Spatial-Depth Super Resolution for Range Images

Qingxiong Yang* Ruigang Yang* James Davis† David Nistér*
*University of Kentucky †University of California, Santa Cruz
http://vis.uky.edu/~liiton/

Abstract

We present a new post-processing step to enhance the resolution of range images. Using one or two registered and potentially high-resolution color images as reference, we iteratively refine the input low-resolution range image, in terms of both its spatial resolution and depth precision. Evaluation using the Middlebury benchmark shows across-the-board improvement for sub-pixel accuracy. We also demonstrated its effectiveness for spatial resolution enhancement up to $100\times$ *with a single reference image.*

1. Introduction

There exists a variety of range measuring technologies to acquire 3D information about our world. For example, laser range scanners can provide extremely accurate and dense 3D measurement over a large working volume [4, 5, 9, 11, 13, 15]. However, most of these high-quality scanners measure a single point at a time, limiting their applications to static environments only. The options to capture depth at video rate are rather limited: the main contender–stereo vision–is known to be quite fragile in practice.

Recently new sensors [1, 2, 16] have been developed to overcome this limitation. By using extremely faster shutter (on the order of nanosecond), these sensors measure time delay between transmission of a light pulse and detection of the reflected signal on an entire frame at once. While the technology is promising, in the current generation, these sensors are either very expensive or very limited in terms of resolution. For example the Canesta EP DevKit sensors can provide range images only up to 64×64. Their applications are therefore limited to background segmentation and user interface control.

In this paper we present a framework to substantially enhance the spatial and depth resolution of low-quality and highly quantized range maps, e.g., those from stereo vision or the Canesta sensor. Our approach takes advantage of the fact that a registered high-quality texture image can provide significant information to enhance the raw range map.

Most related to our work is a range-enhanced method by Diebel and Thrun [8], in which a Markov Random Field (MRF) is first designed based on the low resolution depth maps and high resolution camera images. The MRF is then solve with the well-known conjugate gradient (CG) algorithm [12]. This method gives promising spatial resolution enhancement up to $10\times$. Our formulation has demonstrated spatial resolution enhancement up to $100\times$. Although our work is in the multi-sensor fusion scope, we are different from most of the other approaches [10] because the resolutions of our sensors are quite different from each other.

Key to our success is the use of a bilateral filter [17], inspired by several state-of-the-art stereo algorithms [3, 18, 19]. In essence, we consider that the input range map provides a probabilistic distribution of depth, from which we can construct a 3D volume of depth probability, typically referred to as the *cost volume* in the stereo vision literature. Then we iteratively apply a bilateral filter to the cost volume. The output high-resolution range image is produced by taking the winner-takes-all approach on the weighted cost volume and a sub-pixel refinement afterward.

This simple formulation turns out to be very effective. As demonstrated with a variety of real-world objects, it can provide not only visually compelling range images up to $100\times$ resolution, but also a numerically more accurate depth estimate. We have applied our framework to *all* the algorithms reported on the Middlebury stereo benchmark site [14]. Our depth-enhanced disparity maps, when compared to their original counter parts, are *superior in overall ranking for each and every algorithm listed*, including those already having sub-pixel disparity refinement.

The paper is organized as follows: Section 2 presents an overview of our super resolution framework and the details about spatial resolution enhancement using a bilateral filter and depth resolution enhancement by quadric polynomial interpolation. In Section 3 we then discuss how to enhance the depth resolution for general two-view stereo vision problems through a sub-pixel refinement step. The experimental results are reported in Section 4, followed by a conclusion in Section 5.

1-4244-1180-7/07/$25.00 ©2007 IEEE

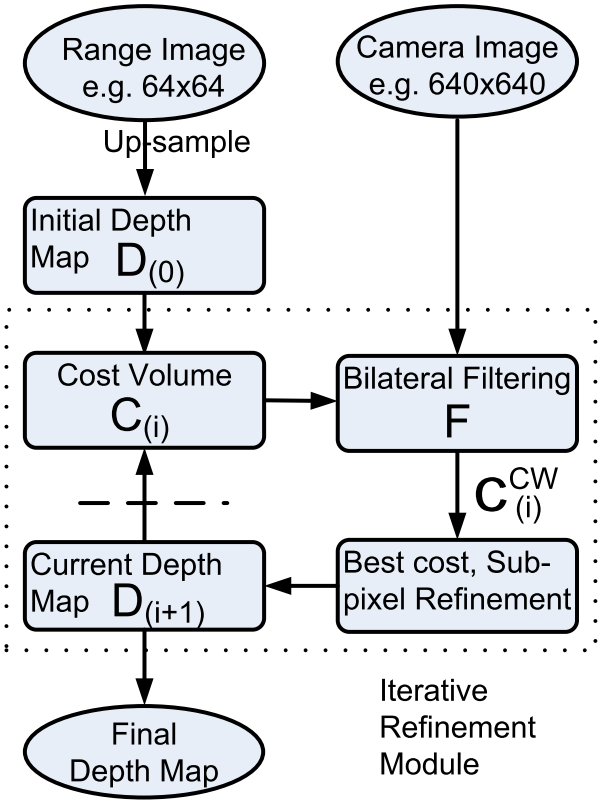


Figure 1. Framework of our post-processing approach. The range image is up-sampled to the same size as the camera image, and serves as the initial depth map hypothesis. The following is an iterative refinement process. A cost volume is built according to the current depth map hypothesis. A bilateral filter is then applied to the cost volume to handle the fattening problem near depth discontinuities. A winner-take-all and sub-pixel estimation procedure is used to produce a new depth map hypothesis, which is fed back into the process.

2. Approach

An overview of the framework of the approach is provided in Figure 1. First, up-sample the low-resolution depth map from the range image to the same size as the high-resolution camera image, save it as $D_{(0)}$. Then follows an iterative refinement module. A cost volume C_i is built based on the current depth map $D_{(i)}$, then a bilateral filtering is performed throughout each slice of the cost volume to produce the new cost volume $C_{(i)}^{CW}$. The refined depth map $D_{(i+1)}$ is generated based on this cost volume by first selecting the depth hypothesis with the minimal cost and a sub-pixel estimation afterwards.

2.1. Construction and refinement of the cost volume

At first, a coarse cost volume is built based on the current depth map. In order to allow large depth variations, as the current depth values are not guaranteed to be correct, the cost function should become constant as the differences become large. One such common function is the truncated quadric model, where the cost increases quadratically based on the distance between the potential depth candidate d and the currently selected depth $D_{(i)}(\mathbf{y},\mathbf{x})$

$$C_{(i)}(\mathbf{y},\mathbf{x},d) = \min(\eta * L, (d - D_{(i)}(\mathbf{y},\mathbf{x}))^2) \quad (1)$$

L is the search range, η is a constant. The square difference is selected as the cost function since we will use quadratic polynomial interpolation for sub-pixel estimation later. This cost function can help to preserve the sub-pixel accuracy of the input depth map.

Bilateral filtering is then applied to each slice of the cost volume based on the following prior assumptions:

- World surfaces are piecewise smooth.
- The pixels with similar colors around a region are likely to have similar depth.

Stereo matching based on bilateral filtering was first presented in [3], and then integrated into the stereo algorithm proposed in [19] which is one of the Middlebury top algorithms. The experimental results in both papers show that the bilateral filter works very well near discontinuities, which is the main challenge for spatial super-resolution discussed in this paper. For the smooth areas, after up-sampling, all the missed sampling areas are filled in correctly by interpolation. However, this is generally not true for the discontinuous areas. The missed sampling areas are blurred after up-sampling. But by using the color information provided by the registered camera images, we demonstrate that it is possible to get sharp/true depth edges for stereo spatial super resolution. This is the central theme of the paper.

The bilateral filter used in the paper is designed as following:

$$\begin{aligned}
F(\mathbf{y}+u,\mathbf{x}+v) &= f_c(W_c(\mathbf{y},\mathbf{x},u,v))f_s(W_s(u,v)), \quad (2)\\
f_c(\mathbf{x}) &= exp(-\frac{|\mathbf{x}|}{\gamma_c}),\\
f_s(\mathbf{x}) &= exp(-\frac{|\mathbf{x}|}{\gamma_s}),\\
W_c(\mathbf{y},\mathbf{x},u,v) &= \frac{1}{3}(|R(\mathbf{y}+u,\mathbf{x}+v) - R(\mathbf{y},\mathbf{x})|\\
&+ |G(\mathbf{y}+u,\mathbf{x}+v) - G(\mathbf{y},\mathbf{x})|\\
&+ |B(\mathbf{y}+u,\mathbf{x}+v) - B(\mathbf{y},\mathbf{x})|),\\
W_s(u,v) &= \sqrt{u^2+v^2}.
\end{aligned}$$

$\mathbf{y}$, $\mathbf{x}$ are the indices of the current pixel in the camera image, and u, v are two variables. R, G, B are the RGB channels of the camera image. γ_c and γ_s are two constants used as

the thresholds of the color difference and the filter size. The bilateral filter works as soft color segmentation in the super resolution framework, which aggregates the probabilities of each depth candidates of the pixels around a region based on the color similarity of the central pixel and its neighbors.

As it is shown in Figure 1, the bilateral filter is iteratively applied to the current cost volume to smooth the cost volume while preserving the edges, then we search through all the depth hypotheses and select the one with the minimal cost. Finally, sub-pixel estimation is performed based on the current cost volume and the depth hypotheses with the minimal cost.

2.2. Sub-pixel Estimation

To reduce the discontinuities caused by the quantization in the depth hypothesis selection process, a sub-pixel estimation algorithm is proposed based on quadratic polynomial interpolation. If the cost function is continuous, the depth with the minimum matching cost can be found. However, the cost function is discrete in practice. The search range is limited, which results in discontinuous depth maps. In order to eliminate this effect, we use quadratic polynomial interpolation to approximate the cost function between three discrete depth candidates: d, d_- and d_+. d is the discrete depth with the minimal cost, $d_- = d - 1$, and $d_+ = d + 1$.

$$f(x) = ax^2 + bx + c, \tag{3}$$

$$x_{min} = \frac{-b}{2a}, \tag{4}$$

$f(x_{min})$ is the minimum of function $f(x)$. Thus, given d, $f(d)$, $f(d_-)$ and $f(d_+)$, the parameters a and b of the continuous cost function can be calculated. Thus:

$$x_{min} = d - \frac{f(d_+) - f(d_-)}{2(f(d_+) + f(d_-) - 2f(d))}, \tag{5}$$

x_{min} is the depth with the minimum of the quadric cost function $f(x)$. Figure 2 provides a visual comparison of the depth maps and their synthesized views before and after sub-pixel estimation. Notice that the quantization effect on the man's face and the background on the synthesized view is removed after sub-pixel estimation.

3. Extended depth super resolution with two views

The main difference between one-view super resolution and two-view super resolution is the construction of the cost volume. In two view case, general stereo matching algorithm can be performed, together with the range image, to provide a more accurate cost volume. At first, three depth

(a) Depth maps. (b) Synthesized views.

Figure 2. (a) Depth maps generated with the **DoubleBP** algorithm [19] reported on the Middlebury website. (b) Synthesized views using (a). First row shows results without sub-pixel refinement, second row shows results with sub-pixel refinement. Notice that the quantization effect on the man's face and the background on the synthesized view before sub-pixel is removed after sub-pixel estimation.

candidates d, d_-, d_+ are computed from the input depth map for each pixel. d is extracted from the input depth map, $d_- = d - 1$ and $d_+ = d + 1$. To perform depth enhancement with two views, three slices of matching cost are calculated based on the three depth candidates. The calculation of matching cost is implemented according to the symmetric correlation approach presented in [19]. First, project the pixel in the reference view to the other view using the depth candidates calculated from the input depth map, and the matching cost is the pixel dissimilarity of the corresponding pixels. To reduce the noise, Birchfield and Tomasi's pixel dissimilarity [6] is used. Second, a symmetric bilateral filtering is applied to the cost slices:

$$F_{symm}(\mathbf{y}+u, \mathbf{x}+v) = F(\mathbf{y}+u, \mathbf{x}+v)F(\mathbf{y}'+u, \mathbf{x}'+v), \tag{6}$$

$F(\mathbf{y}+u, \mathbf{x}+v)$ is the filter defined in Equation 3, $\mathbf{y}$, $\mathbf{x}$ is the index of the current pixel in the reference view, and $\mathbf{y}'$, $\mathbf{x}'$ is the index of the corresponding pixel in the other view.

The sub-pixel depth enhancement is performed by a quadratic polynomial interpolation with the symmetric correlation volume as it is described in Section 2.2. Finally, an adaptive box-car filter (G) is applied to smoothen the depth map:

$$G(\mathbf{y}+u, \mathbf{x}+v) = \begin{cases} 1.0 & \text{if } |D_0(\mathbf{y}, \mathbf{x}) - D_0(u, v)| < 1 \\ 0 & \text{else} \end{cases}$$

D_0 is the input depth map. The size of the box-car used in this paper is relatively small (9x9).

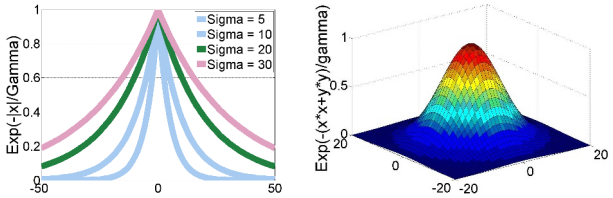
Algorithms	Average Rank			
	Two Views		Single View	
	Before	After	Before	After
DoubleBP	21.5	**5.75**	18.42	**11.33**
AdaptingBP	15.33	**6.92**	**12.58**	**9.17**
C-SemiGlob	**11.75**	**7.25**	**7.75**	**5.25**
Segm+visib	16.08	**9**	**14.67**	**11.17**
SymBP+occ	25	**12**	22.42	**14.92**
SemiGlob	15	**12.33**	**12**	**9.83**
AdaptWeight	29.25	**12.42**	25.58	14.92
RegionTreeDP	32.75	**14**	29.75	18.17
GC+occ	25.33	**14.42**	24	19.08
TensorVoting	24.83	17.67	22.08	16.33
MultiCamGC	28.17	17.83	27	23.08
Layered	34.83	18.08	32.25	25
SegTreeDP	28.33	18.17	26.33	17.42
RealtimeBP	34	18.92	31.42	21.92
CostRelax	23.58	20.17	21.58	21.17
GenModel	22.17	20.83	20.25	17.58
ReliabilityDP	40.17	23.5	37.5	29.5
RealTimeGPU	38.42	24	36.58	24.67
GC	33.67	24.5	32.42	29.25
TreeDP	44	30.5	42.5	36.67
DP	43.92	31.33	42.17	31.17
SSD+MF	46.08	34.75	45.17	41.75
STICA	44.67	35.25	44	35.58
SO	45.17	37.83	43.42	36.75
Infection	44.83	38.75	43.42	38.08

Figure 3. The scores on the last four columns are the average ranks with error threshold 0.5. The scores with bold font are among the top 10 performers. The entries with blue highlighting are stereo algorithms originally without sub-pixel estimation, the others are algorithms originally having sub-pixel estimation. The scoring scheme is the same as the Middlebury benchmark [14].

To validate our sub-pixel refinement approach, an off-line stereo benchmark that has the same scoring scheme as the Middlebury benchmark [14] is built. Every result reported on the Middlebury website is used with our sub-pixel refinement approach, and evaluated on the sub-pixel benchmark. Figure 3 shows that our approach is very robust, it works for all the algorithms, even for those originally having sub-pixel refinement. The completed version of Figure 3 is provided in the supplemental materials, which gives more details about the ranks on different datasets and error thresholds.

4. Experimental Results

Our experimental system consists of a Canesta EP DevKit camera [1] and a FLEA digital camera at [7]. The EP DevKit camera can produce range images with size up to 64×64 of the objects in its view, and the FLEA camera



(a) (b)

Figure 4. (a) $\gamma_c \in \{5, 10, 20, 30\}$. (b) $\gamma_s = 10$.

can produce color images with resolution up to 1024×768. These two cameras are placed very close to each other and image registration is achieved by a 3×3 homographic warp. The warping function is dependent on the average range to the object. A better setup would be to use an beam-splitter to align the optical axes of both sensors to guarantee image alignment.

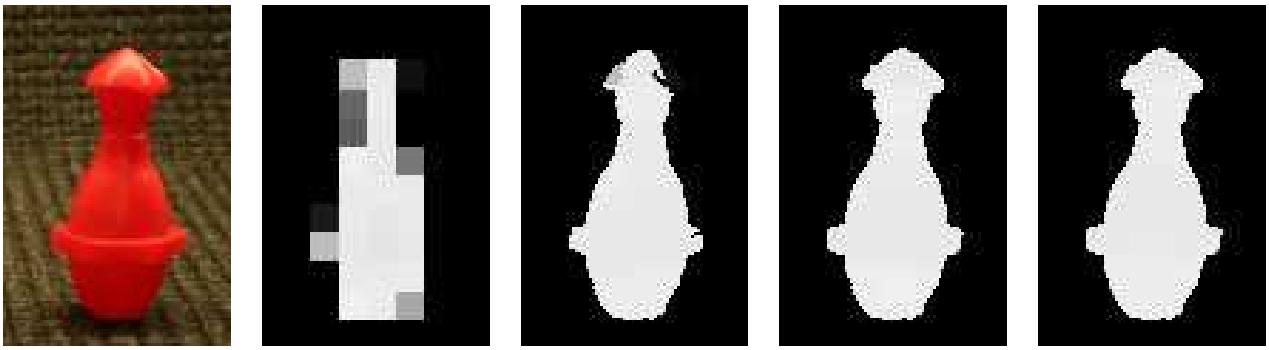
Three main parameters are involved in the experiment, they are η, γ_c and γ_s. η is the constant used in Equation 1, it is set to 0.5 experimentally. To allow large depth variations, the cost function is truncated by $\eta \times L$, where L is the search range. Two parameters are involved in the bilateral filter, they are γ_c and γ_s. In this paper, they are both set to 10. A visual explanation about how these parameters control the shape of the weighting functions in Equation 3 is provided in Figure 4. The experimental results show that γ_c is relatively sensitive, it should be decreased around the low texture area.

4.1. Spatial super resolution

To show the power of the iterative bilateral filtering, a series of intermediate depth maps are provided in Figure 5 in a coarse-to-fine manner architecture.

In Figure 5, D_0 is the depth map after up-sampling from the low-resolution range image. The quality of D_0 is unacceptable. D_1 is the depth map after iteration 1. The quality has been improved a lot, but the areas around part of the discontinuities are incorrect. D_3 is the depth map after iteration 3, the discontinuities are well detected, and the algorithm has almost converged. D_{10} is the depth map after iteration 10. By visual comparison, the difference between D_{10} and D_3 is tiny. Others experimental results are shown in Figure 6. The input depth maps are up-sampled from the 64×64 range images, and the resolution of the output depth maps is 640×640.

Table 1 evaluates the performance of our approach and the MRF approach presented in [8] on the Middlebury datasets on three different scales. On each scale, the depth image is down-sampled by a factor of 2 gradually. On scale 0, the depth image is the ground truth. By comparing the bad pixel percentages before and after bilateral filtering refinement, we show that our approach improves the stereo



(a) Camera image (b) D_0 (c) D_1 (d) D_3 (e) D_{10}

Figure 5. Intermediate results from iterative bilateral filtering refinement module. (a) Camera image. (b) The initial depth map. (c) Depth map after one iteration. (d) Depth map after three iterations. (e) Depth map after ten iterations.

Algorithms	Tsukuba Scale			Venus Scale			Teddy Scale			tsukuba Scale		
	1	2	3	1	2	3	1	2	3	1	2	3
Before Refinement	2.67	5.18	9.66	0.61	1.34	2.79	2.92	8.64	14.7	3.92	7.85	14.7
MRF Refinement [8]	2.51	5.12	9.68	0.57	1.24	2.69	2.78	8.33	14.5	3.55	7.52	14.4
Bilateral filtering Refinement	1.16	2.56	6.95	0.25	0.42	1.19	2.43	5.95	11.5	2.39	4.76	11.0

Table 1. Experimental results on the Middlebury datasets. The numbers in the last twelve columns are the percentages of the bad pixels with error threshold 1.

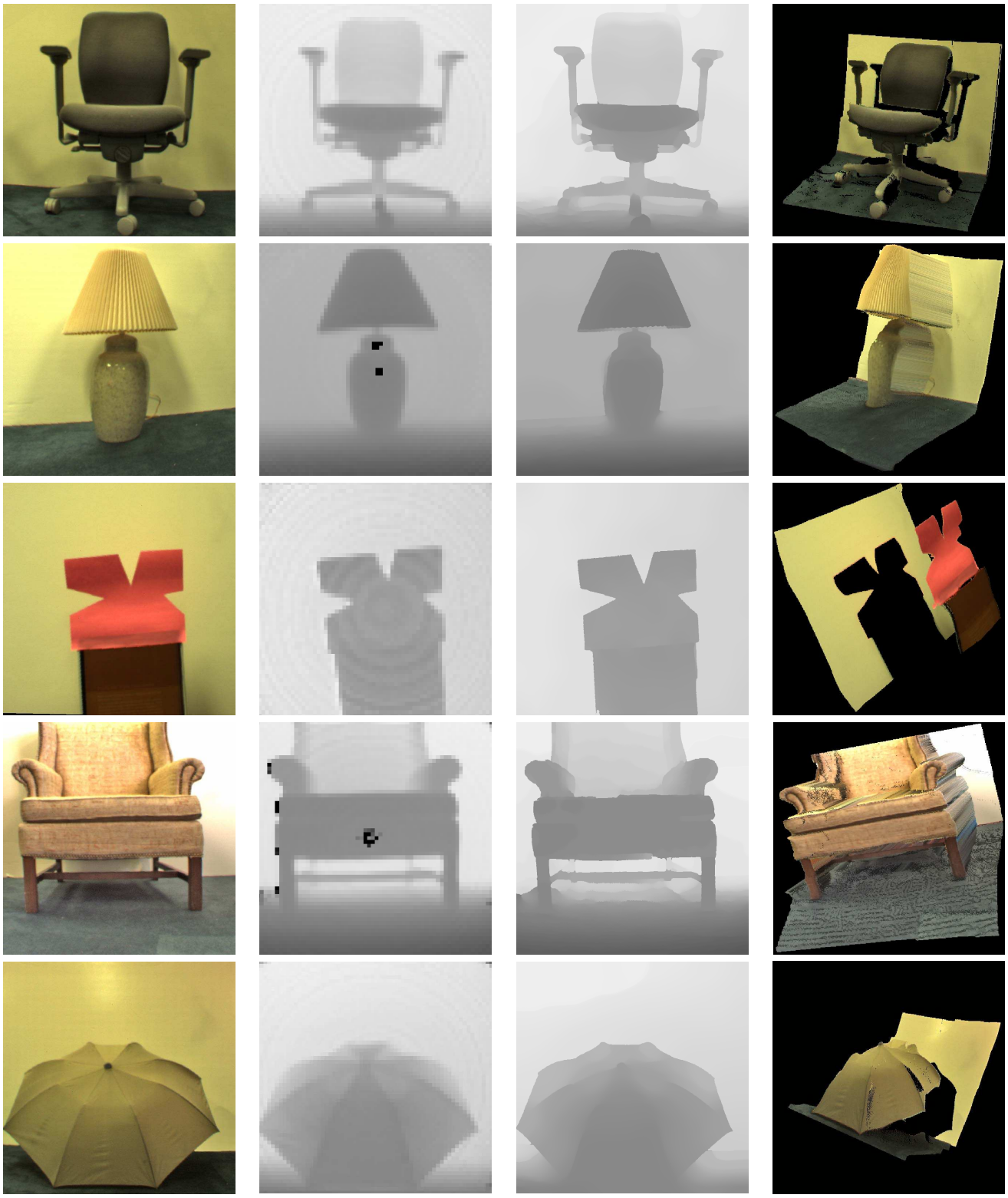
quality of all data sets. The MRF approach in [8] also improves the stereo quality, but the improvement is relatively small compared to our approach. A visual comparison of the depth maps of the Middlebury datasets on Scale 3 are provided in Figure 7. Clearly, the results using our approach have more clean edges than the input depth maps and the results using MRF approach. For further comparison, Figure 8 provides the experimental results of the Cones data set from scale 1 to scale 4 using the MRF approach and our approach. By visual comparison, our approach outperforms the MRF approach as the resolution of the range sensor keeps on decreasing. On the last row in Figure 8, we show that even with tiny sensors (23×28), we can still produce decent high-resolution range images.

4.2. Sub-pixel estimation with one or two reference image(s)

Besides the enhancement of the spatial resolution of range images, our approach also provides sub-pixel estimation for general stereo algorithms with either one or two camera image(s). To evaluate the performance of our sub-pixel estimation approach, we established an off-line stereo benchmark. The scoring scheme is the same as the Middlebury benchmark. In our off-line benchmark, all the algorithms reported to the Middlebury benchmark and their sub-pixel refinement results are evaluated, thus the total number of algorithms evaluated is twice the number on the Middlebury benchmark [14]. Figure 3 provides the average ranks for all the algorithms. With either one or two view(s), we achieve across-the-board improvement for sub-pixel accuracy. The 10 entries with bold font are the top 10 performers. In two-view case, nine of them are the algorithms with our sub-pixel refinement approach. All the entries without blue highlighting in Figure 3 are average ranks of those algorithms using its own sub-pixel refinement techniques. The experimental results show that our sub-pixel estimation approach works for all of these algorithms, however the improvement is naturally a bit smaller than for the cases that originally don't have any kind of sub-pixel refinement. A set of synthesized views built from the **DoubleBP** algorithm [19] are shown in Figure 9, providing a visual comparison of the algorithms with and without sub-pixel refinement. The depth enhancement is obvious. The results shown in column (a) are quantized to discrete number of planes. After sub-pixel estimation, the quantization effect is removed, as it is shown in column (b).

5. Conclusion

In this paper, we present a new post-processing step to enhance the spatial resolution of range images up to 100x with a registered and potentially high-resolution color image as reference. We have validated our approach on several real datasets, including the Middlebury data set, demonstrating that our approach gives clear improvements. In addition, the depth super resolution is extended to two-view



(a) Camera images. (b) Input depth maps. (c) Refined depth maps. (d) Synthesized views.

Figure 6. Experimental results of depth super resolution. (a) Camera images. (b) Input depth maps. (c) Refined depth maps. (d) Synthesized views by using (c). The input depth maps are up-sampled from range image with resolution 64×64, and resolution of the refined depth maps is 640×640. The spatial resolution is enhanced $100\times$.

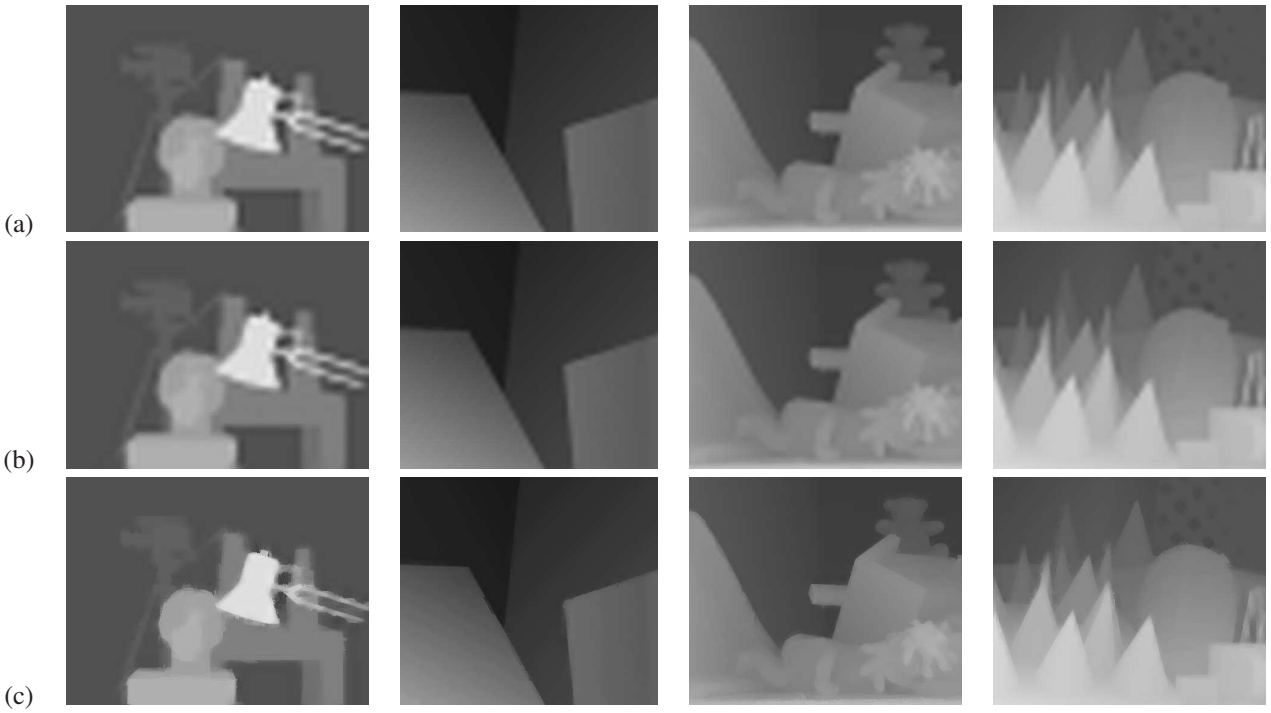


Figure 7. Super resolution on Middlebury datasets. (a) Before refinement. (b) Using MRF approach [8]. (c) Using our approach.

case. To evaluate the effectiveness of our depth-enhanced approach, we first built an off-line stereo benchmark that has the same scoring scheme as the Middlebury benchmark, then tried our approach on all the stereo algorithms reported to the Middlebury benchmark. Together with all the results submitted to Middlebury benchmark, we evaluated all the depth-enhanced results on the off-line benchmark with different error thresholds, and showed across-the-board improvement in sub-pixel accuracy. We are hoping to release an on-line sub-pixel benchmark in the near future.

References

[1] Canestavision^{TM} electronic perception development kit, canesta inc. http://www.canesta.com/html/$development_kits$.htm. 1, 4

[2] Z-cam, 3dv systems. http://www.3dvsystems.com/home/index.html. 1

[3] Adaptive support-weight approach for correspondence search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):650, 2006. K.-J. Yoon and S. Kweon. 1, 2

[4] J. Batlle, E. Mouaddib, and J. Salvi. Recent progress in coded structured light as a technique to solve the correspondence problem: A survey. *Pattern Recognition*, 31(7):963–982, 1998. 1

[5] P. Besl. *Active Optical Range Imaging Sensors, in Advances in Machine Vision, chapter 1*, pages 1–63. 1989. 1

[6] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4):401–406, apr 1998. 3

[7] F. Camera. Point grey research. http://www.ptgrey.com/products/flea/index.asp. 4

[8] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In *NIPS*, 2005. 1, 4, 5, 7

[9] R. Jarvis. A perspective on range finding techniques for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):122–139, 1983. 1

[10] G. Piella. A general framework for multiresolution image fusion: from pixels to regions. *Information Fusion*, 4(4):259–280, 2003. 1

[11] D. Poussart and D. Laurendeau. *3-D Sensing for Industrial Computer Vision, in Advances in Machine Vision, chapter 3*, pages 122–159. 1989. 1

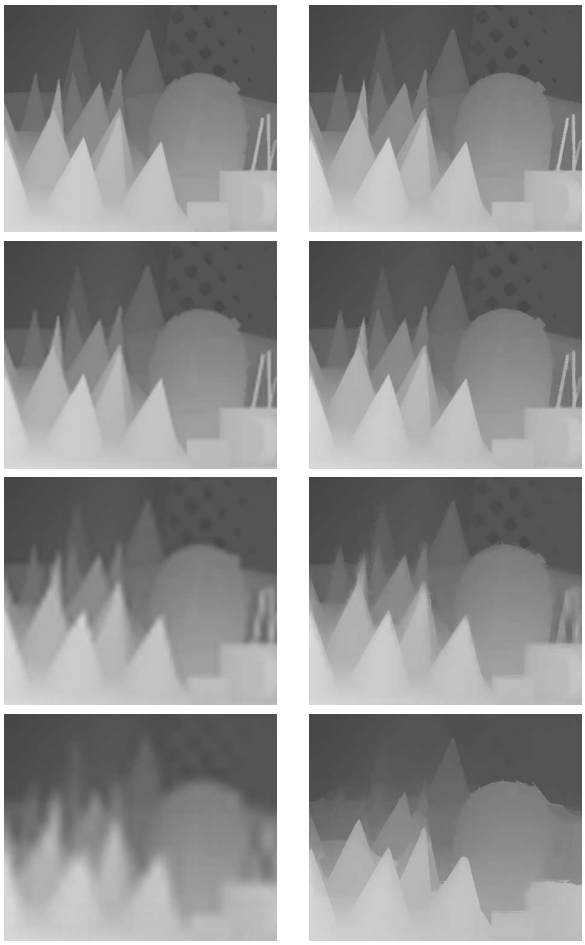
[12] W. H. Press. *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, New York, 1988. 1

[13] J. Salvi, J. Pagès, and J. Batlle. Pattern codification strategies in structured light systems. *Pattern Recognition*, 37(4):827–849, 2004. 1

[14] D. Scharstein and R. Szelisk. Middlebury stereo vision research page. http://bj.middlebury.edu/ schar/stereo/newEval/php/results.php. 1, 4, 5

[15] T. C. Strand. Optical three-dimensional sensing for machine vision. *Optical Engineering*, 24(1):33–40, 1985. 1

[16] C. S. Swiss Ranger SR-2. The swiss center for electronics and microtechnology. http://www.csem.ch/fs/imaging.htm. 1



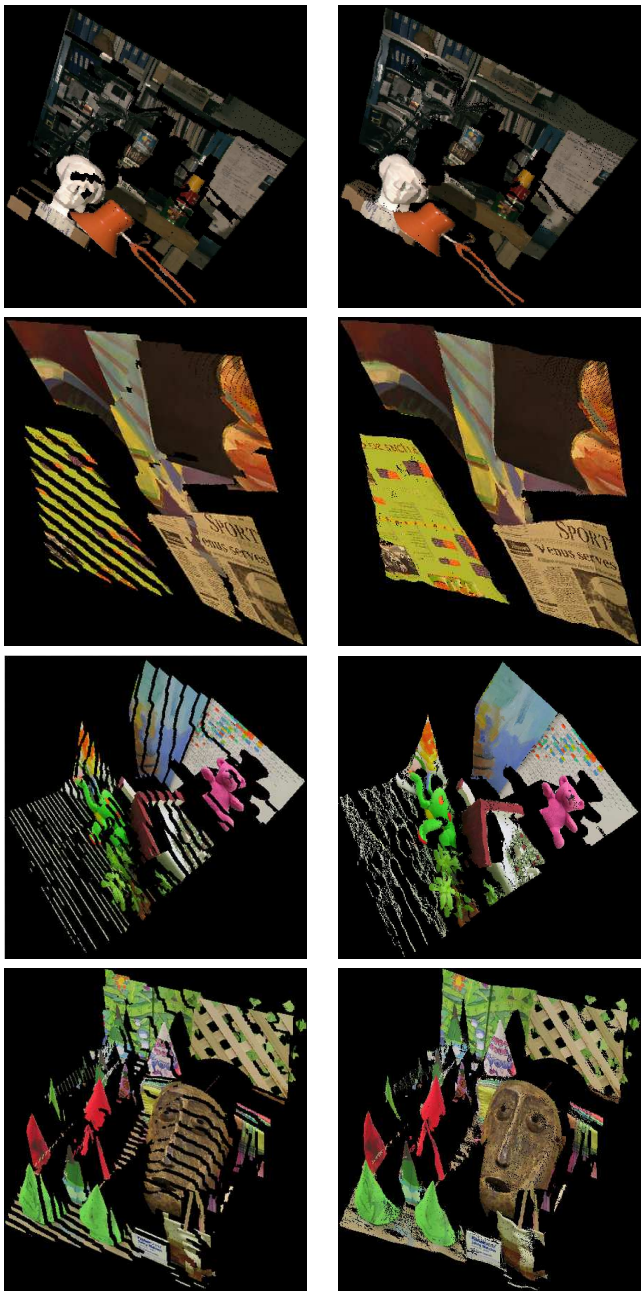
(a) Using MRF approach. (b) Using our approach.

Figure 8. Super resolution on Cones datasets. From up to bottom: Experimental results on scale 1 (resolution: 187×225), Experimental results on scale 2 (resolution: 93×112), Experimental results on scale 3 (resolution: 46×56), Experimental results on scale 4 (resolution: 23×28). This figure shows that, by visual comparison, our approach performs better than the MRF approach as the resolution of the range sensor continues to drop.

[17] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV*, pages 839–846, 1998. 1

[18] L. Wang, M. Liao, M. Gong, R. Yang, and D.Nistér. High-quality real-time stereo using adaptive cost aggregation and dynamic programming. In *Third International Symposium on 3D Processing, Visualization and Transmission (3DPVT 2006)*, June 2006. 1

[19] Q. Yang, L. Wang, R. Yang, H. Stewénius, and D. Nistér. Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. In *CVPR (2)*, pages 2347–2354, 2006. 1, 2, 3, 5, 8



(a) Before. (b) After.

Figure 9. (a) Synthesized views produced by the DoubleBP algorithm [19] reported on the Middlebury website. (b) Synthesized views after sub-pixel refinement. The results shown in column (a) are quantized to discrete number of planes, after sub-pixel estimation, the quantization effect is removed, as it is shown in column (b).