Disocclusion via Motion-parallax and Color-segmentation

***Abstract*—In this work, we propose a novel method for addressing the disocclusion problem of removing foreground occlusions from the depth maps as well as color images for 3D scenes. Our method exploits motion-parallax and color-segmentation cue to coherently fill-in the foreground regions with intensity/depth labels of background which is occluded by the foreground object in the observations. The motion-parallax allows the background information, occluded in some images, to be visible in other images. This includes stereo-correspondence information for depth disocclusion as well as color information for image disocclusion. The color-segmentation cue exploits a natural tendency of depth discontinuities to coincide with intensity edges in the image of the same scene. We also show that the color-segmentation cue is useful, on its own, for achieving disocclusion in range maps.**

I. Introduction

Unwanted foreground obstructions/occlusions are very common images of general 3D scenes. Given the immense amount of vision research in 3D structure estimation, removing such occlusions (also termed as *disocclusion*) seems quite a natural issue to address. Surprisingly, there are only a few methods which exploit the 3D nature of the scene to address this problem of disocclusion.

In this work we address the problem of disocclusion for 3D scenes. More specifically, we propose an approach to remove user-defined foreground objects which obstruct a part of the background. Importantly, our method estimates a disocclued depth map as well as an dissoccluded image given multiple stereo observations of the scene, all of which contain the foreground occluder. Furthermore, we also apply (a part of) our approach to achieve disocclusion of range images captured from laser range scanner or time-of-flight range scanners. Thus, our disocclusion method can be used for both digital cameras and range cameras.

The main challenge in achieving disocclusion is to fill-in the user-defined foreground regions with the *relevant* background intensities, so that the resultant depth map and image is visually *coherent*. Our approach involves using the motion cue, present in the stereo images for disocclusion. The primary idea is that the objects in a 3D scene undergo an apparent motion depending on their depth (which is known as motion parallax). Thus, the foreground occluder pixels will have a different motion across images, relative to the background pixels. Our approach exploits this phenomena to *discover* background pixels hidden in the reference view, in other views. Thus, using multiple images enables us to compute correspondence/color information for the pixels which are hidden in the reference image, thus allowing us to disocclude the (reference) depth map and image. Our parallax-based disocclusion approach uses the belief-propagtion method for estimating depth and color labels. In addition we also the cue from color-segmentation of the observed images to improve the depth estimation [1]. Indeed, the segmentation cue plays a more important role in our method than merely improving the depth estimation as discussed below.

Note that we need the occluded reference-image pixels to be visible in at least two (and usually more than two) other images for computing reliable correspondences for depth disocclusion. [1] If camera motion is restricted or the number of images are limited, reliable correspondences for all pixels may not established. As a result, some of the occluded pixels may remain unlabeled. For labeling such pixels, we use the color-segmentation cue mentioned earlier. In addition to the segmentation cue serving in conjunction with our parallax-based method, we also show that it can be used, on its own, for disoccluding range-images. The segmentation-cue exploits natural properties of 3D scenes viz. depth discontinuities being coincident with image edges and, depth variation over small regions (segments) being locally planar.

A. Related work

The dis-occlusion problem which we consider is closely related to that of inpainting. However, the latter has been mainly been addressed for single color images without considering any 3D structure dependency [9], [10]. These approaches operate on single images and compute plausible color values to be filled in the missing regions typically using neighbourhood information in some sense. Unlike, our approach their task does not include depth estimation/disocclusion.

Handling occlusions in traditional stereo depth estimation [12] may also be treated as dis-occlusion, since it also involve removing occlusions. However, the fundamental definitions of *occluded pixels* in traditional stereo works and those in ours are precisely complementary. While the occluded pixels in stereo are those which are seen in the reference image but not other images (which we as *stereo occlusions*), in our case, the occluded pixels (also called *hidden pixels*) are those which are not visible in the reference image but may be seen in the other images. In fact, stereo depth estimation is a integral part of our method (as elaborated in the next section), we stereo occlusions are also handled in our approach. Thus, our approach addresses both types of disocclusions. Moroever,

[1] Relatively, color computation for image disocclusion is easier since one needs the hidden reference-image pixels to be visible in at least one other image.

traditional stereo works are only concerned with depth estimation, whearas our approach also involves image disocclusion.

Disocclusion in both images and depth has received relatively less attention. To our knowledge, only the recent reported works in [7], [6] closely relate to ours, which exploit the pixel motion for disocclusion. The authors in [6] address the problem in a binocular stereo setting [6]. Their approach requires a priori computation of complete depth maps (with the occluders) from both the views. In contrast to [6], our work handles multiple stereo images and our depth estimation is only with respect to one (reference) view. The work in [7] consider the effects of pixel motion under a shape-from-focus setting. Their approach handles the disocclusion problem implicitly as opposed to our approach which explicitly checks for missing pixels. Moreover, unlike in [7] we also consider the stereo occlusions. Moreover, our color-segmentation based approach to fill in the unlabeled regions and its application to disoccluding range images is completely novel which is not considered in [6], [7].

Some works address the disocclusion problem in depth/range maps [13] and 3D meshes [3]. However, their main goal is to restore damaged range data. Moreover, these approaches work directly with range maps/3D meshes as input and do not involve disocclusion of color images.

II. The approach

We now describe our disocclusion method in detail. Our depth and image disocclusion approach works in two stages. We first compute the disoccluded depth map using multiple stereo observations where the user has marked, in one (reference) image, the foreground regions to be removed. Given the estimated disoccluded depth map we then perform image disocclusion image the same multiple stereo observations.

For depth map disocclusion, we first discuss the parallax-based approach for *discovering* correspondences for the hidden reference-image pixels across other images, followed by a segmentation-based method for disoccluding those pixels which are unlabeled in the parallax-based approach. The image disocclusion is based solely on the parallax-based approach, which disoccluded almost all the hidden image pixels. For the remaining unlabeled image pixels (which are very few in number and span small regions), we use the exemplar-based inpainting approach [10] for estimating their labels.

Our parallax-based approach employs the efficient-BP algorithm [8], which involves computing messages and beliefs on an image-sized grid where a label is assigned to each node. The messages and beliefs are expressed as data and prior costs which are to be minimized. The data cost uses the observations, while the prior cost regularizes the solution. In the subsequent sections, we describe these cost definitions for our problem.

A. Depth disocclusion

Our method begins with the user marking the foreground pixels to be removed. To minimize the user-interaction, the occluder is to be marked *only in the reference image*. The occluder pixel locations in other images are marked depending on the knowledge of those in the reference image. Since occluder is also a part of the scene, its pixels will also undergo motion across images and this motion depends on the depth of the occluder. Hence, to enable the marking of occluder pixels in the other images, we first compute the depth of the complete scene including that for the occluder. Thus, the depth disocclusion process has three main components 1) Depth estimation for the complete scene using multi-image stereo 2) Motion-parallax-based depth disocclusion 3) Segmentation-based depth disocclusion.

1) Stereo depth estimation: The complete depth estimation is carried out by using a belief-propagation-based stereo technique as described below. Since the BP algorithm considers computing costs for every label at each node, the following costs for a particular label at a particular node (pixel).

The data cost at a pixel (l_1, l_2) in the reference image for the stereo matching problem is

$$E_{di}(l_1, l_2) = |g_1(l_1, l_2) - g_i(\theta_{1i}, \theta_{2i})| \quad (1)$$

where $(\theta_{1i}, \theta_{2i})$ is the warped location in the i^{th}, which depends on the camera motion and the depth label Z for which the cost is computed. Her, without loss of generality, out of all observations g_i, $i = 1...N$, we use g_1 as the reference.

To consider *stereo-occlusions* we modulate the data cost with a visibility term $V_i(l_1, l_2)$ which switches on/off depending on whether a pixel is visible or occluded in the i^{th} image. The resultant data cost is

$$E_{di}(l_1, l_2) = V_1(l_1, l_2) \cdot |g_1(l_1, l_2) - g_i(\theta_{1i}, \theta_{2i})| \quad (2)$$

Since stereo-occlusions is not the focus of this paper and due to space-constraints, we refer the reader to [11], [1] for more details on computing and updating visibility V_i via the *geo-consistency* concept. The total data cost between the reference image ($i = 1$) and all other images ($i > 1$) is then computed by summing the individual costs E_{di} for all $i > 1$.

The regularization cost which smoothes the solution while preserving prominent discontinuities is defined as a truncated absolute function

$$E_p(n_1, n_2, m_1, m_2) = \min(|Z(n_1, n_2) - Z(m_1, m_2)|, T) \quad (3)$$

where, $Z(n_1, n_2)$ and $Z(m_1, m_2)$ are depth labels for neighbouring nodes and T is the threshold for truncation.

The estimation using only the above defined costs may yield some of the pixels as labeled incorrectly. As mentioned above, the image segmentation can be used to mitigate such errors. Note that the use of segmentation cue here is *not* used for disocclusion but for improving the depth estimates. Again, due to space constraints we describe it here briefly. More details can be found in [1]. Initially, we color-segment the reference image using the mean-shift algorithm [2]. We classify the pixels as reliable/unreliable based on an initial coarse depth estimate. The first BP iteration is run without using the segmentation cue. We then compute a plane-fitted depth map that uses the current estimate, the segmented image

and the reliable pixels. To regularize the estimates for the unreliable pixels, we feed this plane-fitted depth back to the iteration process by defining a more general data term as

$$E_{ds}(l_1, l_2) = E_d(l_1, l_2) + w(l_1, l_2) \cdot |Z(l_1, l_2) - Z_p(l_1, l_2)| \quad (4)$$

where $E_d(l_1, l_2)$ is the previously defined data cost. Z_p denotes the plane-fitted depth map and the weight w is 0/1 if the pixel is reliable/unreliable. We use this data cost in subsequent iterations after the first.

2) Depth disocclusion using motion-parallax: The above method yields a complete depth map which also includes the depth labels of the foreground occluders. Using these depth labels and the user-marked occluders in reference image, we locate the occluder pixels in the other images. At the end of this process, we have marked the pixels to be disoccluded in all the images. We denote these set of *hidden* pixels as M and proceed as follows to compute the *background* depth labels for such pixels.

We arrange the images in an (arbitrary) order $(g_1, g_2, ..., g_N)$ with g_1 being the reference image. For a hidden pixel, $g_1(l_1, l_2) \in M$ and is not visible in the reference image. Hence, we need to compute the depth value at (l_1, l_2) by searching for the correspondence between images other than the reference.

We compute the coordinates $(\theta_{1i}, \theta_{2i})$ and $(\theta_{1j}, \theta_{2j})$ for a depth label. If $g_i(\theta_{1i}, \theta_{2i}) \notin M$ and $g_j(\theta_{1j}, \theta_{2j}) \notin M$, the matching cost between them is defined as

$$E_{di}(l_1, l_2) = V^c_{ij}(l_1, l_2) \cdot |g_i(\theta_{1i}, \theta_{2i}) - g_j(\theta_{1j}, \theta_{2j})| \quad (5)$$

where $1 < i < j$ and the visibility V^c_{ij} is a *compound* visibility term defined as

$$V^c_{ij}(l_1, l_2) = V_i(l_1, l_2) \cdot V_j(l_1, l_2) \quad (6)$$

The *compound* visibility defined above is also takes into account the stereo-occlusions as well as the user-defined hidden pixels. The idea behind defining the *compound* visibility is that the data cost is not computed if a pixel is not observed in either the i^{th} or the j^{th} view.

The corresponding total data cost for $g_1(l_1, l_2)$ involving all images for a depth label is then computed by summing the matching costs as

$$E_d = \frac{1}{N_i} \sum_i E_{di} \quad (7)$$

Here N_i are the number of pairs of images g_i and g_j such that $g_i(\theta_{1i}, \theta_{2i}) \notin M$ and $g_j(\theta_{1j}, \theta_{2j}) \notin M$ and $V_i(l_1, l_2) \neq 0$. Thus, the cost for a pixel missing in the reference image is computed by using those images in which the pixel is visible.

3) Segmentation-based depth disocclusion: In the above procedure, there may be some pixels for whom $N_i = 0$. The pixel correspondences for such hidden pixels are not found. This results in such pixels being left unlabeled. We invoke the segmentation cue to estimate labels for such pixels.

As mentioned earlier, the segmentation cue exploits the behaviour of depth discontinuities coinciding with the image edges. This behaviour allowed the segmentation-cue to improve the stereo-depth estimation (as discussed previously) when all pixels in a segment are visible, having some estimates of depth, which are used for plane-fitting.

However, for the disocclusion problem, the scenario is quite different. Here, we wish to compute the hidden background depth values for which we have no depth estimates in a segment. In fact, the segments for these hidden regions are not even the true segments of the scene. These segments are themselves unlabeled and are hence erroneous. We denote a set of such segments by S_m. Each such segment will span across largely different depth values, thus disobeying the very premise for the use of the segmentation cue of locally planar depth variation. Hence, we cannot use such segments to compute the plane-fitted depth map.

To address this issue, we assign the segment-labels to each pixel in S_m as that of the segment $\notin S_m$ closest to that pixel. This essentially extends the segments neighbouring to those in S_m by adopting the pixels in S_m. The *closeness* is determined by searching in eight directions from a pixel. Thus, this process assigns segment-labels to all the pixels in S_m. Moreover, since these labels are assigned according to the closest neighbouring segments to the pixels in S_m, they are very similar to their actual natural labels which they would have been assigned, if the foreground occluder would not have been present.

Note that these expanded segments already have disoccluded pixels for which the depth is estimated using the parallax-method described earlier. These can now be used to compute the labels for the newly added unlabeled pixels (from S_m). We use the disoccluded pixels to fit a plane via the RANSAC method [4]. We then use this plane-fit and the disoccluded pixels to define a local cost C_p for assigning a depth label for each invisible pixel p.

$$C_p = |z - z_{pl}| + \lambda_p \sum_{q \in V_p} |z - z_q| \quad (8)$$

Here, z_{pl} is the plane-fitted range at a pixel. V_p is the set of *visible* neighbours in the second order neighbourhood of pixel p that belong to its segment. The second term on the RHS of equation 8, weighted by λ_p, enforces similarity between neighbours. The depth label minimizing C_p is chosen as the label for pixel p.

For some segments, the number of disoccluded pixels are below a threshold (which generally occurs for small segments). Hence, the plane-fitting based labeling of equation 8 may not be robust. For such segments, we compute the median z_m depth over the disoccluded pixels. We also compute the median depths z_{ma} of the disoccluded pixels for the adjacent connected segments. We then label all the unlabeled pixels for such small segments according to the cost

$$C_s = |z - z_m| + w_a \sum_{z_{ma} \in M_a} |z - z_{ma}| \quad (9)$$

where M_a is the set containing medians z_{ma} of the visible pixels in adjacent segments. In equation 9, the second term enforces similarity over neighbouring segments. with the quantity w_a weighing their contribution. Note that this assigns all

the unlabeled pixels in a segment with a constant (median) depth label. However, as mentioned earlier, this occurs for very small segments for which a constant depth approximation is also quite valid.

B. Image disocclusion via motion-parallax

Given the estimated disoccluded depth map, we now wish to estimate the color labels for the hidden pixels in the reference image. With the estimated depth label for each pixel, we map the location of a hidden reference pixel in other images. If the pixels at the mapped locations $\notin M$, we use them in our data cost computation.

The data cost for image disocclusion compares the intensities of $g_i(\theta_{1i}, \theta_{2i})$ $i > 1$ with an intensity label, if $g_i(\theta_{1i}, \theta_{2i}) \notin M$. This data cost for a particular $g_i(\theta_{1i}, \theta_{2i}) \notin M$ and an intensity label L is defined as

$$E_{di}(l_1, l_2) = V_i(l_1, l_2) \cdot |L - g_i(\theta_{1i}, \theta_{2i})| \tag{10}$$

The total data cost is sum of the N_i data costs similar to that described by equation 7, where N_i is the number of images where $g_i(\theta_{1i}, \theta_{2i}) \notin M$ $i > 1$. The smoothness cost for the image is also defined similar to that in equation 3, except that the intensity labels instead of depth labels are used.

Lastly, there may be missing pixels in g_1 for which $g_i(\theta_{1i}, \theta_{2i}) \in M$ $\forall i$. Such pixels are left unlabeled. The extent of such unlabeled pixels depends on the original extent of the missing region and pixel motion. We observe in our experiments, that for most of the image the pixel motion is sufficient to leave no missing region unlabeled. The maximum extent of such unlabeled regions, if they exist at all, is relatively very small as compared to that of the original hidden regions. Such small unlabeled regions can be filled by any inpainting algorithm (for instance, the exemplar-based inpainting [10]).

III. Experimental results

We validate our approach via real experiments on multiple stereo images from the Middlebury stereo dataset [12]. Moreover, as mentioned earlier, we also show results for disocclusion of range images from the USF dataset [5]. These datasets contain complex scenes and serve to test our approach quite extensively. We mark some of the prominent foreground objects for removal. For the stereo dataset the object selection is carried out only in the reference image. The foreground objects cover a considerable-sized region.

The parameters in our method are typically set as follows: The smoothness weights in both the motion-parallax and segmentation-based methods are chosen so the smoothness costs approximately balances the data cost. The truncation threshold is set to half the maximum depth/intensity label. The minimum number of visible pixels for robust plane-fitting is chosen to be 20-30. The belief propagation algorithm for parallax-based disocclusion is run for about 4-5 iterations. The convention for displaying the depth map is that the nearer objects are darker than the farther ones.

We first provide the results on the stereo images of the Middlebury dataset. These results are generated using our complete method described section II for both depth and image disocclusion. This is followed by the depth disocclusion results on USF range-image dataset which only uses the segmentation-based method of section II.A.3.

A. Disocclusion in stereo

The images in the results on stereo-disocclusion are ordered as follows: The sub-figures (a,b) in all the experiment show two of the four observations used of which the first observation is the reference. Sub-figure (c) shows the reference image with foreground objects to be removed, marked in black. Sub-figures (d,e) show the complete depth-map including the foreground occluders and, the dissocluded depth map, respectively, while (f) shows the disoccluded reference image.

Figs. 1, 2 and 3, the disocclusion results on three stereo datasets for the *Drumsticks*, *Dwarves* and *Reindeer* scenes, respectively. Note that, basically, the complete depth maps shows very good localization of discontinuities and a very plausible depth variation. Note that the correctness is localization and accuracy of depth estimation are important for marking the foreground objects in the all images based on that in the reference image. The correctness in marking in turn dictates that of the disocclusion process.

One can observe that the disoccluded depth map does not show any trace of the foreground occluders. Particularly, the background edges which the occluders cross are do not show any visual distortion. Similarly, the smoother depth regions also do not show any visible artifacts, thus demonstrating the successful disocclusion.

The image disocclusion also shows high visual coherence. Particularly, for the *dwarves* and *reindeer* scenes where the foreground objects (the plant and the reindeer, respectively) span a fairly large region, the disocclusion is quite appreciable. A (much) closer look shown minute distortions in some of the edges, which is hardly visible and may be discounted. Moreover, the background texture is also restored quite well, which is an important concern in disocclusion/inpainting methods.

B. Disocclusion in range images

We now provide results to show the effectiveness of the segmentation-based method for removing occlusions. As mentioned, we use the range images for demonstrating this. Range images are captured from laser scanners or time-of-flight range scanners [14], [13]. These are increasing in popularity due to their accuracy and high depth-resolution.

Our segmentation-based approach uses a registered intensity-image and range-image pair to compute the disoccluded range image. Given the availability numerous range-intensity image registration techniques and the fact that many range cameras also capture a corresponding intensity image, acquiring such a registered pair is not a major issue.

The user marks the occluder in either the range or the intensity image. Once the occluder is marked and we know the unlabeled pixels, we are effectively in the same stage as that in the stereo-disocclusion case when the motion-parallax

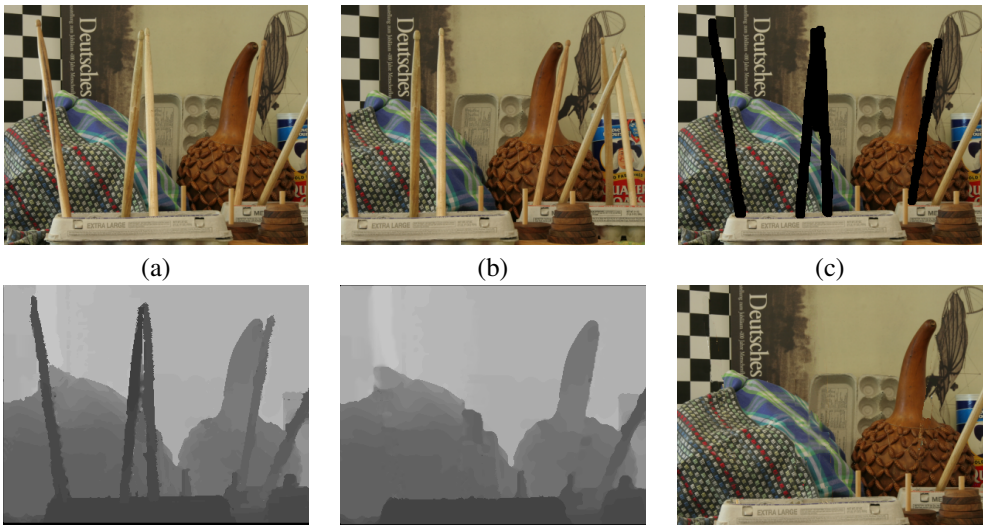

Fig. 1. Drumsticks scene: (a,b) Two of the four observations used in the experiment. (c) Reference observation with the marked occluders. (d) Estimated complete depth including the occluders. (e,f) Disoccluded depth and image, respectively.

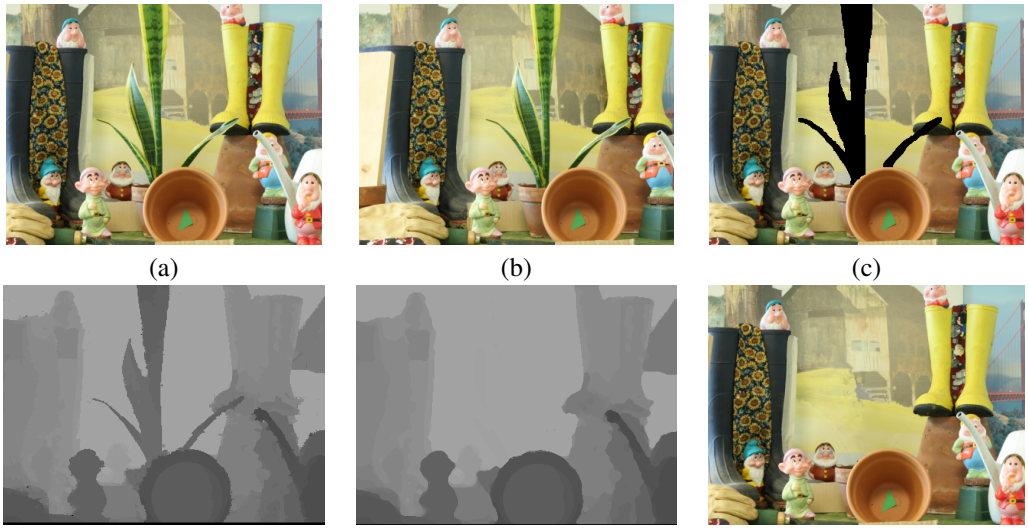

Fig. 2. Dwarves scene: (a,b) Two of the four observations. (c) Reference observation with occluders marked. (d) Estimated complete depth map. (e,f) Disoccluded depth and image, respectively.

based disocclusion leaves some pixels unlabeled. Thus, range disocclusion only needs the segmentation-based approach.

In Figs. 4 and 5, we show two results on range images. Sub-figure (a) shows the original intensity image where the occluders are marked with black in sub-figures (b) (The cone-like object at the back in Fig. 4(b) and the table in Fig. 5(b)). Sub-figures (c,d) give the (complete) ground-truth range map and the disoccluded range map, respectively.

Again, we notice high fidelity in depth disocclusion, with the occluders all but removed. Notice that background edges of the walls and floors in Fig. 4(d) which are masked by the occluder are restored quite well. Moreover, in both examples, the gradual background range variation is also captured with negligible visible artifacts. This sufficiently validates our segmentation-based range disocclusion.

IV. Conclusion

We proposed an approach disocclusion of depth maps and images based on motion-parallax and color-segmentation. The motion-parallax based approach uses the fact that pixels occluded in one view can be seen in others. The segmentation-based approach enforces depth discontinuities to coincide with the image discontinuities and approximates depth variations to be locally planar. We provide results on stereo as well as range images to validate our method.

References

[1] A.V. Bhavsar and A.N. Rajagopalan. Depth estimation with a practical camera. In *British Machine Vision Conference (BMVC 2009)*, 2009.

[2] C. Dorin and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 24(5):603–619, 1999.

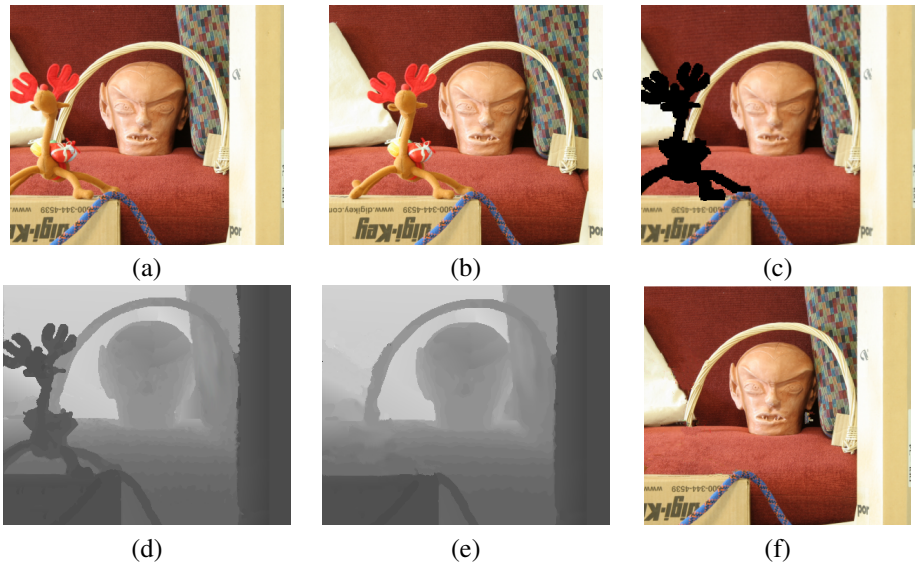

Fig. 3. Reindeer scene: (a,b) Two of the four observations. (c) Reference observation with the marked occluders. (d) Estimated complete depth including the occluders. (e,f) Disoccluded depth and image, respectively.

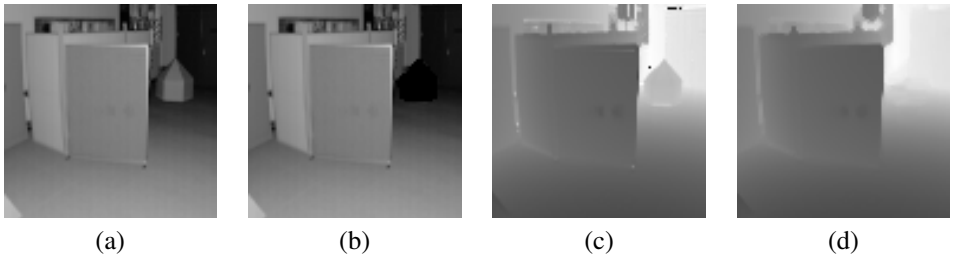

Fig. 4. Range data 1: (a) Original intensity image, (b) with the occluder marked. (c,d) Ground-truth and disoccluded range map, respectively.

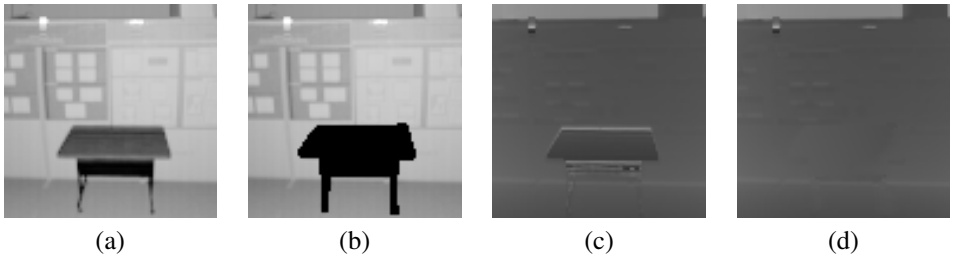

Fig. 5. Range data 2: (a) Original intensity image, (b) with the occluder marked. (c,d) Ground-truth and disoccluded range map, respectively.

[3] J. Davis, S. Marschner, M. Garr, and M. Levoy. Filling holes in complex surfaces using volumetric diffusion. In *International Symposium on 3D Data Processing Visualization and Transmission*, pages 428–438, 2002.

[4] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. 24(6):381–395, 1981.

[5] University of Southern Florida, USF range image database: http://marathon.csee.usf.edu/range/DataBase.html (1997).

[6] L. Wang, H. Jin, R. Yang, and M. Gong. Stereoscopic inpainting: Joint color and depth completion from stereo images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pages 1–8, 2008.

[7] Sahay, R., Rajagopalan, A.N.: Inpainting in shape from focus: Taking a cue from motion parallax. British Machine Vision Conference (BMVC 2009) (2009)

[8] Felzenszwalb, P., Huttenlocher, D.: Efficient belief propagation for early vision. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004). (2004) 1: 261–268

[9] Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. (2000) 417–424

[10] Criminisi, A., Perez, P., Toyama, K.: Object removal by exemplar-based inpainting. Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003) (2003) 721–728

[11] Drouin, M., Trudeau, M., Roy, S.: Geo-consistency for wide multi-camera stereo. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005). (2005) 1: 351–358

[12] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1).

[13] Bhavsar, A.V., Rajagopalan, A.N.: Range map with missing data - joint resolution enhancement and inpainting. Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2008) (2008) 359–365

[14] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pages 1–8, 2008.