Kinect Shadow Detection and Classification

Teng Deng[1], Hui Li[1], Jianfei Cai[1], Tat-Jen Cham[1], Henry Fuchs[2]
[1]Nanyang Technological University, [2]University of North Carolina at Chapel Hill
{dengteng,lihui,asjfcai,astjcham}@ntu.edu.sg, fuchs@cs.unc.edu

Abstract

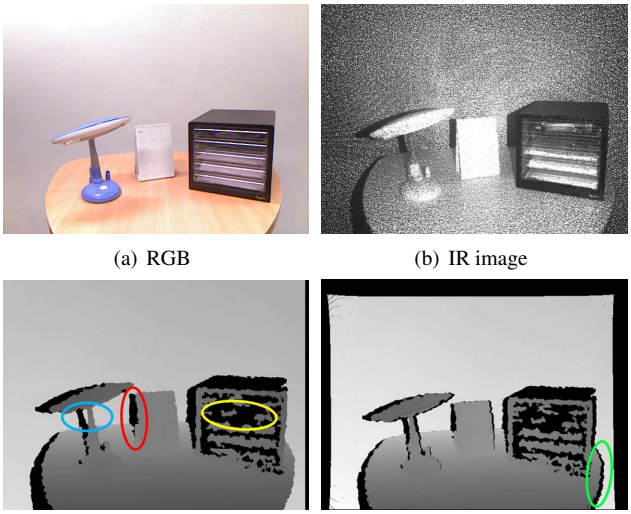
Kinect depth maps often contain missing data, or "holes", for various reasons. Most existing Kinect-related research treat these holes as artifacts and try to minimize them as much as possible. In this paper, we advocate a totally different idea – turning Kinect holes into useful information. In particular, we are interested in the unique type of holes that are caused by occlusion of the Kinect's structured light, resulting in shadows and loss of depth acquisition. We propose a robust detection scheme to detect and classify different types of shadows based on their distinct local shadow patterns as determined from geometric analysis, without assumption on object geometry. Experimental results demonstrate that the proposed scheme can achieve very accurate shadow detection. We also demonstrate the usefulness of the extracted shadow information by successfully applying it for automatic foreground segmentation.

1. Introduction

The Kinect sensor uses a structured light technique to generate a depth map of the scene, wherein a dot pattern is projected by an IR light projector and captured by a displaced IR camera (see Fig. 1(b)). By calculating the disparities of the projected dots between the IR projector and camera, the depth of the scene is attained (see Fig. 1(c)).

However, the depth map usually contains significant artifacts, among which the most notable ones are missing depth regions, or "holes". In some instances, the holes are due to partial occlusion of the structured light by foreground objects, leading to shadows in some background regions which are visible in the IR camera but unreachable by the IR projector pattern. In other instances, specular or low albedo surfaces may remove the visibility of the projector pattern in the IR camera [10].

Most existing Kinect-related research [5, 12] treat Kinect holes as artifacts and try to minimize them as much as possible using various filters. In this research, we advocate a totally different idea of turning Kinect holes into useful information. In particular, we are interested in the type of holes that result from structured light shadows, which often occur near the boundary of the objects in the scene, as illustrated in Fig. 1(c)&(d). Such shadow information is very useful in providing high-level structural information of the scene.



(a) RGB (b) IR image (c) Unaligned depth (d) Aligned depth

Figure 1. An example of Kinect data with different types of depth holes marked in red, blue, green and yellow colors for attached shadow, detached shadow, alignment shadow and non-shadow holes, respectively. (a) RGB image. (b) IR image with intensity amplified by 10 times. (c) Unaligned depth map. (d) Aligned depth map.

Utilizing shadow information is not new in the computer vision community. For example, the technique of shape from shadows [2] has long been studied. In some cases, shadows were deliberately created. For example, in [8, 3], Raskar *et al.* designed and developed a special camera with multiple flashes strategically positioned around the camera to capture cast shadows along depth discontinuities in the scene. Conversely, the shadows in Kinect depth maps are simply an expected but unwanted consequence of using structured light. Because only a single fixed projector is used, it is challenging to automatically detect shadows in

the depth map since they can take different shapes and must be distinguished from other causes of Kinect holes.

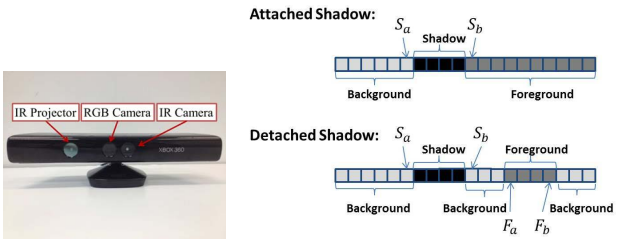
In this paper, we propose to detect different types of shadows using their distinct local shadow patterns desired from geometric analysis. In particular, we consider three types of shadows: *attached shadows*, *detached shadows* and *alignment shadows*, as shown in Fig. 1(c)&(d). An attached shadow is defined as a shadow that is directly adjacent to its occluding foreground object in the depth map, while a detached shadow is the one that is separated from its occluder by a short distance, *i.e*. there is some intervening visible background surface. An alignment shadow is a synthetic artifact that occurs when the original depth map as observed in the IR camera is rendered from the viewpoint of the RGB camera. Experimental results show that our proposed shadow detection algorithm can achieve very accurate shadow detection. To demonstrate the usefulness of our method, we further incorporate the shadow detection results into two state-of-the-art algorithms for automatic foreground segmentation. Experimental results show that the proposed shadow-assisted segmentation methods can achieve fully automatic foreground cutout with superior segmentation performance.

The main contributions of this paper are two-fold. First, we present a robust scheme to automatically detect different types of shadows in Kinect depth maps. As far as we know, there has been little work done on Kinect shadow detection, the closest being the work of Berdnikov and Vatolin [1] which proposes a preliminary occlusion classification method for real-time occlusion filling, presented without proof nor quantitative evaluation. Second, we propose to use the classified shadow information for fully automatic foreground segmentation, unlike existing RGBD object segmentation methods that require either some user input [11] or a training dataset for segmenting specific objects [4].

2. Proposed Kinect Shadow Detection

As mentioned earlier, we consider three types of shadows in Kinect depth maps: attached shadows, detached shadows and alignment shadows. We first focus on detecting shadows in the original depth map captured by the IR camera, where we only need to consider attached shadows and detached shadows, since alignment shadows only occur when the depth map is rendered in the the RGB camera viewpoint.

As the Kinect has the IR projector positioned to the left of the IR camera (see Fig. 2(a)), it is clear that the shadows in the original depth map will always be located to the left of any occluding foreground objects. This arrangement allows for the detection of shadows by scanning each horizontal line in the unaligned depth map. As a result, specific patterns of the shadows can be retrieved, as shown in Fig. 2(b).



(a) Kinect (b) Shadow Patterns

Figure 2. An illustration of (a) Kinect projector and cameras and (b) shadow patterns retrieved from horizontally scanning the unaligned depth map.

2.1. Geometric analysis of attached shadow

For the attached shadow illustrated in Fig. 3(a), let us imagine there is a virtual camera at IR projector location with the same image plane as the IR camera. Denote the IR camera center and the virtual camera center as O_1 and O_2 respectively, and denote their orthogonal projections to the image plane as O_a and O_b respectively. Let X denote the projection of the foreground edge point F_1 to the virtual camera. By drawing an auxiliary line $X'O_1$ parallel to the line XO_2, where X' is the intersection point with the image plane, we can easily get $\overrightarrow{X'O_a} = \overrightarrow{XO_b}$. Let $disp(p)$ measure the disparity between pixel p in the depth map and its corresponding pixel in the virtual camera. we can have

$$disp(S_b) = |\overrightarrow{S_bO_a} - \overrightarrow{XO_b}| = |\overrightarrow{S_bO_a} - \overrightarrow{X'O_a}| = |\overrightarrow{S_bX'}| \tag{1}$$

and

$$disp(S_a) = |\overrightarrow{S_aO_a} - \overrightarrow{XO_b}| = |\overrightarrow{S_aO_a} - \overrightarrow{X'O_a}| = |\overrightarrow{S_aX'}| \tag{2}$$

where S_a and S_b are respectively the left and the right nearest valid pixels to the shadow. In other words, the shadow in the depth map is in from S_a to S_b.

We can then estimate the length of an attached shadow as

$$\widetilde{w}_{as} \doteq |S_aS_b| = |S_bX'| - |S_aX'| = disp(S_b) - disp(S_a), \tag{3}$$

where $disp(S_a)$ and $disp(S_b)$ can be easily computed from the depth values of S_a and S_b according to the triangles of O_1BO_2 and $O_1F_1O_2$. Eq. (3) essentially states that the attached shadow length of a foreground edge point F_1 is fully determined by the depths of F_1 and the closest background point B lying on the 3D ray of O_2F_1.

Based on the above geometric analysis, we propose to detect an attached shadow by comparing whether there is a match between the observed shadow length w_{as} and the estimated shadow length $\widetilde{w}_{as}$ defined in (3). Most specifically, given a horizontal scanline in the depth map, the length of

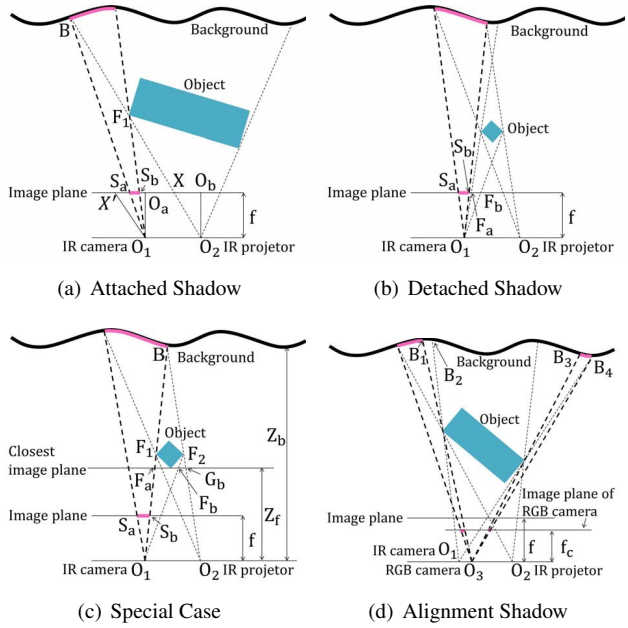


Figure 3. Imaging geometry of Kinect illustrating different types of shadows.

any contiguous region of missing depth can be measured as w_{as}, with the nearest valid pixels S_a and S_b at both ends identified. We then calculate the estimated shadow length $\widetilde{w}_{as}$ using (3). If $\widetilde{w}_{as} > 0$ and $\widetilde{w}_{as}$ is close to w_{as}, we will classify the region of missing depth as an attached shadow.

2.2. Geometric analysis of detached shadows

The formation of a detached shadow is due to the comparatively narrower width of the foreground object, as illustrated in Fig. 3(b). Let F_a and F_b denote respectively the projections of the left and the right edge points of the foreground object in the IR camera. Using the same procedure for deriving (3), we obtain

$$\widetilde{w}_{sb} \doteq |S_aF_a| = disp(F_a) - disp(S_a), \quad (4)$$

$$\widetilde{w}_{bf} \doteq |S_bF_b| = disp(F_b) - disp(S_b), \quad (5)$$

where $\widetilde{w}_{sb}$ is the estimated length of the shadow plus the portion of intervening background (sandwiched between the shadow and the foreground) in the depth map, while $\widetilde{w}_{bf}$ is the estimated length of the foreground plus the same portion of background, as illustrated in Fig. 2(b).

Based on the above geometric analysis, we propose to detect detached shadows by using their distinctive features as expressed in (4) and (5). More specifically, given a horizontal scanline of the depth map with some contiguous region of missing depth, if it is determined to *not* be an attached shadow, we will then test to see if it is a detached shadow. We identify the two nearest shadow-bounding pixels as S_a and S_b, and also attempt to find the width of the foreground occluder. This is done by seeking for a sequential pair of large negative and positive depth deltas to the right of the shadow, located at F_a and F_b respectively, as illustrated in Fig. 2(b). With these four identified pixels, we are able to determine the length of shadow w_s, the length of the intervening background portion w_b and the length of the foreground w_f in the horizontal scanline. We then calculate $\widetilde{w}_{sb}$ and $\widetilde{w}_{bf}$ using (4) and (5). Finally, if $\widetilde{w}_{sb}$ is close to $(w_s + w_b)$ and $\widetilde{w}_{bf}$ is close to $(w_b + w_f)$, we classify the depth-missing region as a detached shadow.

2.3. Discussion

From Fig. 3 (a)&(b), it can be seen that a long foreground length leads to an attached shadow while a short foreground length results in a detached shadow. We are interested to find out the particular foreground length that separates the two types of shadows. That is the special case where the 3D line O_1F_1 intersects the 3D line O_2F_2 at the point B, as illustrated in Fig. 3(c). Let F_a and F_b denote respectively the projections of the left and the right edge points of the foreground object on its closest image plane of the IR camera. And G_b denotes the projection of the right edge point of the foreground object on its closest image plane of the virtual camera. Since $\overrightarrow{F_aF_b}$ is parallel to $\overrightarrow{O_1O_2}$, we have $\frac{|O_1O_2|}{|F_aG_b|} = \frac{z_b}{z_b - z_f}$, where $|O_1O_2|$ is the baseline between the IR projector and the IR camera. Considering the Kinect baseline is a constant value of 7.5 cm and it has a valid working range of $0.8\text{m} < z_f < z_b < 4\text{m}$, we obtain

$$|F_aF_b| < |F_aG_b| = 7.5(1 - \frac{z_f}{z_b})\ \text{cm}, \quad (6)$$

where $0.2 < \frac{z_f}{z_b} < 1$. Thus if the foreground and background objects have constant depth, a foreground length (the length projected on its closest image plane respect to IR camera, which is F_aF_b) less than $7.5(1 - \frac{z_f}{z_b})$cm will result in a detached shadow in the depth map. Taking into account $0.2 < \frac{z_f}{z_b} < 1$, we get $|F_aF_b| < |F_aG_b| < 6$ cm, which means any foreground length larger than 6 cm will generate an attached shadow in the depth map.

2.4. Imaging geometry of alignment shadows

Fig.3(d) illustrates the imaging geometry of alignment shadows, which occur when the original depth map from the IR camera is rendered from the viewpoint of the RGB camera, which we call the aligned depth map. Considering that the RGB camera is positioned between the IR projector and the IR camera, it is possible that parts of a shadow visible in the IR camera, such as B_1B_2 in Fig.3(d), will disappear behind the foreground object from the viewpoint of the RGB camera. Conversely, parts of the background such as B_3B_4 will be deoccluded and become visible in the RGB camera, resulting in additional holes to the right of the foreground objects in the aligned depth map. We call these

additional regions of missing depth as alignment shadows. Alignment shadow can be further classified into attached-alignment shadow and detached-alignment shadow. Using the similar discussions as Section2.3, we get that for any object with foreground length larger than 2cm will generate an attached-alignment shadow as Fig.3(d). Our detection algorithm for the alignment shadow is capable to detect both kinds of alignment shadows.

To detect the shadows in the aligned depth map, the shadows to the left of foreground objects can be directly identified by analyzing the original depth map. The shadows to the right of foreground objects are detected by computing if a split would occur in the aligned depth map for foreground and background pixels that would have been adjacent to each other in the original depth map. Note that the camera parameters for both RGB and IR cameras are calibrated using a method similar to that in [6].

3. Experiments for Kinect Shadow Detection

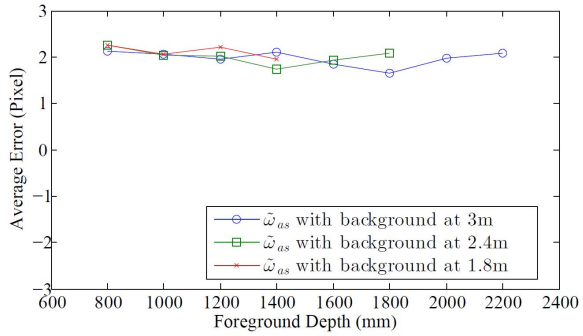
3.1. Accuracy of estimated feature lengths

We first compare the three computed feature lengths, $\widetilde{w}_{as}$, $\widetilde{w}_{sb}$ and $\widetilde{w}_{bf}$ using (3), (4) and (5) respectively, with the corresponding measured lengths w_{as}, w_{sb} and w_{bf} to quantitatively evaluate the accuracy. In particular, we shot different scenes with foregrounds and backgrounds located at different distances using three different Kinects. To focus on the accuracy evaluation, we use simple setup to generate only one type of shadows in one scene.

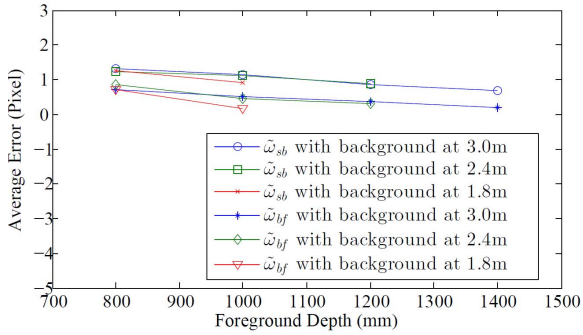
Fig. 4 shows the average errors between the estimated lengths and the corresponding measured lengths over the total number of corresponding shadows captured by the three Kinects. It can be seen that our length estimations are quite accurate since the average errors are always around $1 \sim 2$ pixels for all foreground and background distances. For attached shadows, there is a constant overestimation of around two pixels on the shadow length w_{as}, regardless of foreground and background distances. For detached shadows, we observe there is an overestimation of around one pixel for w_{sb} and about 0.5 pixel for w_{bf}. The constant overestimation is possibly due to the window operation used to convert IR image to depth map and the accumulated error from the inaccurate calibration for focal length.

3.2. Shadow detection performance

To evaluate the performance of the proposed shadow detection, we constructed a dataset consisting of 20 RGBD images with manually labeled ground truth for shadows. As far as we know, there is no existing dataset providing unaligned Kinect depth maps with intrinsic camera parameters in conjunction with ground truths for shadows. The 20 captured RGBD images can be classified into two categories: simple scenes and complex scenes. A scene in the former category contains only one object, while the images in the latter category contain multiple objects.



(a) Attached Shadow

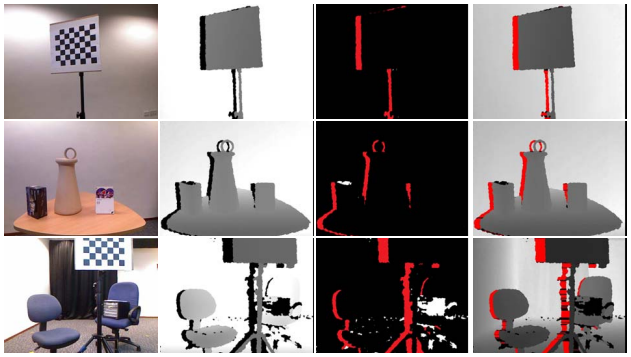


(b) Detached Shadow

Figure 4. Results of the average error between the estimated feature lengths and the corresponding measured lengths under different foreground and background distances.

Precision (P) and recall (R) measures are used as evaluation metrics. Let S_x be the detection result and S_{gt} be its ground truth. P is computed by $\frac{S_x \bigcap S_{gt}}{S_x}$, which measures the fraction of detected shadow pixels that were correct detections. R is computed by $\frac{S_x \bigcap S_{gt}}{S_{gt}}$, which measures the fraction of ground truth shadow pixels found. F-measure defined as $\frac{2PR}{R+P}$ is given as the overall score of each detection.

Fig. 5 shows a few examples for visually comparing the ground truth shadows and the detected shadows. We can see that the detected shadows match the ground truth quite well. More visual results of shadow detection can be found in Fig. 6(b). The precision and recall values are listed in Table 1. It can be seen that our proposed shadow detection achieves almost perfect precision scores for both simple and complex scenes, which means almost every shadow detected is a true shadow. Although there is a decline in the recall scores when going from simple to complex scenes, the recall performance is still acceptable. This decline is mainly due to situations when the shadows themselves are partially occluded by other objects and their full extent cannot be measured.



(a) RGB (b)Unaligned depth (c) GT (d) Result

Figure 5. Examples of shadow detection results. (c) Ground truth of the shadows. (d) The detection results. All ground truth and detected shadows are labeled in red color.

Table 1. Performance of the proposed shadow detection.

	Precision	Recall	F-measure
Simple	1.0000	0.9307	0.9641
Complex	0.9991	0.8369	0.9109
Overall	0.9996	0.8838	0.9381

3.3. An Application on Automatic Foreground Segmentation

In this section, we demonstrate the usefulness of the detected and classified shadow information by simply applying it for automatic foreground segmentation. Here, the foreground is considered to be one or more salient objects that are located closer to the Kinect, and that (more importantly) cast local shadows. To the best of our knowledge, there is no work that uses shadow for automatic general RGBD foreground segmentation.

Given an aligned depth map with classified shadow information, we can generate initial seeds based on the local patterns of each type of shadows. We select the seeds by moving a fixed distance in the direction given by shadow type from the likely edge pixels towards the foreground(background) side for foreground (background) seeds. We then perform K-means clustering on initial seeds to find the correct seeds on the salient object. Subsequently, two state-of-art segmentation algorithms, GrabCut (GC) [9] and Convex Active Contours (CAC) [7] are modified to segment RGBD image by introducing depth value as the 4th channel in the data term.

To evaluate the segmentation performance on Kinect RGBD images, we constructed a dataset of 50 RGB images with their depth maps and ground truths. Note that the existing Kinect RGBD datasets, such as [10], are either not designed for foreground segmentation or do not provide the original unaligned depth maps that are needed for our shadow detection.

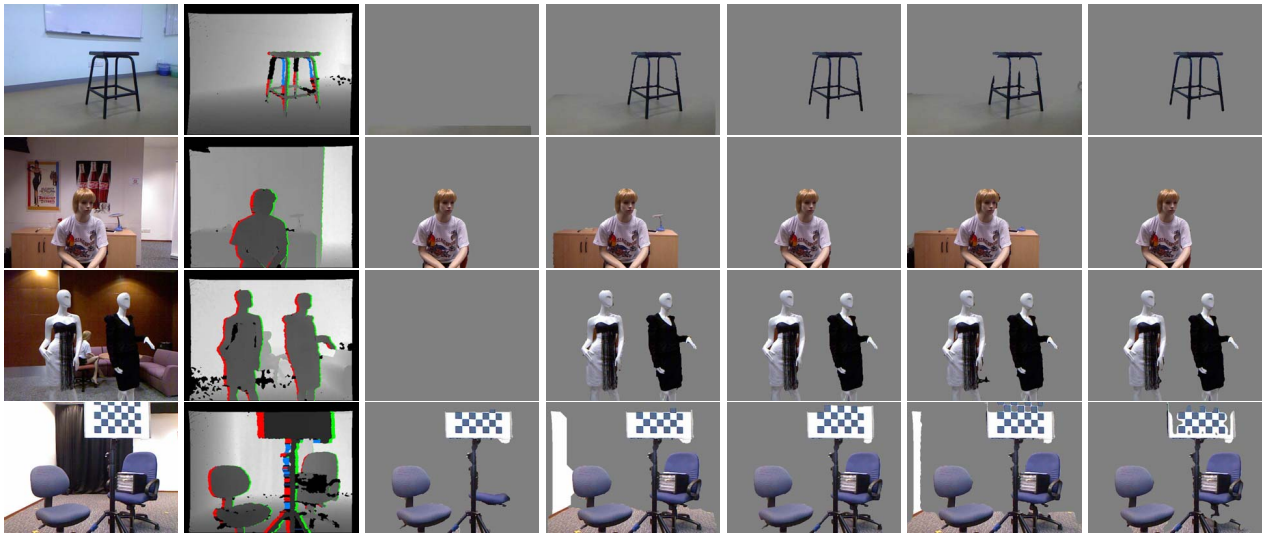
Since we are not aware of any existing work that can achieve fully automatic foreground segmentation on a single Kinect RGBD images, we compare our segmentation with an intuitive threshold-based automatic image segmentation method. Specifically, for the threshold-based method, the seeds are generated based on two thresholds T_f and T_b, where pixels with depth values smaller than T_f are classified as foreground seeds while pixels with depth values larger than T_b are considered as background seeds. The threshold-based method uses the same segmentation algorithms as ours. Fig. 6 shows the visual comparisons of the foreground segmentation results of different methods. It can be seen that the threshold-based methods can achieve high-quality segmentation for some cases when the thresholds happen to lead to accurate seeds, while there is no fixed set of thresholds (see Fig. 6(c)&(d)) that works for most of the images. In contrast, with the assistance of the shadow detection results illustrated in Fig. 6(b), our methods are able to smartly extract highly accurate seeds, which lead to very good segmentation results for almost all the images in the dataset. This clearly demonstrates the usefulness of the extracted shadow information for segmentation.

3.4. Limitations

Although our proposed shadow detection achieves very accurate detection results and the proposed shadow assisted segmentation achieves very good segmentation performance, there are still some limitations. The last row of Fig. 6 shows a partially successful case, in which quite a few shadows were not detected and the segmentation results contained considerable errors. The failure to detect shadows is mainly a result of the shadows either overlapping with another region of missing depth or partially blocked by other foreground objects. The noise in the shadow edges may also cause problems in the proposed shadow detection. For the shadow-assisted segmentation, problems can arise when there is erroneous depth, or when non-shadow Kinect holes are co-located with regions with little color contrast between foreground and background. In such circumstances, our shadow assisted segmentation methods cannot achieve a clean foreground cutout.

4. Conclusions

In this paper, we presented a robust Kinect shadow detection scheme which can detect three types of shadows accurately, which are attached shadows, detached shadows and alignment shadows. We also demonstrated the usefulness of the extracted shadow information by applying it to automatic foreground segmentation, where the shadow detection results are used to generate accurate foreground and background seeds. We believe the extract shadow information can be used in many other applications such as depth re-



(a) RGB (b) C-Depth (c) GC-threshold1 (d) GC-threshold2 (e) GC-shadow (f) CAC-threshold2 (g) CAC-shadow

Figure 6. Visual comparisons of foreground segmentation results of different automatic approaches. (a) Color image. (b) Aligned depth map with classified shadows in red, blue and green colors for attached, detached and alignment shadows respectively. (c) Threshold-based GrabCut with $T_f = 0.3$ m and $T_b = 0.5$ m. (d) Threshold-based GrabCut with $T_f = 0.5$ m and $T_b = 0.6$ m. (e) Shadow assisted GrabCut. (f) Threshold-based CAC with $T_f = 0.5$ m and $T_b = 0.6$ m. (g) Shadow assisted CAC.

covery and scene understanding since it conveys high-level structure information.

Acknowledgement

This research, which is carried out at BeingThere Centre, is supported by MoE AcRF Tier-1 Grant RG30/11 and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

[1] Y. Berdnikov and D. Vatolin. Real-time depth map occlusion filling and scene background restoration for projected-pattern based depth cameras. In *Graphic Conf., IETP*, 2011.

[2] M. Daum and G. Dudek. On 3-d surface reconstruction using shape from shadows. In *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pages 461–468. IEEE, 1998.

[3] R. Feris, R. Raskar, L. Chen, K.-H. Tan, and M. Turk. Multiflash stereopsis: Depth-edge-preserving stereo with small baseline illumination. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(1):147–159, 2008.

[4] V. Gulshan, V. Lempitsky, and A. Zisserman. Humanising grabcut: Learning to segment humans using the kinect. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1127–1133, 2011.

[5] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011.

[6] K. Khoshelham and S. O. Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012.

[7] T. N. A. Nguyen, J. Cai, J. Zhang, and J. Zheng. Robust interactive image segmentation using convex active contours. *Image Processing, IEEE Transactions on*, 21(8):3734–3743, 2012.

[8] R. Raskar, K.-H. Tan, R. Feris, J. Yu, and M. Turk. Non-photorealistic camera: depth edge detection and stylized rendering using multi-flash imaging. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 679–688. ACM, 2004.

[9] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[10] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012*, pages 746–760. Springer, 2012.

[11] K. Vaiapury, A. Aksay, and E. Izquierdo. Grabcutd: improved grabcut using depth information. In *Proceedings of the 2010 ACM workshop on Surreal media and virtual cloning*, pages 57–62. ACM, 2010.

[12] J. Yang, X. Ye, K. Li, and C. Hou. Depth recovery using an adaptive color-guided auto-regressive model. In *Computer Vision–ECCV 2012*, pages 158–171. Springer, 2012.