed fully-connected CRF based hybrid Kinect refinement algorithm that achieves high accuracy depth estimations on any kind of challenging surfaces in a scene. Different from state-of-the-art algorithms, our algorithm provides dense depth estimations on transparent objects as well as precise accuracy near depth discontinuities. The improved depth maps can be used to perform accurate 3D depth reconstruction and depth-guided segmentation of the scenes. Further accuracy improvements can be achieved using additional pairwise priors and more accurate stereo matching algorithms. The main limitation of stereo matching between IR and RGB cameras in Kinect is the short baseline distance between the sensors.

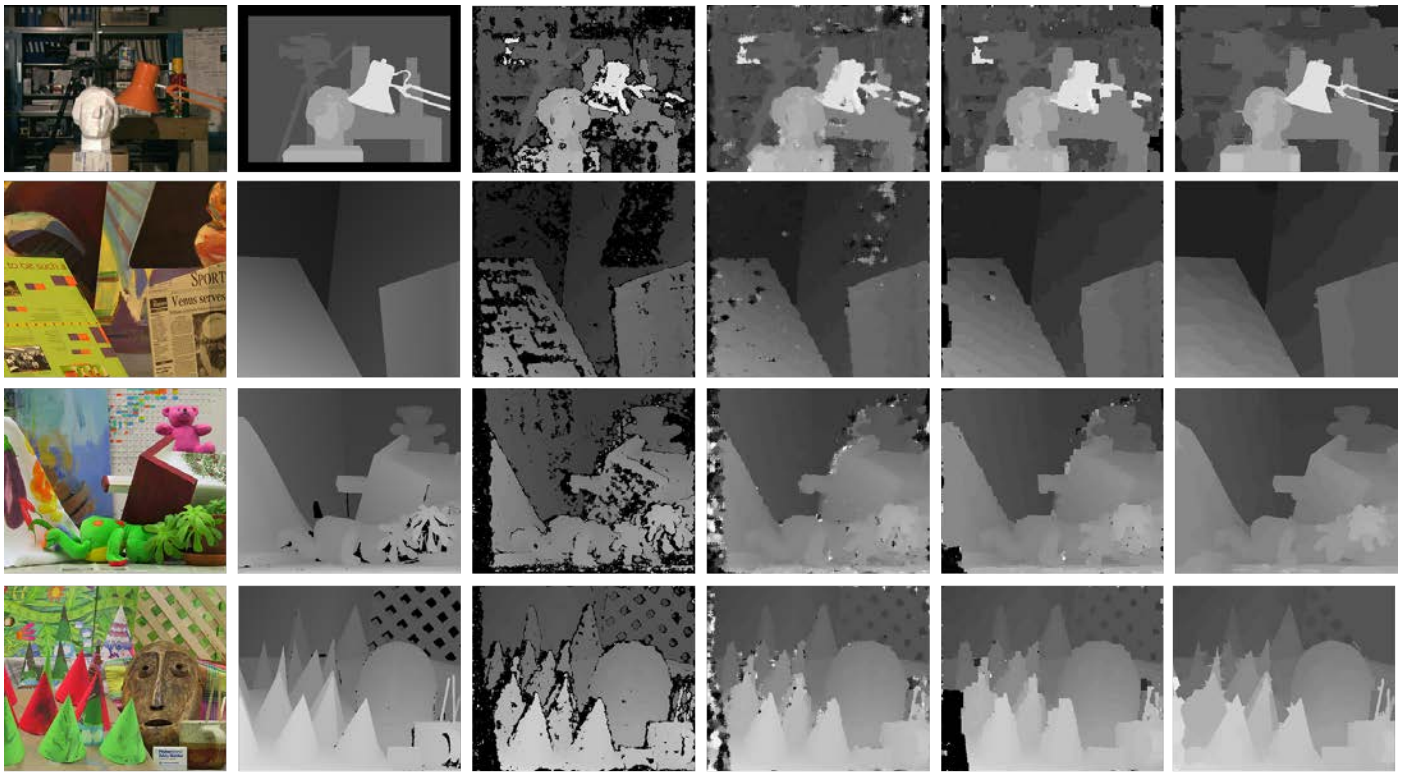

Fig. 9. Middleburry image results: (a) color image, (b) ground truth depth, (c) raw stereo matching results with filtered disparities, (d) inpainting, (e) 4-connected MRF, (f) fully-connected CRF refinement respectively.

V. Acknowledgement

This work was supported by the EU-AAL / ZonMW project SALIG++.

References

[1] L. Cruz, D. Lucio, and L. Velho, "Kinect and rgbd images: Challenges and applications," in *SIBGRAPI-T*, 2012, pp. 36–49.

[2] W. C. Chiu, U. Blanke, and M. Fritz, "Improving the kinect by cross-modal stereo," in *BMVC*, 2011, pp. 1209–1214.

[3] 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," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 7, pp. 1400–1414, 2011.

[4] W. C. Chiu, U. Blanke, and M. Fritz, "I spy with my little eye: Learning optimal filters for cross-modal stereo under projected patterns," in *ICCV Workshops*, 2011, pp. 1209–1214.

[5] F. Qi, P. Wang, G. Shi, and F. Li, "Structure guided fusion for depth map inpainting," *Pattern Recognition Letters*, 2012.

[6] V. Lazcano, P. Arias, G. Facciolo, and V. Caselles, "A gradient based nighborhood filter for disparity interpolation," in *ICIP*, 2012, pp. 873–876.

[7] L. Chen, H. Lin, and S. Li, "Depth image enhancement for kinect using region growing and bilateral filter," in *ICPR*, 2012, pp. 3070–3073.

[8] M. Camplani, T. Mantecon, and L. Salgado, "Accurate depth-color scene modeling for 3d contents generation with low cost depth cameras," in *ICIP*. IEEE, 2012, pp. 1741–1744.

[9] S. Lee and Y. Ho, "Real-time stereo view generation using kinect depth camera," in *in Proc. APIPA Annual Summit and Conference (APSIPA ASC)*, 2011, pp. 1–4.

[10] S. Zhang, C. Wang, and S. Chan, "A new high resolution depth map estimation system using stereo vision and kinect depth sensing," *Journal of Signal Processing Systems*, pp. 1–13, 2013.

[11] D.-Y. Chan and C.-H. Hsu, "Regular stereo matching improvement system based on kinect-supporting mechanism," 2012.

[12] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Computer VisionECCV'94*. Springer, 1994, pp. 151–158.

[13] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[14] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 508–515.

[15] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *arXiv preprint arXiv:1210.5644*, 2012.

[16] S. Zhang, C. Wang, and S. Chan, "A new high resolution depth map estimation system using stereo vision and kinect depth sensing," *Journal of Signal Processing Systems*, pp. 1–13, 2013.