Consistent Depth Maps Recovery from a Trinocular Video Sequence

Supplementary Material

Wenzhuo Yang[1] Guofeng Zhang[1] Hujun Bao[1] Jiwon Kim[2] Ho Young Lee[2]

[1]State Key Lab of CAD&CG, Zhejiang University [2]Samsung Advanced Institute of Technology

1. Iterative Optimization of Trinocular Stereo Matching Model

In this section, we will discuss how to solve the proposed trinocular stereo matching model in details.

We apply an iterative optimization algorithm to minimize the energy function (1) defined in our paper, and compute disparity and occlusion by Belief Propagation algorithm. Similar to [2], the optimization process iterates between two steps: 1) estimate occlusion given disparity, and 2) estimate disparity given occlusion. For simplicity, $O(\mathbf{x})$ and $D(\mathbf{x})$ are denoted as $o_{\mathbf{x}}$ and $d_{\mathbf{x}}$, respectively.

1.1. Estimate occlusion given disparity

Given the estimated disparity map D, the original energy function (i.e. Equation (1) in our paper) can be simplified as:

$$\begin{aligned} E_O = &\sum_{\mathbf{x}} (1-O_L(\mathbf{x}))(1+O_R(\mathbf{x}))\rho(\mathbf{x},d_{\mathbf{x}};I_L,I_M)+ \\ &\sum_{\mathbf{x}} (1-O_R(\mathbf{x}))(1+O_L(\mathbf{x}))\rho(\mathbf{x},d_{\mathbf{x}};I_R,I_M)+ \\ &\sum_{\mathbf{x}} O_L(\mathbf{x})O_R(\mathbf{x})\eta+ \\ &\sum_{\mathbf{x}} \beta_\omega(|O_L(\mathbf{x})-W_L(\mathbf{x};D)|+|O_R(\mathbf{x})-W_R(\mathbf{x};D)|)+ \\ &\sum_{\mathbf{x}}\sum_{\mathbf{y}\in N(\mathbf{x})} \beta_o(|O_L(\mathbf{x})-O_L(\mathbf{y})|+|O_R(\mathbf{x})-O_R(\mathbf{y})|), \end{aligned} \tag{1}$$

where $N(\mathbf{x})$ is the set of all adjacent pixels for $\mathbf{x}$. The first three equations constitute the data term $E_d(D,O;I)$, and the last two equations constitute the visibility term $E_v(D,O;I)$.

Since this energy function contains two variables O_L, O_R and a quadratic term $O_L(\mathbf{x})O_R(\mathbf{x})$, it cannot be minimized conveniently, so that we combine O_L and O_R into a single variable O defined as $O(p) = 2O_R(p) + O_L(p)$. Therefore, the data term at pixel $\mathbf{x}$ has the following four cases:

1. $O(\mathbf{x}) = 0$ when $O_L(\mathbf{x}) = 0$ and $O_R(\mathbf{x}) = 0$, then $E_d^0(\mathbf{x},d) = \rho(\mathbf{x},d;I_L,I_M) + \rho(\mathbf{x},d;I_R,I_M)$.
2. $O(\mathbf{x}) = 1$ when $O_L(\mathbf{x}) = 1$ and $O_R(\mathbf{x}) = 0$, then $E_d^1(\mathbf{x},d) = 2\rho(\mathbf{x},d;I_R,I_M)$.
3. $O(\mathbf{x}) = 2$ when $O_L(\mathbf{x}) = 0$ and $O_R(\mathbf{x}) = 1$, then $E_d^2(\mathbf{x},d) = 2\rho(\mathbf{x},d;I_L,I_M)$.
4. $O(\mathbf{x}) = 3$ when $O_L(\mathbf{x}) = 1$ and $O_R(\mathbf{x}) = 1$, then $E_d^3(\mathbf{x},d) = \eta$.

And the first term in the visibility term at pixel $\mathbf{x}$ also has four cases:

1. $O(\mathbf{x}) = 0$, then $E_v^0(\mathbf{x},d) = \beta_\omega(W_L(\mathbf{x};D) + W_R(\mathbf{x};D))$.
2. $O(\mathbf{x}) = 1$, then $E_v^1(\mathbf{x},d) = \beta_\omega(1 - W_L(\mathbf{x};D) + W_R(\mathbf{x};D))$.
3. $O(\mathbf{x}) = 2$, then $E_v^2(\mathbf{x},d) = \beta_\omega(1 + W_L(\mathbf{x};D) - W_R(\mathbf{x};D))$.
4. $O(\mathbf{x}) = 3$, then $E_v^3(\mathbf{x},d) = \beta_\omega(2 - W_L(\mathbf{x};D) - W_R(\mathbf{x};D))$.

We denote $E_i(\mathbf{x},d)$ as $E_i(\mathbf{x},d) = E_d^i(\mathbf{x},d) + E_v^i(\mathbf{x},d)$. Then the energy function (1) can be rewritten as:

$$\begin{aligned} E_O = &\sum_{\mathbf{x}} (T(o_{\mathbf{x}}=0)E_0(\mathbf{x},d_{\mathbf{x}}) + T(o_{\mathbf{x}}=1)E_1(\mathbf{x},d_{\mathbf{x}}))+ \\ &\sum_{\mathbf{x}} (T(o_{\mathbf{x}}=2)E_2(\mathbf{x},d_{\mathbf{x}}) + T(o_{\mathbf{x}}=3)E_3(\mathbf{x},d_{\mathbf{x}}))+ \\ &\sum_{\mathbf{x}}\sum_{\mathbf{y}\in N(\mathbf{x})} \beta_o(\phi(o_{\mathbf{x}},o_{\mathbf{y}})), \end{aligned} \tag{2}$$

where $N(\mathbf{x})$ is the set of all adjacent pixels for $\mathbf{x}$, and T is the indicator function that returns 1 if its argument is true and 0, otherwise. Function $\phi(x,y) = \lceil (x \otimes y)/2 \rceil$, $x,y \in \{0,1,2,3\}$, where $\otimes$ is the "exclusive or" operator. With our definition of $E_i(\mathbf{x},d)$, the unary terms in

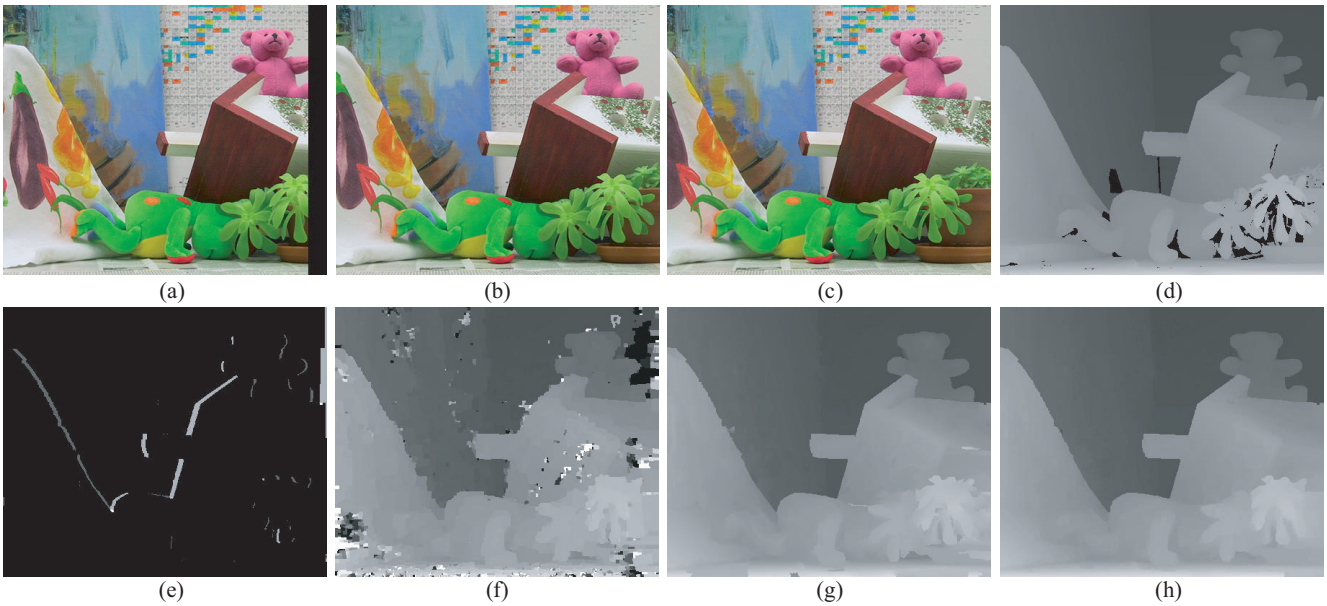

Figure 1. "Teddy" example. (a) The left image. (b) The center image. (c) The right image. (d) The ground truth of disparity map. (e)The estimated occlusion map. White pixels are occluded in the right image and gray pixels are occluded in the left image. (f) The estimated disparity map without plane fitting. (g) The estimated disparity map by plane fitting with one segmentation result. (h) The final refined disparity map after fusion refinement.

Equation (1) and Equation (2) are equivalent. To see the binary terms are also equivalent, we give a simple proof. Because $O_L(\mathbf{x}), O_R(\mathbf{x}) \in \{0,1\}$, $|O_{L/R}(\mathbf{x}) - O_{L/R}(\mathbf{y})| = O_{L/R}(\mathbf{x}) \otimes O_{L/R}(\mathbf{y})$. We denote $|O_L(\mathbf{x}) - O_L(\mathbf{y})| + |O_R(\mathbf{x}) - O_R(\mathbf{y})|$ as $S(\mathbf{x}, \mathbf{y})$, so that we have

$$\begin{aligned} S(\mathbf{x},\mathbf{y}) &= (O_L(\mathbf{x}) \otimes O_L(\mathbf{y})) + (O_R(\mathbf{x}) \otimes O_R(\mathbf{y})) \\ &= (o_\mathbf{x} \otimes o_\mathbf{y}) - 2\lfloor (o_\mathbf{x} \otimes o_\mathbf{y})/2 \rfloor + \lfloor (o_\mathbf{x} \otimes o_\mathbf{y})/2 \rfloor \\ &= (o_\mathbf{x} \otimes o_\mathbf{y}) - \lfloor (o_\mathbf{x} \otimes o_\mathbf{y})/2 \rfloor \\ &= \lceil (o_\mathbf{x} \otimes o_\mathbf{y})/2 \rceil \\ &= \phi(o_\mathbf{x}, o_\mathbf{y}), \end{aligned} \tag{3}$$

Therefore, Equation (1) and Equation (2) are equivalent. Then we can apply the BP algorithm to approximately minimize (2) to solve the occlusion.

1.2. Estimate disparity given occlusion

Similar to [2], given the estimated occlusion O, the original energy function can be simplified as:

$$\begin{aligned} E_D =& E_d(D; O, I) + \\ & \sum_\mathbf{x} \beta_\omega (O'_L(P_{M\to L}(\mathbf{x}, d_\mathbf{x})) + O'_R(P_{M\to R}(\mathbf{x}, d_\mathbf{x}))) + \\ & \sum_{\mathbf{x},\mathbf{y}} \lambda(\mathbf{x},\mathbf{y}) \rho_s(d_\mathbf{x}, d_\mathbf{y}). \end{aligned} \tag{4}$$

The first term is the data term and the second term is the visibility term. $P_{M\to L}(\mathbf{x}, D(\mathbf{x}))$ is a projection function, projecting $\mathbf{x}$ onto the left view based on the disparity $D(\mathbf{x})$. Binary map $O'_L(\mathbf{x}')$ indicates whether pixel $\mathbf{x}'$ in the left view is occluded in the center view. $O'_L(\mathbf{x}')$ equals 1 if $\mathbf{x}'$ is occluded and 0, otherwise. O'_L can be computed by warping all the pixels in the middle view by using the estimated disparity map in the last iteration. To reduce noises in O'_L, a mean filter is applied to it. If the value $O'_L(\mathbf{x})$ after applying the filter is less than a threshold, it is set to 0. $P_{M\to R}(\mathbf{x}, D(\mathbf{x}))$ and O'_R are defined in a similar way. The third term is the smoothness term.

As we have mentioned in the paper, we use BP algorithm to minimize Equation (4) to solve the disparity map in this iteration. Then the estimated disparity map is refined by fitting disparity segments to a set of 3D planes, using the same plane fitting technique introduced in [3]. However, the refined disparity map may contain errors if the segmentation information is imperfect. Finally, we propose to fuse different disparity maps estimated under a variety of segmentation results generated by different segmentation parameters. Specifically, we choose a set of different mean-shift parameters to generate k disparity maps $\{D_1, D_2, \cdots, D_k\}$ which form the disparity candidate set $\hat{D}$. In addition, we also compute the average value (i.e. $D_{k+1}(\mathbf{x}) = \sum_{i=1}^{k} D_i(\mathbf{x})/k$), and add it into $\hat{D}$. Finally, with these proposals, the disparity map is re-estimated by minimizing (4).

In summery, our iterative optimization algorithm alternates between these two steps. The occlusion O are initially set to zeros, and the initial disparity map D is computed by minimizing Equation (4) without the visibility term.

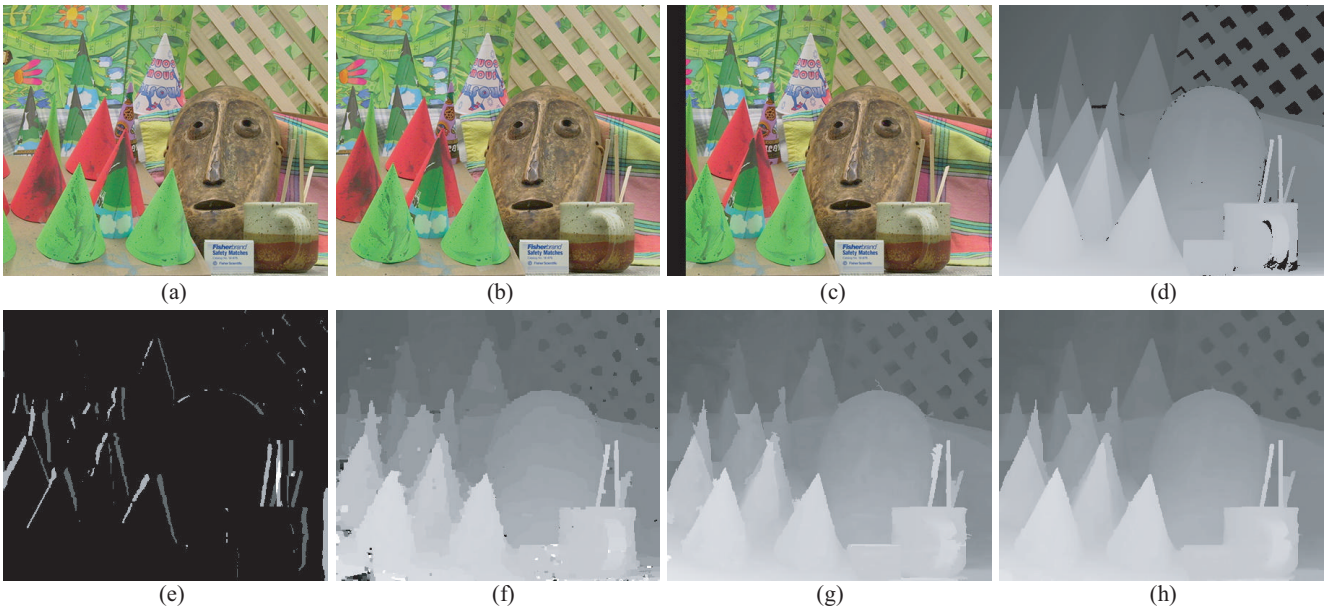

Figure 2. "Cones" example. (a) The left image. (b) The center image. (c) The right image. (d) The ground truth of disparity map. (e)The estimated occlusion map. White pixels are occluded in the right image and gray pixels are occluded in the left image. (f) The estimated disparity map without plane fitting. (g) The estimated disparity map by plane fitting with one segmentation result. (h) The final refined disparity map after fusion refinement.

1.3. More Results of Trinocular Stereo Matching

We show more results with Middlebury stereo data [1]. Figure 1 shows the results of "Teddy" example. Figures 1(a)-(c) are the left, center and right images, respectively. Although there are some pixels without color information in the left image, our method still can recover a high-quality disparity map, as shown in Figure 1(h).

Figure 2 shows the "Cones" example. Figures 2(a)-(c) are the left, center and right images. Figures 2(f)-(h) are the estimated disparity maps of "Cones" example in different procedures. As can be seen, after fusion refinement, the disparity map is significantly improved.

Faithfully, the disparity maps estimated by our method are very closed to the ground truths. Besides high-quality results, our method is very efficient. The running time of these two examples is about 40s. In our implementation, we use GPU to accelerate the process of computing stereo matching data cost.

2. Modified Bundle Optimization Model

The data term $E_{d_2}(D;I)$ for depth refinement of static pixels is defined as follows:

$$E_{d_2}(D;I) = E_d(D,O;I) + \sum_{\mathbf{x}}(1-u(\mathbf{x})\cdot \sum_{t'} p_c(\mathbf{x},D(\mathbf{x}),I,I_{t'})p_v(\mathbf{x},D(\mathbf{x}),D_{t'})),$$

[1] http://vision.middlebury.edu/stereo/

where p_c is the same to Equation (2) in [3], and geometric coherence term p_v is similar to Equation (8) in [3]. $u(\mathbf{x})$ is the normalization factor defined the same as [3]. The minor difference is that geometric coherence term in [3] is defined in image space. Here we propose to measure the geometric coherence in disparity space by

$$p_v(\mathbf{x},D(\mathbf{x}),D_{t'})) = \exp(-\frac{||h(D(\mathbf{x}))-D_{t'}(\mathbf{x}')||^2}{2\sigma_v^2}), \quad (5)$$

where $\mathbf{x}'$ is the projected pixel in frame t', and $D_{t'}(\mathbf{x}')$ is its estimated disparity. Here $h(D(\mathbf{x}))$ is the transformed disparity value considering the camera parameters, i.e. the corresponding disparity value after projecting pixel $\mathbf{x}$ to frame t'. Assuming the projective matrix from reference frame to frame t' is $[R|t]$, the intrinsic matrix of reference frame is K, the 2D position of $\mathbf{x}$ is (u,v). Then $h(D(\mathbf{x}))$ can be computed by

$$h(D(\mathbf{x})) = \frac{1}{(\frac{1}{D(\mathbf{x})}RK^{-1}(u,v,1)^\top + t)[3]}, \quad (6)$$

where [3] denote the third element of the vector. The intuition is that the corresponding pixels in different frames should have the same 3D position. σ_d is a parameter. In our experiments, $\sigma_d = 0.02(d_{\max}-d_{\min})$, where $[d_{\min}, d_{\max}]$ is the disparity range.

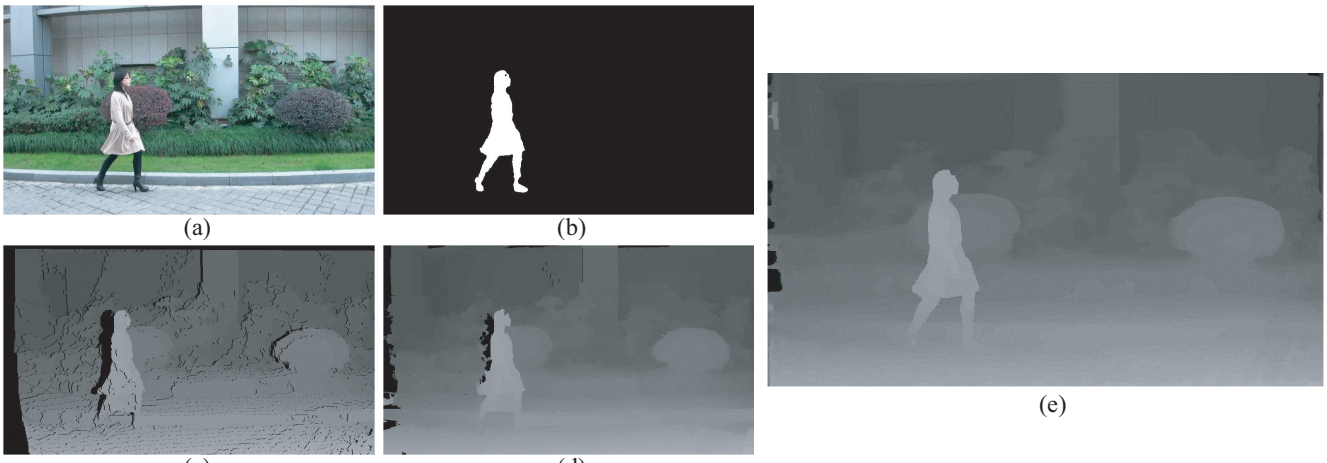

Figure 3. Disparity estimation of the left sequence. (a) One selected image of the left sequence. (b) The warped segmentation mask from the center view to the left view. (c) The warped disparity map from the center view to the left view. (d) The refinement disparity map by plane fitting. (e) The optimized disparity map after spatio-temporal optimization.

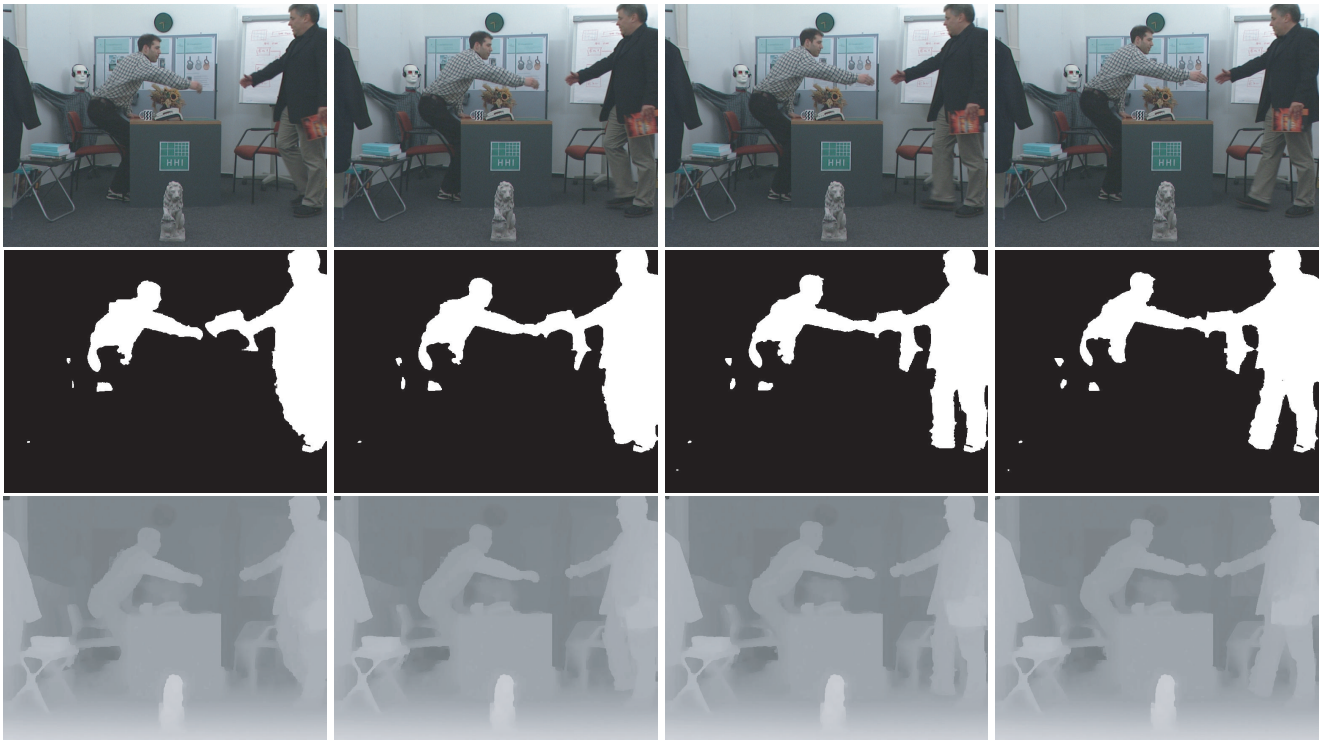

Figure 4. "Book Arrival" example. The images in the first row are the four frames in the center sequence. The segmentation results of the dynamic regions are shown in the second row. The third row shows the estimated disparity maps.

3. Intermediate Results of Depth Estimation for Left and Right Sequences

Besides recovering the depth maps of the center sequence, we also need to estimate the depth maps of the left and right sequences. The estimated segmentation masks and depth maps in the center sequence are warped to the left and right sequences. Figures 3(b)-(c) show the warped results. The warped depth map contains some holes due to the missing pixels caused by occlusion. We combine segmentation information to infer the depth values of these pixels. As shown in Figure 3(c), the missing pixels are mostly in the static regions. Then mean-shift algorithm is applied to segment the left/right image. If a segment contains both static and dynamic pixels (based on the warped moving ob-

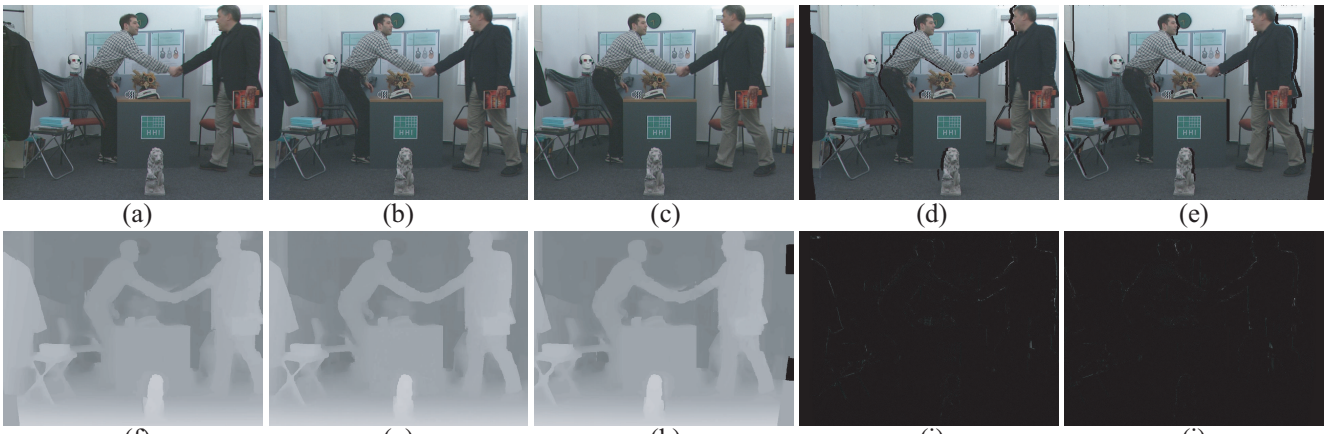

Figure 5. Depth accuracy verification. (a) One selected image in the left sequence. (b) One selected image in the center sequence. (c) One selected image in the right sequence. (d) Warping the center image to the left one. (e) Warping the center image to the right one. (f) The estimated depth map of (a). (g) The estimated depth map of (b). (h) The estimated depth map of (c). (i) The difference between (a) and (d). (j) The difference between (c) and (e).

ject masks), it further splits. By assuming the 3D surface of each segment can be approximated by a 3D plane, we fit a 3D plane for each segment which has missing pixels, so that the disparities of the missing pixels can be inferred. As shown in Figure 3(d), most missing pixels are completed. However, some segments in the occluded regions are mistakenly fitted due to the lack of sufficient pixels with accurate disparities. Finally, we use our spatio-temporal optimization method to further refine the left and right depth maps. Figure 3(e) shows an estimated depth map of one image from the left sequence.

4. Results of Video Sequences Captured by Stationary Trinocular Cameras

Figure 4 shows our estimated results of "book arrival" example [1] where the capturing trinocular cameras are stationary. Since the dynamic people do not have sufficient movement, the estimated static background information is incomplete, so that some regions could not be accurately segmented. Even with this imperfect segmentation result, our method still can faithfully recover the high-quality depth maps due to the following two reasons. First, our spatio-temporal optimization for dynamic pixels also can be used for static ones. Therefore, if a background region is recognized as moving region, it may not harm the depth estimate. Second, in bundle optimization, the data term only associates one frame with about 6-8 neighboring frames. For a dynamic pixel with small movement, since it can be approximated as static pixels in a short time, its disparity still can be effectively optimized by bundle optimization. To verify the accuracy of the recovered depths, we warp one center image to the left/right views with the recovered depth map. Figures 5(d)-(e) shows the warped images, and Figures 5(i)-(j) shows the difference images, which demonstrate the accuracy of the estimated depth maps.

References

[1] I. Feldmann, M. Mller, F. Zilly, R. Tanger, K. Mller, A. Smolic, P. Kauff, and T. Wiegand. HHI test material for 3D video. *ISO/IEC JTC1/SC29/WG11, MPEG08/M15413*, 2008.

[2] J. Sun, Y. Li, S. B. Kang, and H.-Y. Shum. Symmetric stereo matching for occlusion handling. In *CVPR*, 2005.

[3] G. Zhang, J. Jia, T.-T. Wong, and H. Bao. Consistent depth maps recovery from a video sequence. *TPAMI*, 31(6), 2009.