STEREO RANDOM FIELD FOR BI-LAYER IMAGE SEGMENTATION

Kuo-Chin Lien and Jerry D. Gibson

Department of Electrical and Computer Engineering, University of California, Santa Barbara, USA
{kuochin, gibson}@ece.ucsb.edu

ABSTRACT

Stereo image segmentation usually incorporates depth cues to achieve high quality. However, previous methods that pointwise propagate information within stereo pairs could suffer from a poorly estimated depth map. In this paper, we introduce a novel graphical model where a greater amount of reliable messages can be conveyed during two-view joint segmentation. This model leads to a strongly coupled stereo pair, thus improving robustness, accuracy and consistency of stereo segmentation. Additionally, we augment a depth map to a novel correspondence matrix which is suitable for the proposed stereo segmentation model. Our experiments on a public stereo dataset show that the proposed correspondence method and stereo model outperforms state-of-the-art stereo segmentation algorithms.

Index Terms— stereo, bi-layer segmentation, random field

1. INTRODUCTION

Obtaining a consistent segmentation to the visual objects in a stereo pair is the key to accurate 3D manipulation. Bi-layer segmentation algorithms developed in the Microsoft i2i project have been applied to automatic background substitution and view generation in 3D visual communication [4]. Lim et al. [6] proposed a stereo inpainting algorithm based on stereo bi-layer segmentation. Lo et al. [5] introduced an interactive image editing system which enables users to select and clone objects of interest in 3D images. More recently, Price and Cohen developed an interactive object selection tool that further allows users to refine their selection in both views [3].

To separate the foreground object from the background scene, current stereo segmentation algorithms usually consider depth as a strong cue. In the i2i system [4], a pre-computed depth map was used to calculate the likelihood of pixels belonging to the foreground. This likelihood, combined with color and contrast cues, can then be used to infer and extract the foreground layer. However, the quality of the bi-layer segmentation and consequent 3D image rendering strongly relies on the accuracy of input depth maps. Lo et al. exploited depth maps in a different way [5]. In their system, the foreground region in the left view is first segmented with user's interaction and then propagated to the right view by using a depth map and Bai et al.'s local window technique [1]. To be robust against an imperfect depth map, only the pixels with coherent disparities between the two views are propagated. This approach did not propagate any information from the right view to help the segmentation in the left.

Instead of segmenting one view and then the other, Price and Cohen's StereoCut system [3] performed a joint segmentation of the two views. Disparities were used to define a pairwise term in a conditional random field (CRF) framework where the two views are modeled as a single graph. After a global optimization, foreground and background labels were then simultaneously assigned to all pixels in the two views. Again, to avoid the harm of an imperfect depth map, those pixels not passing a consistency check (hereinafter referred to as "lone pixels") are not recommended to be considered during the construction of the pairwise term. The problem of this method is that while a consistency check rejects unreliable disparities, the remaining depth information may be insufficient to make stereo segmentation consistent. Even though a perfect disparity map is provided, occlusions may still prevent some pixels from finding correspondence in the other view.

Addressing these issues, we propose a high-order approach that more effectively utilizes available depth information on a depth map than previous methods. In the proposed approach, a dense depth map is first translated to a higher order correspondence matrix. With this correspondence matrix, the proposed stereo high-order model encourages label consistency among groups of pixels between the two views so as to improve the accuracy of stereo segmentation. Enjoying the same advantage of high-order single-view segmentation algorithms [7, 8], our stereo segmentation can handle complicated object boundaries in each view. More importantly, compared to prior work on stereo segmentation, a larger coverage of pixels can find a reliable set of corresponding points in the other view for maintaining good consistency during segmentation.

In the next section, we will describe our stereo random field which models high-order interaction between two views while performing segmentation. In Section 3, we describe how to augment a given depth map to achieve this

high-order correspondence. In Section 4, we verify our approach on a public dataset.

2. HIGH-ORDER CONSISTENCY FOR STEREO BI-LAYER SEGMENTATION

In this section, we present our high-order stereo segmentation model, which maintains better consistency on the segmentation of stereo image pairs. We rely on a well-known observation that a locally quantized segment (superpixel) is more meaningful and robust to noise than a pixel [9, 10]. This observation implies that a more stable relation between the two views can be established on local clusters.

2.1 Segmentation using CRF models

CRF [11] is a powerful framework which builds graphical models to solve image segmentation problems in a probabilistic way. In this framework, each node is usually associated with a pixel, and segmenting an image is equivalent to assigning proper labels to the nodes in a graph. A commonly used second order model is defined as follows:

$$E(x)=\sum_{i\in I}\psi_i(x_i)+\sum_{\substack{i\in I\\ j\in Neighbor}}\psi_{ij}(x_i,x_j) \qquad (1)$$

where ψ_i is the unary potential function, inferring the likelihood of label x_i by the associated observation z_i on the i^{th} pixel in image I, and ψ_{ij} is a pairwise term that models the relation between two adjacent pixels. By minimizing the energy function E(x) via graphcut [13], an optimal segmentation boundary can be determined.

For bi-layer segmentation, to label $x\in$ {foreground, background} a popular choice of the observation z is color. With this choice, the unary potential can be written as follows:

$$\begin{aligned}\psi_i &= w^{color}\cdot(\psi^{CF}-\psi^{CB})\\ \psi^{CF} &= -\log(P^{CF})\\ \psi^{CB} &= -\log(P^{CB})\end{aligned} \qquad (2)$$

where P^{CF} and P^{CB} are the likelihood of a pixel belonging to the foreground and background, which can be obtained from examining foreground and background color models on input image I. The parameter w^{color} is a weight. While the unary term describes how a label is affected by the color observation of its associated pixel, the pairwise term ψ_{ij} can be used to encourage neighboring pixels with similar color to have the same label:

$$\psi_{ij}=\left|x_i-x_j\right|(\lambda_1+\lambda_2(\exp(-\left\|\frac{z_i-z_j}{\beta}\right\|)))$$

where λ_1 and λ_2 are two parameters.

2.2 High-order CRF for stereo segmentation

To extend this model for bi-layer stereoscopic segmentation, we note that the pairwise term in this framework serves as a smoothness function, smoothing the labels of two corresponding pixels. Thus, a natural approach to maintain the consistency of stereoscopic segmentation is adding a pairwise energy term to smooth the labels associated with the two pixels that are related via a depth map. The energy function can be written as follows:

$$\begin{aligned}E(x)=&\sum_{i\in I_R,I_L}\psi_i(x_i)+\sum_{\substack{i\in I_R,I_L\\ j\in Neighbor}}\psi_{ij}(x_i,x_j)\\ &+\sum_{\substack{i\in I_R,I_L\\ k\in Correspondence}}\psi_{ik}(x_i,x_k)\end{aligned} \qquad (3)$$

where the last term is governed by the given depth maps.

In order to be robust against a noisy depth map, either due to occlusion or the errors in stereo matching algorithms, [3] further applies a consistency check between the two depth maps estimated for the left and the right views. This check removes unreliable depth correspondence between the two views if one can correspond a pixel *i* to a pixel *k* via the left depth map but cannot map it back via the right depth map. While this check indeed improves the segmentation quality, a vast number of non-corresponding pixels (lone pixels) cannot directly benefit from this model.

Addressing this issue, we propose the following high-order stereoscopic model, which more generally associates a group of pixels to another group, or more than one group, in the other view. With the proposed correspondence scheme detailed in Section 3, we can then help lone pixels find a set of proper correspondences in the other view, thus improving the consistency of the pixel-level segmentation.

$$\begin{aligned}E(x)=&\sum_{i\in I_R,I_L}\psi_i(x_i)+\sum_{\substack{i\in I_R,I_L\\ j\in Neighbor}}\psi_{ij}(x_i,x_j)\\ &+w_h\cdot\sum_{m,n\in C}\psi^C_{mn}(\mathbf{x}_m,\mathbf{x}_n)\end{aligned} \qquad (4)$$

In our model, with a weight the high-order term ψ^C encourages a group of pixels $\mathbf{x}_m$ to have the same label as a group of pixels $\mathbf{x}_n$ in the other view. This many-to-many linking relies on a more stable and robust correspondence map C (detailed in Section 3) and allows us to build a stronger connection between the two views, thus achieving better segmentation consistency and accuracy. An illustration of our model is shown in Fig. (1a).

2.3 Relation to the other CRF models

The proposed stereo model can be seen as a generalization of [3]. We argue that there might be two reasons why previous models cannot effectively associate two views: 1) while consistency checking rejects unreliable linking between the two views, the resultant, loosely coupled graph

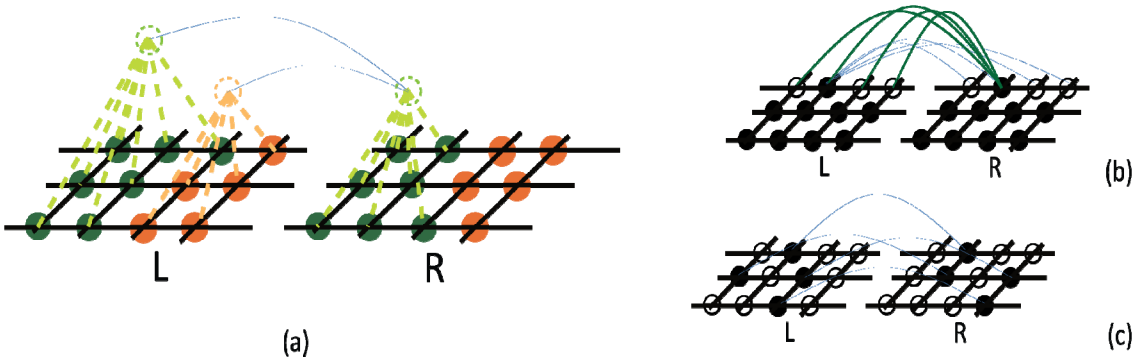


Fig. 1. Our and previous stereo random fields. Our model (a) encourages cross-view group consistency instead of pixel consistency, which was pursued in (b) and (c). To efficiently build connections among groups of pixels, we create an auxiliary node (the dashed circle) as a representative for a group of pixels. Note that a group can be softly related to multiple groups. While (b) and (c) suffer from the poor depth estimation (hollow circles indicate no reliable depth information), the proposed model is more robust to noisy depth maps and maintains better consistency for the stereo segmentation.

is not enough to make the segmentation consistent; 2) as one variety in [3] (named "Consistent PDF") also tried to associate one pixel to multiple possible corresponding pixels in the other view, the results were not significantly improved. This is because Consistent PDF only searches for correspondence along epipolar lines by using the output distribution of stereo matching. It is not surprising that the above-mentioned two correspondence schemes fail wherever the stereo matching algorithm fails, though multiple correspondence mitigates the hurt of incorrect stereo matching. In summary, previous methods strongly rely on the result of stereo matching but lack a mechanism to compensate the error of the stereo matching algorithm. With this consideration, we have designed our stereo model as a strongly coupled graph where nearly every pixel can be associated to multiple correspondences, not restricted on the epipolar line, in the other view. A comparison of our and previous stereo segmentation models is shown in Fig. 1. We will detail our correspondence map between the two views in the next section to replace the conventional depth map.

Once we obtain a mapping function to correspond a group of pixels in the right view to another group in the left, a computational issue may arise if the two groups both have large sizes, say M and N. The subgraph with M×N nodes will have complicated M×N links and exhausts computational power to solve it. Instead of building M×N pairwise links between the two groups containing N and M pixels respectively, we create an auxiliary node for each group similar to [13], as they did it for quickly solving a given graph and we aim to define a strongly coupled graph. Similar techniques also have been utilized in [7, 8], both of which deployed a node for each superpixel. An essential difference is that [7] and [8] target intra-frame smoothness, and we explore cross-view consistency.

3. CONSTRUCTING STRONG CONNECTIONS BETWEEN STEREOSCOPIC VIEWS

In order to benefit from the above high-order stereo model, we construct the following group-to-group correspondence by augmenting a given depth map.

3.1 Local clustering

Our strategy is to form locally coherent segments on both views and then associate them. Grouping pixels locally has been empirically proven to be helpful for subsequent single-image segmentation [7, 8] and video segmentation [10] while a segment usually has more stable and meaningful characteristics than a pixel. This also helps in our case: exploring stable correspondences between the two views rather than directly using a noisy depth map. In our implementation, the mean-shift algorithm [12] is used to perform this pixel grouping in Lab color space. Typically, about one thousand segments can be formed in one image.

3.2 Group correspondences guided by depth map

With locally grouped pixels, our aim is to find the best match among the clusters in the two views so as to associate them in our stereo random field. To achieve this long-range association, one generally needs to define a similarity function between two groups of pixels and exhaustively compare each pair of groups to find the best match [14]. For stereo, however – a special case where the two images inherently strongly relate to each other – reliable cross-view matching can be conducted by using an imperfect depth map.

Given a stereo pair partitioned into locally coherent regions $\{c^L_m\}$ and $\{c^R_n\}$ indexed by *m* and *n* for the left and right views, we now detail how to correspond them and accordingly approximate the high-order potential ψ^C_{mn} by auxiliary variables. As the first step, for every segment *c*, we create an auxiliary node *y* and bind this node to all the containing pixel-level nodes **x** with strong links. This auxiliary node will serve as a representative to interact with the corresponding clusters in the other view. Building the intra-cluster links can be seen as building the following potential function on x and y:

$$\psi(y_m, x) = \sum_{x \text{ in } c_m} w_\theta \cdot | x - y_m | \qquad (5)$$

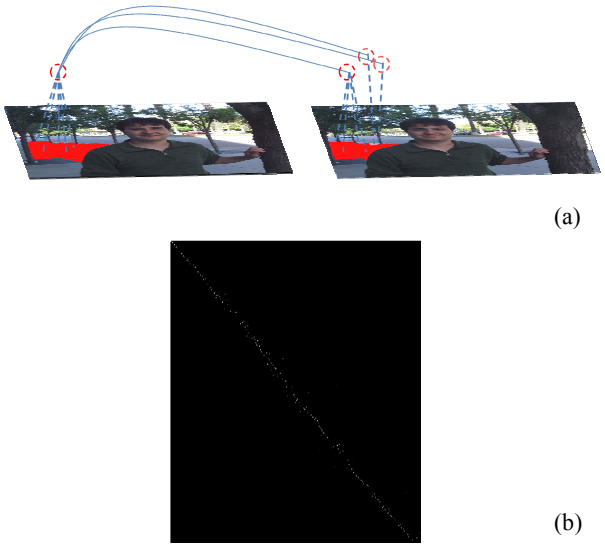
where w_θ is a large number.

We then consider a local cluster c^L_m in the left view containing $M = M^{corr}_m + M^{lone}_m$ pixels, where M^{corr} is the number of the pixels that can find a reliable correspondence via the given depth map and M^{lone} is the number of lone pixels. We define the following potential function on the two auxiliary variables y^L_m and y^R_n representing the two clusters c^L_m and c^R_n:

$$\psi(c^L_m, c^R_n) = w_c(c^L_m, c^R_n) \cdot | y^L_m - y^R_n |$$
$$w_c(c^L_m, c^R_n) = \frac{M^{LR}_{mn}}{M^{corr}_m} \frac{M^{LR}_{mn}}{M^{corr}_n} \qquad (6)$$

where M^{LR}_{mn} is the number of pixels in c^L_m that can find reliable correspondence in c^R_n. This potential function tends to build a strong link in our graph for the segments where the majority of the containing pixels can be related to the same segment. $G_m = \frac{M^{LR}_{mn}}{M^{corr}_m}$ and $G_n = \frac{M^{LR}_{mn}}{M^{corr}_n}$ can be seen as the 'goodness' of the two segments in terms of the consistency of depth cue and color cue. Note that $G_m = G_n = w_c = 1$, as the segments c^R_n and c^L_m are the only corresponding region to each other; in this case, a *hard* correspondence between the two groups of pixels will be constructed. Otherwise, *soft* correspondence among multiple segments will be constructed. Also note that no valid link will be constructed between two groups as $M^{LR}_{mn} = 0$. In other words, two groups of pixels should not affect each other if no group member corresponds to the other group via the depth map. Figure (2a) illustrates how label dependency is built between one group in the left view and three corresponding groups in the right. Figure (2b) is a matrix showing the link weight w_c between the first 407 segments in the left view and the first 420 segments in the right; the brightness indicates weight. The sparsity of this matrix implies that nonzero linking usually directs to quite a few target segments.

For another design, we consider a rounded weight as Equation (7):



(a)

(b)

Fig. 2. An example of our soft correspondence among multiple groups of pixels. In (a), the pixels marked as red in the left view are linked with the pixels marked as red and pink in the right view. The weighting factors of these cross-view links are visualized as a matrix like (b), where the brightness indicates the weight.

$$w_c(c^L_m, c^R_n) = \begin{cases} 1, \text{ if } G_m \text{ or } G_n > 0.5 \\ G_m \cdot G_n, \quad \text{otherwise} \end{cases} . \qquad (7)$$

This design equally weights every segment which can find one major correspondence in the other view, regardless of the goodness of this corresponding segment. Unlike Eq. (6), which implicitly penalizes one-to-many mapping, Eq. (7) may help in cases where one large segment is partitioned into several small segments in the other view by mean-shift.

Finally, by combining Eq. (5) and Eq. (6) (or Eq. (7)), we define our high-order consistency term as follows:

$$\sum_{m,n \in C} \psi^C_{mn} = \sum_{all\ m,n} (\psi(y_m, x) + \psi(y_n, x)) + \sum_{m,n \in C} \psi(c^L_m, c^R_n) \qquad (8)$$

We consider the proposed group-to-group correspondence more informative than conventional depth maps in that a higher coverage of pixels can be properly associated to the other view by intra-cluster consensus. Figures (3b) and (3c) show the typical results of the reliably associated regions via depth maps and the proposed group correspondence, respectively. As the example in Fig. (3b) demonstrates, a conventional depth map can build reliable pixel-to-pixel correspondence for 64% of the pixels, which we mark as white. In contrast, as shown in Fig. (3c), our region-to-region correspondence can construct strong correspondence for 93% of the pixels, thus providing better consistency in the stereo segmentation.

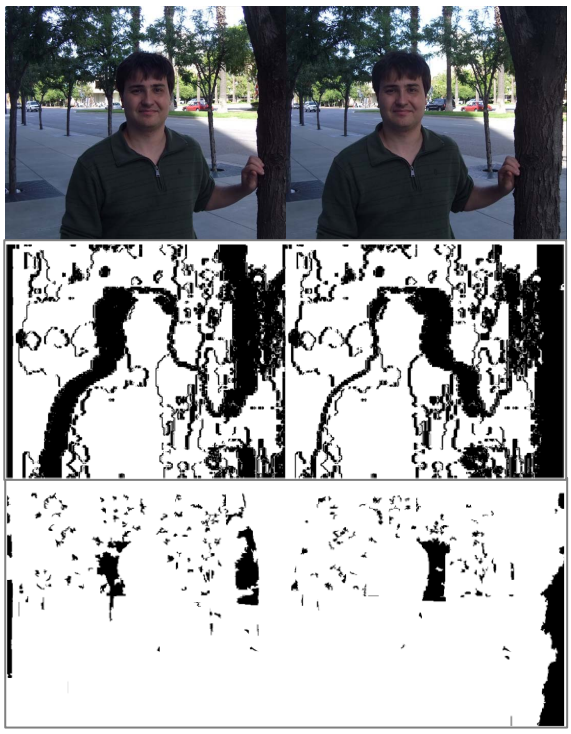


Fig. 3. Effectiveness of the augmented correspondence. (a) The input image pair. (b) About 64% pixels (marked as white) can find reliable correspondence via a depth map. (c) With color-based grouping, 93% pixels can find correspondence in the other view.

4. EXPERIMENTS AND DISCUSSION

We evaluate our stereo random field model on Adobe's stereo object selection dataset. Table 1 shows the results compared to two propagation-based and two joint segmentation methods. The five algorithms are evaluated in three test scenarios where different amounts of user input are provided to partially locate the foreground object and background scene in the left view. According to these user marked pixels, we train Gaussian mixture models in RGB color space for the foreground and background to construct the unary energy term in our random field. Depth maps in our experiments are generated by openvis3d[1]. We empirically determine parameters of our model and use fixed parameters in each test scenario. One can see in Table 1 that the proposed model outperforms prior work in all three scenarios, especially when the provided training data is scarce (in the Boundary and Stroke cases). In the proposed stereo random field, the two methods designed to define the cross-view linking work equally well.

To qualitatively compare our method to the state-of-the-art, we implement "Cons. Delta" proposed in [3]. Sample results shown in Fig. 4 and Fig. 5 demonstrate stereo segmentation produced by our and Price's joint segmentation methods. While our method successively provides better consistency in demarcation of hair (Fig. 4) and a banana (Fig. 5), as background color is indistinguishable from the foreground, two failure cases still can be identified in Fig. 5. First, the ducks in the two views are both classified as foreground. This mistake cannot be resolved in our model since two wrongly segmented regions related in the two views still reach an erroneous consensus. Also, a corner of the yellow box is misclassified as foreground in the left view due to color ambiguity. Although the box is correctly segmented in the right view, our cross-view consistency did not recover the segmentation error in the left view. This is because within the small corner region no valid pixel correspondence can be found via the provided depth map. This produces a "lone segment" in our graph; the pixels in this local cluster still cannot create the cross-view correspondence in our model. In our experiments, there are about 1.7% pixels located in such lone segments that may harm the resultant segmentation. To overcome this deficiency, an integration of overlapped clusters, which can be generated by applying different spatial bandwidth in mean-shift, may be needed. Though compared to directly segmenting with a depth map, which contains 32.7% lone pixels in the same experiments, our group correspondence still provides better consistency and accuracy.

Algo.\ Misclassfication rate	Truth (%)	Boundary (%)	Stroke (%)
SnapCut [1]	0.37	N/A	1.39
LiveCut [2]	1.07	2.87	1.17
Cons. Delta [3*]	0.28	1.33	1.11
Cons. PDF [3]	0.24	1.67	1.02
Proposed (Eq.6)	**0.22**	0.80	**0.64**
Proposed (Eq.7)	**0.22**	**0.78**	0.66

Table 1 (adapted from [3]): A comparison of five stereo segmentation algorithms on Adobe's stereo object selection dataset. [1] and [2] propagate segmentation results in one view to the other to help the consequent segmentation. Our method and the two best versions in [3] segment the two views in one shot. The proposed method achieves the lowest error rate among the five methods in all of the three test scenarios. *The misclassification rate of Cons. Delta is slightly different from that reported in [3] because we run our own implementation for further qualitative comparison.

5. CONCLUSIONS

In this paper, we have presented a novel random field that softly encourages group-level, cross-view consistency while performing pixel-level segmentation. Compared with prior work, the proposed cross-view association more effectively models the two-view dependency in a single graph and leads

[1] Openvis3d: http://code.google.com/p/openvis3d/

Fig. 4. A successful case. *Left top*: Input stereo pair. *Left bottom*: ground truth segmentation. *Right top*: Result of Consistent Delta proposed by [3]. *Right bottom*: our result. The hair of the person is better extracted in our result.

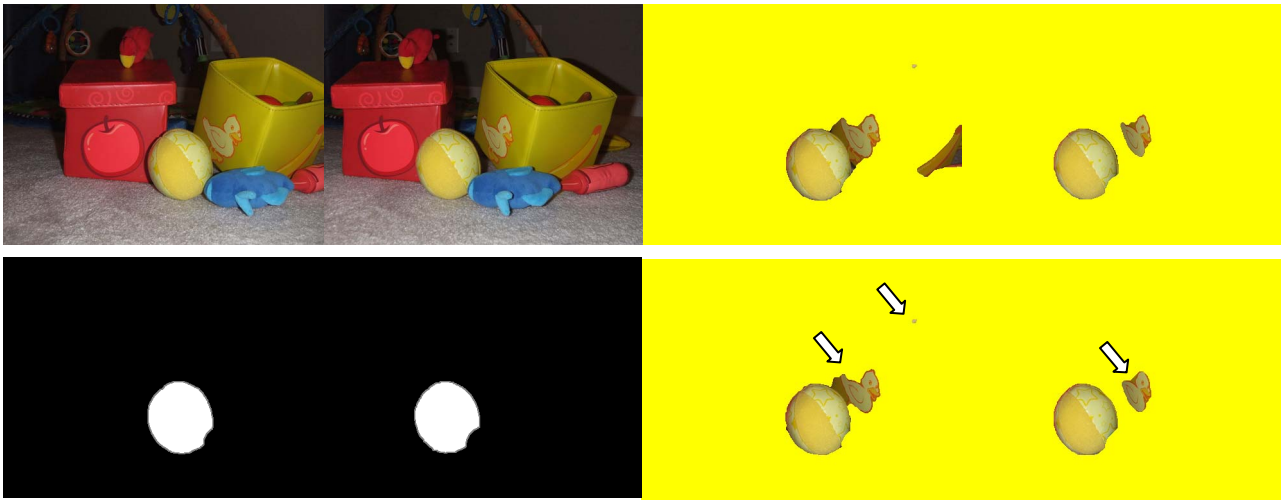


Fig. 5. A failure case; same layout as Fig. 4. The duck and one corner of the yellow box are misclassified as foreground even though our model provides better two-view consistency. See text for the discussion.

to better consistency of bi-layer stereo segmentation. We expect this technique can improve the visual quality of a bi-layer disparity remapping technique for 3D video communications, as described in [15].

6. REFERENCES

[1] X. Bai, J. Wang, D. Simons, and G. Sapiro. "Video snapcut: robust video object cutout using localized classifiers," *ACM SIGGRAPH*, 2009.

[2] B. Price, B. Morse, and S. Cohen. "Livecut: Learning based interactive video segmentation by evaluation of multiple propagated cues." *IEEE ICCV*, 2009.

[3] B. Price and S. Cohen, "StereoCut: Consistent Interactive Object Selection in Stereo Image Pairs," *IEEE ICCV*, 2011.

[4] I2I- http://research.microsoft.com/en-us/projects/i2i/

[5] W.-Y. Lo, J. van Baar, C. Knaus, M. Zwucker, and M. Gross. "Stereoscopic 3d copy & paste," *SIGGRAPH*, 2010.

[6] H. Lim et al., "Bi-layer inpainting for novel view synthesis," *IEEE ICIP*, 2011.

[7] P. Kohli, L. Ladický, and P. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 2009.

[8] L. Ladický, C. Russell, P. Kohli, and P. Torr. "Associative hierarchical crfs for object class image segmentation," *IEEE ICCV*, 2009.

[9] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," *IEEE ICCV*, 2007.

[10] M. Grundmann, V. Kwatra, M. Han, and I. Essa. "Efficient hierarchical graph-based video segmentation," *CVPR*, 2010.

[11] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *ICML*, 2001.

[12] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE TPAMI*, 2002.

[13] Y. Boykov, O. Veksler, and R. Zabih "Fast Approximate Energy Minimization via Graph Cuts," *IEEE TPAMI*, 2001.

[14] S. Gould "Multiclass Pixel Labeling with Non-Local Matching Constraints," *IEEE CVPR*, 2012.

[15] S. Mangiat, K.-C. Lien, and J. D. Gibson "Bi-layer Disparity remapping for Handheld 3D Video Communications," *IEEE ICME*, 2013.