INTERACTIVE VIDEO SEGMENTATION SUPPORTED BY MULTIPLE MODALITIES, WITH AN APPLICATION TO DEPTH MAPS

Jeroen van Baar[1], *Paul Beardsley*[1]

[1]Disney Research Zürich
Clausiusstrasse 49
CH-8092, Zürich

Marc Pollefeys[2], *Markus Gross*[1,2]

[2]Swiss Federal Institute of Technology (ETH)
Institute for Visual Computing
Zürich

ABSTRACT

In this paper we propose an interactive method for the segmentation of objects in video. We aim to exploit multiple modalities to reduce the dependency on color discrimination alone. Given an initial segmentation for the first and last frame of a video sequence, we aim to propagate the segmentation to the intermediate frames of the sequence. Video frames are first segmented into superpixels. The segmentation propagation is then regarded as a superpixels labeling problem. The problem is formulated as an energy minimization problem which can be solved efficiently. Higher-order energy terms are included to represent temporal constraints. Our proposed method is interactive, to ensure correct propagation and relabel incorrectly labeled superpixels. As a final step the initial segmentation boundaries are refined to obtain accurate object boundaries. We then exploit these object boundaries in an application for computing depth maps.

***Index Terms* —** Image segmentation, Video signal processing, Optimization, User-generated content

1. INTRODUCTION

Segmentation is a fundamental operation in image and video processing. Different post-processing applications utilize segmented foreground objects for example for compositing or tracking. In this paper we focus on segmentation of objects in video sequences. Segmentation can be regarded as a labeling problem: given a set of labels representing the foreground objects and background, which label is assigned to each pixel? This labeling is determined from the color similarities between pixels, both within a video frame as well as between video frames. Although many methods have been proposed in the literature, segmentation remains a challenging problem for many scenes.

To reduce the dependency on color discrimination alone, we have proposed a system to capture additional modalities besides color images. These modalities include Time-of-Flight (ToF) depth and far infrared (thermal) images. Figure 1 shows our prototype system. In this paper we propose a flexible interactive segmentation method which exploits the multiple modalities. Given an initial start and end segmentation, our proposed method propagates the segmentations across the intermediate video frames. We formulate the problem as a labeling problem of smaller segments, or superpixels, across the video sequence. Superpixels are matched to superpixels in adjacent images, without requiring the computation of optical flow, or camera motion. We will show that the labeling problem can be solved efficiently, while exploiting temporal coherence. Initial boundaries of the segmented objects are then further refined to obtain accurate object boundaries.

Object boundaries are usually strongly correlated with depth discontinuities in the scene. We can thus exploit these accurate boundaries in computing depth maps. An advantage of accurate object boundaries for a video sequence is temporal consistency for the depth discontinuities. Results of this application will be presented in Section 4.

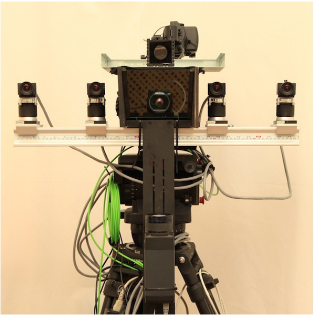

Figure 1. The high quality reference camera is at center (obscured by a beam splitter). The four satellite cameras are arranged approximately along a horizontal line with the reference camera (two on either side). The depth camera is at the top. The beam splitter passes visible light to the reference camera and reflects thermal radiation to the thermal camera at the bottom.

2. RELATED WORK

Video segmentation is a well-studied research area. It is beyond the scope of this paper to provide an exhaustive list of related work. Interactive segmentation methods [1, 2, 3, 4, 5] require the user to provide initial scribbles to indicate fore- and background. Local color models are learned from the initial indication, and pixels are assigned a label according to these models. These methods are limited to segmentation of *single* foreground objects only. Our approach is also interactive: the user provides an initial segmentation for the first and last frame of the sequence, and supervises the propagation to ensure correct labeling. However we define the segmentation as a labeling problem over a video sequence, and segment multiple foreground objects with accurate segment boundaries.

Automatic methods [6, 7, 8, 9, 10] can obtain temporally consistent segmentations, however the segmentation boundaries are inaccurate. The methods proposed by [8, 9] formulate video segmentation as an energy minimization problem. For these methods,

```
Propagation Algorithm
- Given a video sequence $\mathbf{I} = \{I_1, \cdots, I_n\}$.
- Perform superpixel segmentation on frames of $\mathbf{I}$.
- Known segmentations for $I_1, I_n$.
- For each intermediate frame $I_i$:
    - For each superpixel $S_i^k$:
        - Determine label from $I_1, I_n$.
        - Determine matches $S_{i-1}^k$ and $S_{i+1}^k$.
- Define matched sequences $\mathbf{S} = \{S_2^k, \cdots, S_{n-1}^k\}$.
- Formulate as energy minimization problem.
```

Table 1. Outline of the algorithm for propagating known segmentations.

occlusions result into separate segmentations. We also define the problem as an energy minimization problem to propagate the intermediate frames of the video sequence. However, by segmenting each video frame into a set of superpixels and exploiting known segmentations for the first and last frame of a video sequence, we can handle occluding foreground objects.

Segmentation and stereo are correlated and this is exploited to simultaneously compute segmentation and depth maps [11, 12]. These methods operate on stereo image pairs, and the resulting segmentation is used to handle stereo occlusions, as opposed to object occlusions. Background segmentation is exploited in [13] for improving depth maps. In contrast, we propose an interactive approach to accurately segment multiple (possibly occluding) objects in a video sequence. These accurate boundaries may then be exploited as explicit constraints when computing depth maps.

3. VIDEO SEGMENTATION USING MULTIPLE MODALITIES

The outline of our algorithm is given in Table 1. For a video sequence $\mathbf{I}$ consisting of n frames I_i, we first perform superpixel segmentation [14] on each I_i. In the superpixel segmentation, we exploit both color and the thermal signal to obtain superpixel boundaries. Next, known segmentations for frames I_1 and I_n are provided. With known segmentations for both I_1 and I_n we are able to handle occlusions between foreground objects. In our work the known segmentations are obtained by interactive merging of superpixels. The goal is now to propagate these known segmentations over the intermediate frames in the video sequence. We will discuss this in more detail below.

3.1. Segmentation Propagation as Energy Minimization

We formulate the problem of propagating known segmentations for I_1 and I_n as an energy minimization:

$$E = \phi(x_i) + \phi(x_i, x_j) + \phi(\mathbf{x}). \tag{1}$$

Here $\phi(x)$ represents a unary term, $\phi(x_i, x_j)$ represents a binary term between neighboring superpixels x_i and x_j, and finally $\phi(\mathbf{x})$ represents a so-called higher-order clique term [15]. Each superpixel is assigned a label, with the set of labels $\mathcal{L}$ defined by the different segments.

The unary term $\phi(x)$ represents the likelihood of a superpixel taking a particular label. It is defined by the smallest matching cost between a superpixel S_i in I_i, and matching superpixels in I_1 and I_n. The matching cost for S_i is computed for each label. The matching cost between superpixels is the Euclidean distance between feature vectors defined as:

$$f = (\bar{Y}, \bar{C_b}, \bar{C_r}, \bar{Th}, H_Y, H_{Cb}, H_{Cr}, H_{Th})^T. \tag{2}$$

Here Y, C_b, C_r represents the superpixels' color, Th represents the thermal signal, and the H-terms are histograms. The histograms are determined for the superpixel S and its set of neighboring superpixels $S^{\mathcal{N}}$. We take the neighborhood into account since a local neighborhood around a superpixel only changes near object boundaries, but otherwise remains constant. This greatly improves matching robustness. The cost of matching any two superpixels S^i and S^j then becomes:

$$MC_{i \to j} = \|(\triangle Y, \triangle C_b, \triangle C_r, \triangle Th, \chi^2 dist(H))^T\|. \tag{3}$$

Note that we do not enforce uniqueness: superpixels in I_1 or I_n can be matched to multiple superpixels in I_i.

Candidate superpixels are determined by defining a search radius r around the centroid location of S_i in I_1, I_n. All superpixels within radius r are then considered as candidates. Since we do not require flow or camera motion, this matching approach can handle non-rigid motions and moving cameras, at the expense of increased computational complexity. The radius r depends on the motion in the scene. Large motions will require a correspondingly larger search radius.

The binary term $\phi(x_i, x_j)$ in 1 aims to enforce a first-order smoothness prior between neighboring superpixels in a frame, under the assumption that similar color (and thermal signal) superpixels should likely have the same label. Using Graph Cuts [16] we could solve 1 taking only unary and binary terms into account. This would result in a per-frame segmentation, without any temporal consistency between corresponding superpixels across frames.

Using the higher-order term, or clique potential, $\phi(\mathbf{x})$ we aim to impose a temporal smoothness constraint on the superpixel labeling. Clique potentials penalize the assignment of different labels to some collection of variables, i.e. a clique. Kohli et al. [15] define a robust extension to allow some members of the clique to take a different label. The robust clique potential is defined as:

$$\phi(\mathbf{x}) = \min\{\min_{k \in \mathcal{L}} \left(N \cdot \frac{\gamma_{max} - \gamma_k}{Q} + \gamma_k \right), \gamma_{max}\}. \tag{4}$$

Here γ_k is a per-label penalty, γ_{max} is the maximum penalty for the clique, and $N = |c_{\mathbf{x}}| - n_k(\mathbf{x})$, that is the number of variables in the clique which take a different label than l_k. Q is a truncation parameter reflecting how many variables are expected to have a different label.

Cliques are defined by formulating *sequences* of matching superpixel correspondences over the video sequence. Using the matching approach described above, a superpixel S_i in I_i is matched S_i to a superpixel in I_{i-1} resulting in S_{i-1} (and vice versa). The match sequence is then formulated as $\mathbf{S} = \{S_2, \cdots, S_{n-1}\}$, given matching superpixels $S_2, S_3, \cdots, S_{n-1}$ for frames $[2, n]$. For each sequence $\mathbf{S}_j$ we store two sequences of matching costs: C_j^S stores the matching cost of superpixels between adjacent frames, and $C_j^{\mathcal{L}}$ stores the matching cost of superpixels with the labels (from the first and last image in the sequence). We can then determine the mean $\mu_k(C_j^{L_k})$ for each L_k in $\mathcal{L}$, and $\gamma_{k_{best}} = \min_k(\mu_k)$. The remaining γ_k are assigned γ_{max}, which is determined as: $\gamma_{max} = \gamma_{k_{best}} + \varepsilon$. The cost increase ε in turn is determined from the standard deviation of C_j^S. This ensures that some variables in the clique may be assigned a different label with only moderate cost increase. Given $C_j^{\mathcal{L}}$ we can determine the expected number of variables with a different label, denoted N_e. The truncation parameter is then set to $Q = \min\left(\frac{|\mathbf{x}|}{2}, N_e\right)$. However, in the case when $\mu\left(C^{L_j^k}\right)$ have similar value, we set $Q = \frac{|\mathbf{x}|}{2}$.

In our case, some superpixels in a match sequence should be allowed to have a different label. Either because the matching was incorrect or because of an occlusion occurrence. A sequence $\mathbf{S}_k$ may also be split into $\mathbf{S}_k^l, \mathbf{S}_k^r$ if a cost c_i in C_S^k is above some threshold. This typically occurs for occlusions and wrong matches. Each subsequence is then treated individually.

In the case where the propagation on a sequence fails, we could perform the propagation iteratively. After the initial segmentation propagation is performed, the video sequence is broken into smaller sequences. The initial propagated segmentation may then serve as a starting point. The user would then complete these initial segmentations and the propagation is applied to the smaller sequences.

3.2. Interactive Segmentation Correction

Superpixels may have an incorrect label after propagation. It is therefore necessary for the user to correct these incorrect segmentation labels. Rather than requiring to re-label individual pixels, in our case the interactive correction step can be more easily performed on the superpixels directly.

3.3. Segmentation Boundary Refinement

The segmentation boundaries after propagation and interactive correction are not yet accurate, and may include pixels from other foreground objects or the background. We employ a boundary refinement step based on the method of Bai et al. [2]. We similarly define local overlapping classifier windows along the boundaries of each segmented foreground object. We extend the method from [2] by considering multiple segmentations within each window (representing the different foreground objects). For each segmentation we compute a Gaussian Mixture Model (GMM) within the window. In our case we also include the thermal signal when computing the GMM. Each GMM represents the likelihood of the pixels belonging to a particular segmentation. We then refine the pixels near the initial boundary using Graph Cuts. We repeat this process for two to three iterations.

3.4. Exploiting Multiple Modalities

We exploit the thermal signal in the superpixel segmentation and in the matching of superpixels. This is especially helpful for scenes with human actors, since the thermal signal helps to separate the actors from their background, and could also help to separate actors from each other. The thermal signal is also exploited in the boundary refinement as explained in Section 3.3. We can also exploit the depth we obtain from the ToF camera, although the ToF depth is not reliable enough to perform accurate matching between superpixels. Instead we use the ToF depth to merge superpixels if their depths are within some threshold. The advantage of this merging is that we greatly reduce the number of individual superpixels to process.

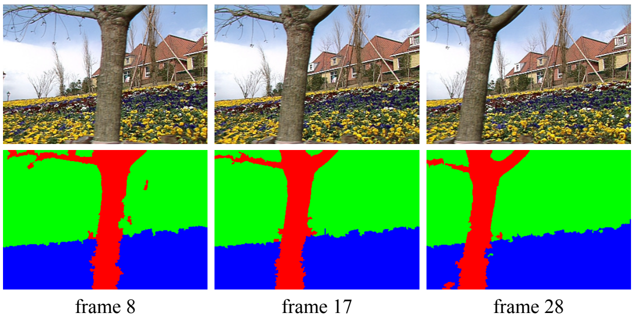

Figure 2. Result of our propagation method for frames 8, 17, and 28 of the flower garden dataset. The video sequence consists of 40 frames. The first and last image of the sequence have been segmented into three layers: tree, flowers, background. The results shown here are prior to interactive correction by the user, and shows the performance of our method on a standard dataset.

4. RESULTS

Figure 2 shows the result of our propagation method for frames 8, 17, and 28 of the standard flower garden video sequence. We used 40 frames in this example, and a three level segmentation was provided for the first and last frame of this video sequence. The propagated segmentations are prior to any interactive correction by the user, and prior to boundary refinement. Although some of the superpixels have been mislabeled, these results demonstrate that we can achieve good segmentation propagation for an arbitrary video sequence.

Figure 3 shows the result of a challenging case where one person occludes another as they walk past. This dataset was acquired with our prototype rig of Figure 1. Figure 3 shows the results just before the occlusion (top row), during the occlusion (middle row), and just after the occlusion occurred (bottom row). The segmentation propagation results are before interactive correction and boundary refinement. By exploiting the thermal signal, and defining clique potentials, the propagation can keep track of both people even though one person is nearly entirely occluded by the other. In particular, frames near the occlusion occurrence require interactive correction of the labels for a small number of superpixels, however this is crucial for accurate results.

The top row of Figure 4 shows the result of boundary refinement for the segmentation in the top-right of Figure 3. By exploiting the thermal signal (see Figure 4 inset), the refinement can also produce a good boundary in the region where the hair overlaps, and there is no color discrimination.

4.1. Application: Depth Maps

The bottom row of Figure 4 shows the result of a depth map computed from the multi-modal sensor information. The propagated and refined segmentation boundaries are used as constraints. We formulate the fusion of the ToF depth and the satellite cameras (Figure 1) using an energy function. A more detailed discussion is beyond the scope of this paper. We minimize the energy formulation using Belief Propagation. The segment boundaries produce accurate depth discontinuities for this challenging case.

5. CONCLUSION

We have described an interactive video segmentation approach based on the propagation of known segmentations for the first and

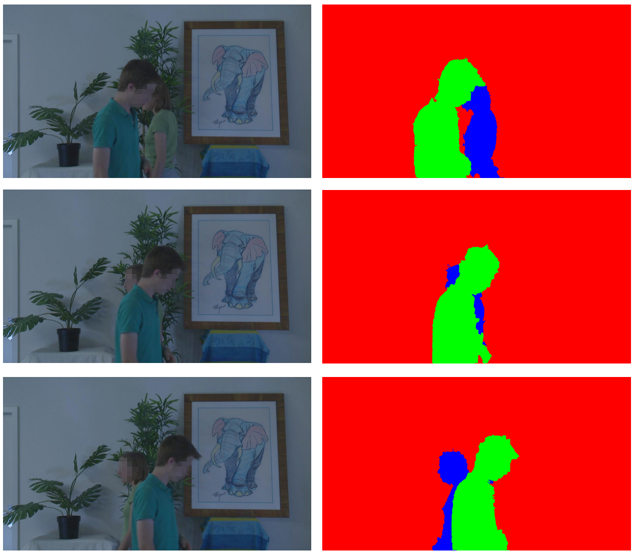

Figure 3. Segmentation propagation for occluding objects. **Left** The input images. **Right** Segmentation propagation results prior to interactive correction and boundary refinement. Our method is able to propagate the segmentation through an occlusion.

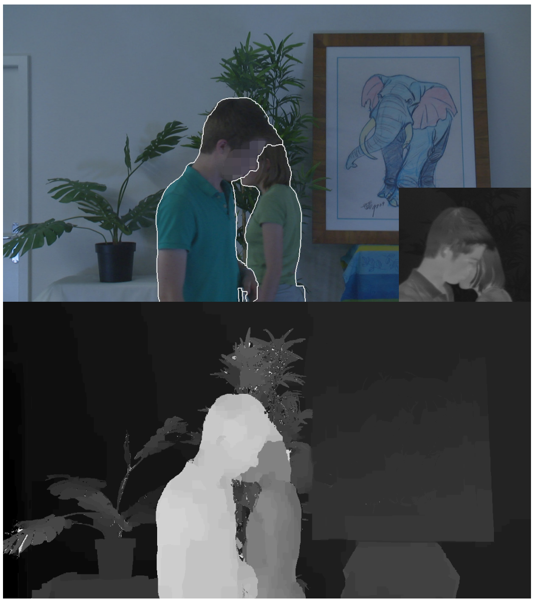

Figure 4. **Top row** The refined boundaries are superimposed on the input image. The inset shows a crop of the thermal image. In combination with the thermal signal, boundary refinement can produce good results. **Bottom row** Depth map corresponding to the image above, with the refined boundaries taken as constraints.

last frame, to the intermediate frames of a video sequence. The straightforward matching of superpixels across the video sequence could easily handle moving cameras and non-rigidly moving objects. The foreground objects within a sequence must be present in the first and last frame. However, foreground objects may then disappear and re-appear for the intermediate frames, which for example is the case in occlusions. Exploiting multiple modalities helps to make the matching of superpixels between frames of a sequence more robust. A user interactively corrects the propagated labeling. Finally, a refinement step produces accurate boundaries which can be used during the computation of depth maps.

6. REFERENCES

[1] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, “”grabcut”: interactive foreground extraction using iterated graph cuts,” *ACM Trans. Graph.*, 2004.

[2] X Bai, J Wang, D Simons, and G Sapiro, “Video snapcut: robust video object cutout using localized classifiers,” *ACM Trans. Graph.*, 2009.

[3] Minglun Gong and Li Cheng, “Foreground segmentation of live videos using locally competing 1svms,” *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2011.

[4] Brian L. Price, Bryan S. Morse, and Scott Cohen, “Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues,” in *IEEE Computer Vision*, 2009.

[5] P Ochs and T Brox, “Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions,” *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[6] Daniel DeMenthon and Remi Megret, “Spatio-Temporal Segmentation of Video by Hierarchical Mean Shift Analysis,” Tech. Rep. LAMP-TR-090,CAR-TR-978,CS-TR-4388,UMIACS-TR-2002-68, University of Maryland, College Park, 2002.

[7] Sylvain Paris, “Edge-preserving smoothing and mean-shift segmentation of video streams,” in *European Conference on Computer Vision (ECCV)*. 2008, Springer-Verlag.

[8] M Grundmann, V Kwatra, Mei Han, and I Essa, “Efficient hierarchical graph-based video segmentation,” *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2010.

[9] A Vazquez-Reina, S Avidan, H Pfister, and E Miller, “Multiple hypothesis video segmentation from superpixel flows,” *European Conference on Computer Vision (ECCV)*, 2010.

[10] J. Lezama, K. Alahari, J. Sivic, and I. Laptev, “Track to the future: Spatio-temporal video segmentation with long-range motion cues,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[11] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha, “Object stereo – joint stereo matching and object segmentation,” *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2011.

[12] Guofeng Zhang, Jiaya Jia, and Hujun Bao, “Simultaneous multi-body stereo and segmentation,” in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[13] Wenzhuo Yang, Guofeng Zhang, Hujun Bao, Jiwon Kim, and Ho Young Lee, “Consistent depth maps recovery from a trinocular video sequence,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[14] R Achanta, A Shaji, K Smith, A Lucchi, P Fua, and S Süsstrunk, “Slic superpixels,” *Technical Report 149300 EPFL*, 2010.

[15] P Kohli, L Ladický, and P.H.S Torr, “Robust higher order potentials for enforcing label consistency,” *Int. Journal Computer Vis.*, 2009.

[16] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” *IEEE Pattern Analysis and Machine Intelligence (PAMI)*, 2001.