USER ASSISTED STEREO IMAGE SEGMENTATION

H. Emrah Tasli and A. Aydin Alatan

Department of Electrical and Electronics Eng, Middle East Technical University, Ankara, Turkey

ABSTRACT

The wide availability of stereoscopic 3D displays created a considerable market for content producers. This encouraged researchers to focus on methods to alter and process the content for various purposes. This study concentrates on user assisted image segmentation and proposes a method to extend previous techniques on monoscopic image segmentation to stereoscopic footage with minimum effort. User assistance is required to indicate the representative locations of an image as object and background regions. An MRF based energy minimization technique is utilized where user inputs are applied only on one of the stereoscopic pairs. A key contribution of the proposed study is the elimination of dense disparity estimation by introducing a sparse feature matching idea. Segmentation results are evaluated by objective metrics on a ground truth stereo segmentation dataset and it can be concluded that competitive results with minimum user interaction have been obtained even without dense disparity estimation.

1. INTRODUCTION

Segmentation, dating back to a couple of decades, has always been a popular field due to the range of applications and performance issues. Many automatic and assisted methods have been proposed, however, no solution has been able to achieve human-like performance yet. Therefore, human assistance is exploited in order to achieve superior performance under varying cases. The seminal work of Boykov [1] has been vastly studied and considerable performance increase has been achieved since then. Rother et al. [2] demonstrates that with minimum user assistance objects can be rapidly segmented .

With the recent increase in 3D capable displays, an interest in 3D content has arisen. This brought the necessity towards 3D content creation and editing tools. This paper presents a method to extend previous techniques on monoscopic image segmentation towards the stereoscopic case with minimum extra cost up. Previous studies have offered methods to incorporate disparity information in stereo segmentation to overcome possible perception, disparity, and occlusion issues. However, the methods proposed for estimating per pixel disparities in two view stereo sometimes lack "ground truth" accuracy for arbitrary scenes. The steroscopic copy-paste idea [3] concentrates on the same problem and offers a method to segment the selected object in stereo image by interactively merging the oversegmented regions. The regions are clustered according to a maximal-similarity merging method, and then refined by graph cut. The propagation of left eye segment on its right eye pair is realized through the disparity information corresponding to the segmented object. A recent study [4] provides details of the joint energy assignment in graph cut method for the stereo image pairs. However, the main constraint on previous methods is the necessity of dense disparity estimation which is a computationally complex step in the whole pipeline.

Since there is only limited literature about stereo image segmentation, mono segmentation techniques are also investigated and utilized for qualitative comparison. The study in [5] proposes an interactive video segmentation method where structure from motion techniques are utilized for information propagation through the succeeding frames. However, quite long processing and interaction times cause this approach to be highly impractical. The study proposed in [6] utilizes many different cues for obtaining segmentation. Color, gradient, color adjacency, shape, temporal coherence, camera and object motion and easily-trackable points are the cues incorporated in the graph-cut optimization framework. The weighting of the cues are achieved automatically in order to boost performance using the most effective cues for segmentation. However, it also requires long execution and interaction time for the final result. In order to reduce the interaction and execution time, we propose an efficient method which requires interaction only with one of the stereo pairs and removes computational burden of dense disparity estimation.

The highlights of the proposed method can be listed as follows: The interactive segmentation framework utilizes MRF based energy minimization. Superpixel primitives are used [7] in the graph generation phase for efficient maximum flow calculation. User assistance is required as input seeds on the representative locations of just one of the stereo image pairs to save user from repeating the procedure for the second image. The information propagation is handled via efficient feature point based stereo matching. Hence, the necessity of a dense disparity estimation module is eliminated. The ground truth stereo database [4] is tested for judging objective stereo segmentation performance. With additional user strokes, the proposed method is shown to generate outstanding results compared to state of the art methods.

2. ALGORITHM DETAILS

The goal of the method is to faithfully segment object and background regions in stereoscopic image pairs. Algorithmic flow is presented in four major steps:

- Estimation of superpixel regions [8] for graph generation.
- User assistance as scribbles on image representative areas.
- Feature matching for information propagation.
- Stereo segmentation via graph cut.

2.1. Graph Initialization and Energy Assignment

Original work [1] introducing energy minimization approach for interactive image segmentation purposes utilizes pixel primitives as graph nodes. This assignment restricts the applicability of the method to higher resolution images due to memory and time constraints. Hence, we have incorporated superpixel idea in the graph generation phase to overcome the mentioned difficulties.

Oversegment patches divide image into color and textural wise homogeneous small regions. Each small region corresponds to the

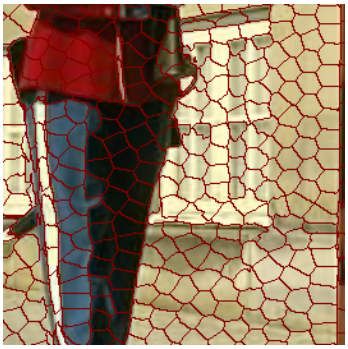

Figure 1. Oversegment Boundary Adaptation

nodes of an undirected graph $G = (V, E)$. Each edge e of the graph is assigned a weight w_e depending on the similarity of the nodes that it connects. The source S and sink T are the terminal nodes and they are connected to each node with weights assigned in accordance with the user input and segment similarity to the object and background regions. Oversegment boundaries as shown in Figure 1 well adapt to the local image edges, and hence, the assumption of assigning same label to the pixels in same superpixel region becomes valid.

2.1.1. Energy Cost for Interactive Segmentation

The general energy cost function to minimize in the binary segmentation problem is given below.

$$E(L) = \sum_{p \in P} D_p(L_p) + \sum_{(p,q) \in N_p} V_{(p,q)}(L_p, L_q) \qquad (1)$$

L is the labelling of image P, D_p is the data penalty, $V_{p,q}$ is the interaction potential and N_p is the neighborhood of node p. Data penalty, as the name indicates, is designed to increase as the probability of assigning a label to the selected node decreases. Interaction potential functions as a smoothness prior, it obtains high values, when two similar intensity neighbors are assigned different labels.

Edge weights between the nodes and between nodes and terminals are assigned according to the energy formula. D_p is the weight of the edge from nodes to terminals. It is either set by the user via input strokes or determined by the system otherwise. Inputs scribbles for object and background regions are used to determine a limited model of the region to be segmented. The proposed similarity metric uses this limited model to measure similarity of a node to the object and background. $D_p(B)$ is the edge weight from node to "Source" (Background) and $D_p(O)$ is the edge weight from node to "Sink" (Object). Similarity measure is related to the input node statistics and proposed location distance of the current patch to the input nodes.

Edge weight between node p and background terminal is equal to:

$$D_p(B) = max\{similarity(p, q), q \in B\} \qquad (2)$$

The weight of edge between node p and object terminal is given as:

$$D_p(O) = max\{similarity(p, q), q \in O\} \qquad (3)$$

The weight of edge between node p and q where $q \in N_p$ is defined as:

$$V_{(p,q)}(L_p, L_q) = similarity(p, q), q \in N_P \qquad (4)$$

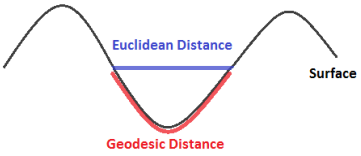

Figure 2. Geodesic Distance Illustration

The similarity between two nodes of any kind is found as:

$$similarity(p, q) = e^{-\lambda_1 * intDiff/locDist} \qquad (5)$$

Intensity difference, $intDiff$ in the above formulation is defined as the similarity of the two oversegment nodes and can be computed by mean, median or sum of intensity differences in a selected color domain. Histogram-based comparison is also powerful, but computationally complex. Location difference, $locDist$ between two nodes is calculated using a geodesic type distance formulation as explained in the prior studies [9], [10].

2.1.2. Energy Cost using Geodesic Distance

Gaussian mixture-based region modelling utilize intensity difference for defining similarity (5) between nodes. However, we utilize a distance term in the similarity equation in addition to the intensity difference term (5). Main motivation for utilization of such a distance information is to differentiate two different regions with similar color distribution. When only color similarity is considered, a superpixel patch outside the object region might gain high similarity value to the object region. However, by the location information, it can be correctly assigned to background. During the selection of distance metric, Euclidean is considered first. However, it is observed that Euclidean-wise closeness usually does not reveal useful information. The distance of a node to a given scribble might be close especially at the object boundaries. However, it does not necessarily imply that this region is similar to the region defined by the closest scribble point. Hence, geodesic type distance is utilized for resolving such ambiguities. As Figure 2 illustrates, the geodesic path needed to be traversed in order to reach the target point might be different than the Euclidean path. In the case of graph node similarity assignment, Figure 3 shows a typical case where object regions close to background scribbles might easily be assigned to background due to

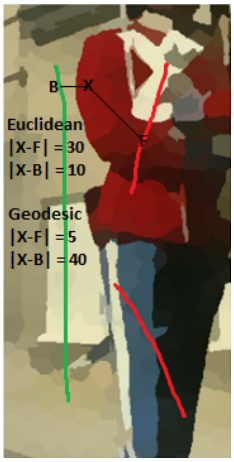

Figure 3. Euclidean vs Geodesic Distance: X is closer to B (F) in Euclid (Geod) distance

Figure 4. ORB feature point match

Euclidean-wise closeness. In the case of geodesic distance, the cost function penalizes traversing through nodes with different intensity values. Hence, greater geodesic distances can be achieved for Euclidean-wise close regions as shown in Figure 3.

2.2. User Interaction for Background and Foreground

User interaction is necessary and sufficient for generating intended region segmentation. Object and background regions are individually selected via input scribbles on the representative locations of the image. Unlike many state-of-the-art methods, the proposed method expects user to set input seeds only on one of the image pairs. This provides an easier interaction by minimizing user assistance. The framework is also designed to post process the final segmentation when the user is not totally satisfied with the output. Figure 5 and 6 shows typical input scribbles on the image with red and green colors for object and background identification.

2.3. Stereo Matching for Transferring Scribbles

Previous methods on stereo segmentation mostly utilize a dense disparity map for estimating pixel correspondences. According to the rankings in Middlebury [11], dense disparity estimation at the original image resolution is a computationally complex process even with an efficient implementation [12]. The estimated results is also prone errors by the best performing method [13]. However, proposed method eliminates this procedure by applying an efficient sparse feature matching method. Stereo feature matches as shown in Figure 4 are used to find the disparity of the segmented object. The estimated object disparity is used to transfer the scribbles supplied from one image to the other.

During selection of the feature descriptor, state-of-the-art methods have been considered. It is observed that the popular methods such as SIFT [14] or SURF [15] rely on costly descriptors for detection and matching. Hence, ORB (Oriented FAST [16] and Rotated BRIEF [17]) feature detector [18] has been selected due to its computational efficiency and performance. It improves FAST by an additional fast and accurate orientation component. Keypoint detection is defined as a binary classification problem where pixels are labelled as keypoint or not. Center pixel is compared with the selected pixels on a 7x7 patch for deciding if the center is brighter, darker or similar to the selected pixel. Decision tree tests has been selected depending on a entropy based information gain maximization principle. Another major improvement in ORB is the efficient computation of oriented BRIEF features. BRIEF is a binary descriptor which utilizes binary tests on an image patch centered on the selected pixel. Similarity between descriptors is then measured by the Hamming distance between the corresponding binary strings . This is quite fast, since the Hamming distance can be computed very efficiently with a bitwise XOR operation followed by a bit count [17].

Table 1. Average Segmentation Error

Methods	Label Error in %
Livecut [6]	1.07
Snapcut [5]	0.37
StereoCut [4]	0.31
Proposed with Euclidean	0.39
Proposed with Geodesic	0.32

2.4. Stereo Segmentation

At the final step of the algorithm, the input seeds and their stereo correspondences are integrated in the graph cut energy formulation as explained in Section 2.1.1. The calculated average disparity in the segmented object is used for transferring seed location information on the stereo pair. The relocated input seeds are used in the binary segmentation framework to create final stereo segmentation output as shown in Figure 5.

Figure 5. Input Scribbles and Proposed Stereo Segmentation

3. EXPERIMENTS

Segmentation results have been evaluated by a ground truth dataset containing binary segmentation results, supplied by [4]. It contains 30 images while some of them are from Middlebury stereo dataset [11]. The images have ground truth labels as foreground and background regions. The segmentation performance is quantitatively evaluated in terms of segmentation accuracy, which measures the ratio of the correctly labelled pixels over the total number of pixels. Table 1 presents the average label error rates in comparison with other methods for all 30 images. Proposed method is tested using geodesic and Euclidean distance in node similarity assignment. Geodesic distance utilization produced lower error rates as expected. Based on the presented results, our technique performs superior to or competitive with the state-of-the-art.

Any interactive system can produce satisfying results with sufficient assistance. The virtue is to keep these interactions at

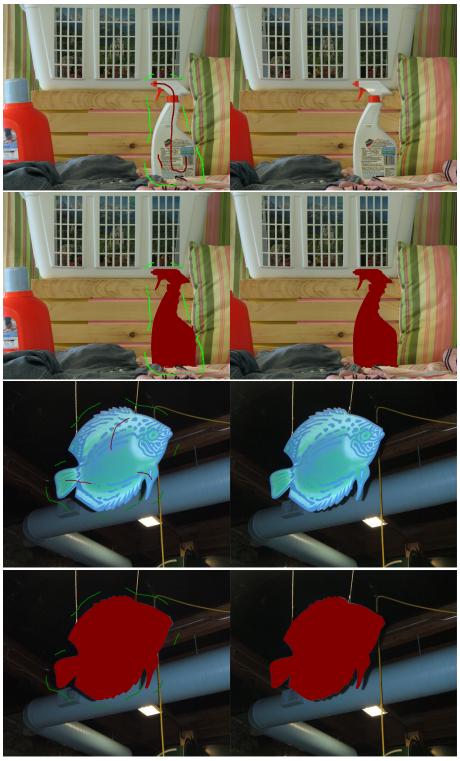

Figure 6. Input Scribbles and Proposed Stereo Segmentation

minimum. We firmly believe that the true performance of a user-assisted segmentation technique cannot be evaluated by pure segmentation results; number of interactions should also be considered. Hence, we recorded the required time for obtaining proposed segmentation results. On the average, our system requires less than one minute per image including user interaction and CPU processing time. Approximately 3 seconds for preprocessing (including sparse feature matching) and 50 ms for graph cut optimization is recorded for 1920x1080 resolution stereo images on a 3.06 GHz PC.

It should be noted that the method in [4] strictly requires a dense depth field to obtain a stereo segmentation. Time required for a dense disparity estimation takes up to minutes for the given resolution even with the most efficient methods [11]. Therefore, the proposed solution with sparse matches is computationally much feasible compared to such a dense approach.

4. CONCLUSION AND FUTURE WORK

This study proposes an efficient method for superpixel based stereo image segmentation. Graph cut optimization is utilized for solving the minimum cut problem. Since the graph nodes are generated from oversegment image patches where region homogeneity and convexity are prioritized, it provides an efficient solution in terms of computational complexity. Main contribution of the study is to provide an easier interaction through single image via propagating user input on the stereo pair efficiently. This approach eliminates the commonly used disparity estimation step from the pipeline. Quantitative results prove the functionality of the proposed method with low segmentation error.

As for future work, proposed method will be applied to video content where there is strong coherency between the succeeding frames.

5. REFERENCES

[1] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Computer Vision, 2001. ICCV 2001.*

[2] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, ""grabcut": interactive foreground extraction using iterated graph cuts," in *ACM SIGGRAPH 2004 Papers*.

[3] Wan Lo, J. van Baar, C. Knaus, M. Zwicker, and M. H. Gross, "Stereoscopic 3d copy & paste," *ACM Trans. Graph.*, 2010.

[4] Brian L. Price and Scott Cohen, "Stereocut: Consistent interactive object selection in stereo image pairs," *Computer Vision, IEEE International Conference on*, 2011.

[5] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video snapcut : Robust video object cutout using localized classifiers," *ACM SIGGRAPH 2009 papers*, vol. 28, 2009.

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

[7] Yin Li, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum, "Lazy snapping," *ACM Trans. Graph.*, 2004.

[8] Cevahir Cigla and A. Aydin Alatan, "Efficient graph-based image segmentation via speeded-up turbo pixels," in *IEEE International Conference on Image Processing, ICIP 2010.*

[9] H. Emrah Tasli and A. Aydin Alatan, "Interactive object segmentation for mono and stereo applications : Geodesic prior induced graph cut energy minimization," *ICCV Workshop on Human Interaction on Computer Vision*, 2011.

[10] A Criminisi, Toby Sharp, and Andrew Blake, "Geos: Geodesic image segmentation," in *ECCV*, 2008.

[11] D. Scharstein and R. Szeliski, "Middlebury stereo repository," in *http://vision.middlebury.edu/stereo/*, 2010.

[12] Cevahir Cigla and A. Aydin Alatan, "Efficient edge-preserving stereo matching," *ICCV Workshop on Live Dense Reconstruction from Moving Cameras*, 2011.

[13] M. Zhou S. Jiao H. Wang X. Mei, X. Sun and X. Zhang, "On building an accurate stereo matching system on graphics hardware," *ICCV Workshop on GPU for Computer Vision Applications, 2011.*

[14] David Lowe, "Object recognition from local scale-invariant features," *International Conference on Computer Vision, ICCV 1999.*

[15] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," *ECCV 2006.*

[16] E. Rosten and T. Drummond, "Machine learning for high speed corner detection," *ECCV*, 2006.

[17] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," *ECCV*, 2010.

[18] e. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb : an efficient alternative to sift or surf," *ICCV 2011.*