VIDEO OBJECT SEGMENTATION BY EXTENDED RECURSIVE-SHORTEST-SPANNING-TREE METHOD

Ertem Tuncel and Levent Onural

Electrical and Electronics Engineering Department,
Bilkent University, TR-06533, Ankara , Turkey
e-mail:onural@ee.bilkent.edu.tr

ABSTRACT

A new video object segmentation algorithm, which utilizes an extension of recursive shortest spanning tree (RSST) algorithm, is introduced. A 2-D affine motion model is assumed, and correspondingly, for each region, the planar approximation for the given dense motion vector field is computed. Starting from each 2×2 block as a distinct region, the algorithm recursively searches for the best pair of adjacent regions to merge. The "best pair" is defined as the one merging of which causes the least degradation in the performance of the piecewise planar motion vector field approximation. The RSST method is fast, parameter-free and requires no initial guess, unlike the existing algorithms. Moreover it is a hierarchical scheme, giving various segmentation masks from coarsest to finest.

The algorithm successfully captures 3-D planar objects in the scene with acceptable accuracy in the boundaries, which can be further improved by utilizing the spatial information. Improvement over the existing European COST 211 Analysis Model (AM) is observed when the motion segmentation submodule of AM is replaced by the proposed algorithm.

Keywords: Video object segmentation, recursive shortest spanning tree method, hierarchical segmentation, multimedia, video processing.

1. INTRODUCTION

The purpose of video object segmentation is to define over each frame of the sequence a partition that corresponds to the collection of *semantically meaningful* objects. A reasonable assumption is that such objects in the 3-D world are making rigid motion. When projected onto 2-D image plane, rigid motion constitutes a parametric model throughout the range of the projection.

Once a parametric model is picked, a good strategy is to search for regions with coherent motion, i.e., to find regions for which a good parameter set, that explains the observed 2-D motion successfully, exists [1], [2], [3], [4]. In other words, the synthetic motion field reconstructed using the computed parameters should be as close to the original estimated field as possible for all the extracted regions. This strategy is generally known as "segmentation through surface fitting."

It is also logical to utilize spatial information to find correct object boundaries [7], [8], [9], because the estimated motion field inevitably has some errors, especially at the object boundaries [6].

In this paper, an extension of the recursive shortest spanning tree (RSST) method [5] is introduced and utilized to extract regions for which 2-D affine motion model is reasonable. The RSST method is very fast and does not require any *ad hoc* parameters or an initial segmentation mask. It yields a hierarchical segmentation tree, i.e., from finest segmentation (all the pixels are distinct regions) to coarsest (all the frame is a single region). Conventional RSST is currently used in the motion segmentation module of the European COST 211 Analysis Model (AM) [9]. The performance of the described algorithm is compared with that of conventional RSST, when used as a stand-alone algorithm and also when used as a submodule of the Analysis Model.

2. THE ALGORITHM

Given a dense motion vector field $\vec{\mathbf{v}}(x, y)$, the goal of the segmentation algorithm is to extract connected and non-overlapping set of regions so that, for each region R_i, there exists a parameter set $\theta(R_i)$ that explains $\vec{\mathbf{v}}(x, y)$ successfully. The parameter set $\theta(R_i)$, defined by the assumed motion model, implies an approximation for $\vec{\mathbf{v}}(x, y)$ inside R_i.

For 2-D affine motion model, which is adopted in this work, $\theta(R_i) = \{\mathbf{S}_i, \mathbf{T}_i\}$, where $\mathbf{S}_i$ is the 2×2 scaling and rotation matrix, and $\mathbf{T}_i$ is the 2×1 translation vector. The approximation is given by

$$\vec{\mathbf{w}}_{R_i}(x, y) = \mathbf{S}_i \begin{bmatrix} x \\ y \end{bmatrix} + \mathbf{T}_i \ . \tag{1}$$

For a fixed segmentation mask, the approximation error

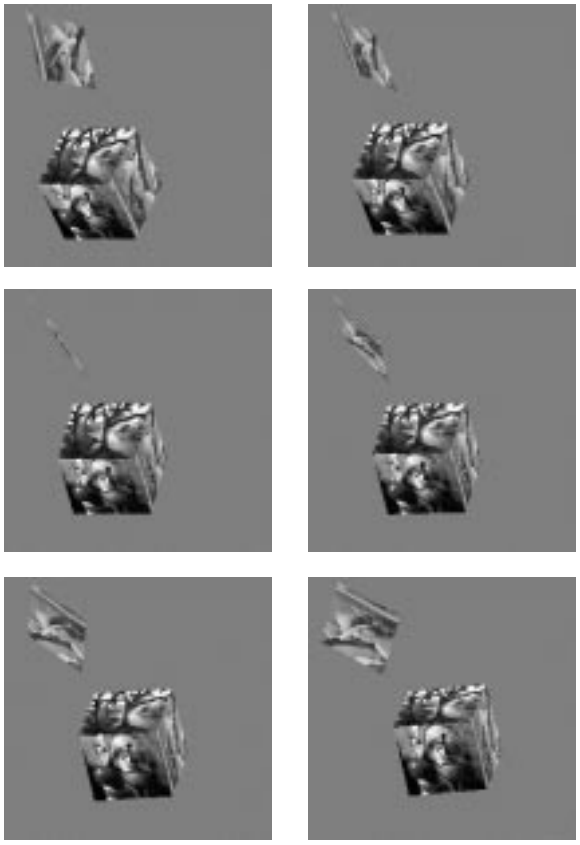

Figure 1: Samples from the artificial sequence.

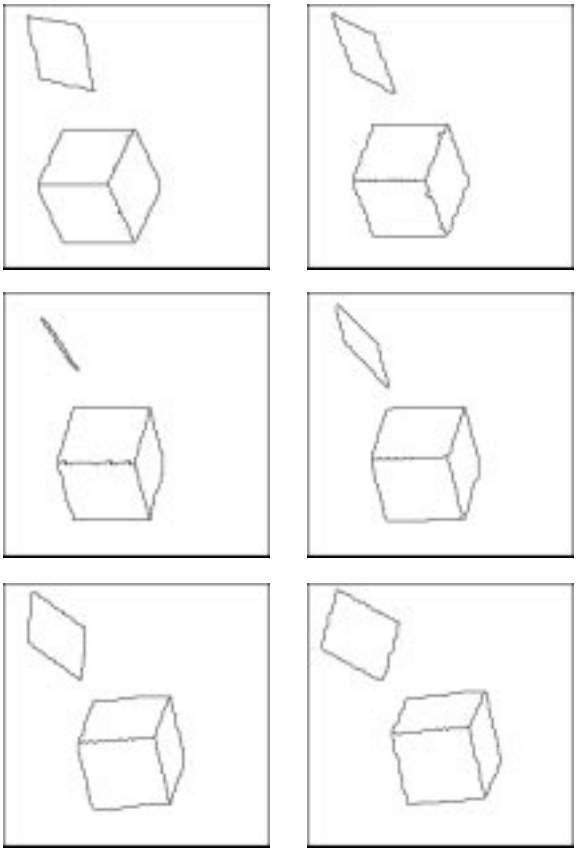

Figure 2: Segmentation of the artificial sequence by the extended RSST algorithm.

is

$$D = \sum_{i=1}^{K} D(R_i) = \sum_{i=1}^{K} \sum_{(x,y)\in R_i} \|\vec{\mathbf{v}}(x,y) - \vec{\mathbf{w}}_{R_i}(x,y)\|^2 \ , \tag{2}$$

where K is the number of regions. Obviously, $D(R_i)$ is minimized by fitting best planes to motion vector components $\mathbf{v}_x(x,y)$ and $\mathbf{v}_y(x,y)$, separately, in the *least squares* sense. However, how to find the optimal set of $\{R_i\}$ minimizing D, is not clear. Searching through all possible segmentation masks is, of course, a solution but it is computationally prohibitive.

There are various methods [1], [2], [3], [4], attacking this problem. A modified K-means algorithm [1], where cluster parameters instead of cluster means are stored and compared, or Bayesian approaches [2], [3], [4], where a term supporting connected regions with smooth boundaries is added to D, and the resultant cost function is minimized using simulated annealing, are examples. These methods are all computationally intensive. Moreover, an initial guess for the segmentation mask and/or some other algorithmic parameters are required from the user.

We propose to use an extension of the recursive shortest spanning tree (RSST) method [5], which represents a segmentation mask as a weighted graph, where regions are considered as nodes and each pair of 4-adjacent regions (R_i, R_j) is connected with a link L_{ij}. The weight d_{ij} associated with link L_{ij} depend only upon $\vec{\mathbf{v}}(x,y)$ inside R_i and R_j.

The graph at the beginning of the algorithm is obtained by dividing the image domain into 2×2 blocks, which are considered as regions. At any intermediate step, if

$$(i^*, j^*) = \arg\min_{i,j} d_{ij} \ , \tag{3}$$

then the link $L_{i^*j^*}$ is removed from the graph, i.e., the corresponding regions R_{i^*} and R_{j^*} are merged. The weights of the links L_{i^*x} and L_{xj^*}, that connect the new node $R_{i^*} \cup R_{j^*}$ to other nodes, are recalculated. Repeating this procedure until there is a single region, we obtain a hierarchical segmentation tree, the K'th level of which has the segmentation mask for K regions.

For the purpose of minimizing D, a logical approach

Figure 3: Samples from the natural sequence.

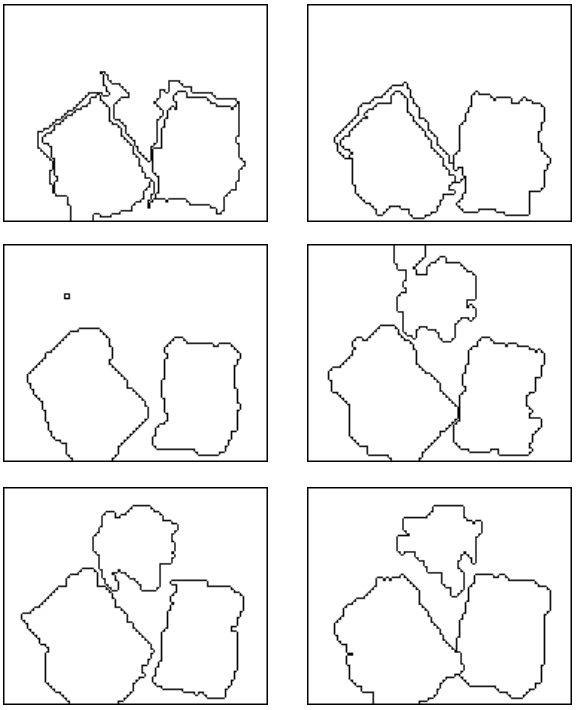

Figure 4: Segmentation of the natural sequence by the extended RSST algorithm.

is to set the weight d_{ij} to

$$\begin{aligned} d_{ij} \quad = \quad & \sum_{(x,y)\in R_i\cup R_j} \|\vec{\mathbf{v}}(x,y) - \vec{\mathbf{w}}_{R_i\cup R_j}(x,y)\|^2 \\ & - \sum_{(x,y)\in R_i} \|\vec{\mathbf{v}}(x,y) - \vec{\mathbf{w}}_{R_i}(x,y)\|^2 \\ & - \sum_{(x,y)\in R_j} \|\vec{\mathbf{v}}(x,y) - \vec{\mathbf{w}}_{R_j}(x,y)\|^2 \ , \end{aligned} \tag{4}$$

where, d_{ij} becomes exactly the amount of increase in D caused by merging regions R_i and R_j.

In the original RSST method, introduced for still image segmentation [5] and later on utilized for motion segmentation [7], [8], d_{ij} is calculated the same way as above. However, it has a much simpler form that involves only the number of pixels and average motion vectors for regions R_i and R_j. That is because the approximation strategy is to fit to each motion vector component a *constant* instead of a *plane*. Fitting constants to motion vectors over regions corresponds to 2-D translational motion model assumption, which is obviously a special case of 2-D affine model.

3. EXPERIMENTAL RESULTS

The first experiment is performed on an 256×256 artificial sequence, shown in Figure 1. The scene, which is orthographically projected onto the 2-D image plane, contains only 3-D planar objects. The purpose of this experiment is to observe the performance of the algorithm under optimal conditions, i.e., the 2-D motion vector field is a priori known since the scene is artificially created, and our model of the motion vector field is exact, because when the projection is orthographic, rigid motion of 3-D planar objects constitute 2-D affine motion vector field.

Figure 2 shows the resulting regions when the number of regions, K, is 5. The 3-D planar surfaces are successfully extracted. This is a promising step towards testing the algorithm with real sequences.

In the second experiment, we used a natural QCIF sequence, shown in Figure 3. In this case, since motion field $\vec{\mathbf{v}}(x,y)$ is not known a priori, a motion estimation algorithm is to be run first. We utilized a regularized Gibbs-formulated motion estimation tool [6], where the cost function to be minimized is a Lagrangian sum of the displaced frame difference and a penalty forcing the smoothness of the dense motion field. Some error in estimated motion is inevitable, especially in untextured areas, and near the object boundaries, due to covered/uncovered background problem and the smoothness term introduced in the cost function.

The proposed extended RSST algorithm is run with $K = 4$. The results are shown in Figure 4. The ob-

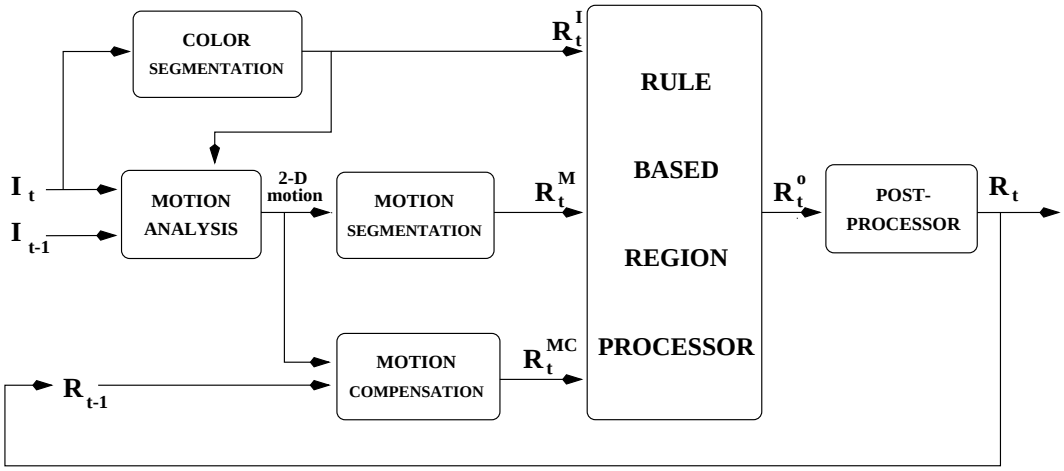

Figure 5: The Block Diagram of the European COST 211 Analysis Model.

jects in the scene whose motion is explainable by the 2-D affine motion model, e.g., the planar objects, are extracted successfully. The inaccuracy of the object boundaries are because of the inevitable error in motion estimation mentioned above.

4. IMPROVEMENTS ON EUROPEAN COST 211 ANALYSIS MODEL

The European COST 211 Analysis Model (AM) version 3.0 is fully described in [9]. The AM offers a new approach for object segmentation and tracking, that is, to fuse motion, color, and accumulated segmentation information at "region level" by rule processing. The algorithm has two modes; the output is either a binary mask describing the "foreground" as a single object, or a segmentation mask describing several objects. The first version of the AM, introduced in [7] and [8], and shown in Figure 5, is embedded in AM 3.0 [9] as an operation mode. The main idea is to utilize the motion segmentation mask to capture the objects in the scene, the color segmentation mask to estimate the true boundaries of objects, and the segmentation mask of the previous frame for tracking the objects.

The motion segmentation block proposed in the AM assumes a 2-D translational model, and hence uses the conventional RSST algorithm, which approximates the motion field $\vec{\mathbf{v}}(x, y)$ by constants instead of planes or higher order surfaces. Figure 6 shows the resultant segmentation masks when our natural sequence is the input to the AM. Here, to demonstrate the object tracking ability of the AM, objects are shown with distinct gray levels. It is observed that 3-D planar objects are not captured as single objects because of the inaccurate motion model. Obviously, a constant motion vector, for the entire 3-D planar object which is rotating, is not a good approximation.

When we replace the motion segmentation module in Figure 5 by the proposed segmentation algorithm, and run AM on the same sequence, we get the results shown in Figure 7. The performance is superior to that of current AM; the books are extracted as a whole, and the head is tracked in all the frames. Note that in either case, the boundaries of the objects are much more accurate compared to the masks in Figure 4. This is because that the AM corrects the boundaries with the help of the color segmentation mask.

5. CONCLUSION

The conventional RSST method searches for the best piecewise *constant* approximation for a given field defined over the 2-D image plane. RSST has various advantages over existing algorithms in the literature; it is fast, free from some ad hoc parameters required from the user, and does not require an initial guess about the segmentation mask. In this work, RSST is extended to a tool that searches for the best piecewise *planar* approximation, and is used to segment the dense motion vector field. Approximation of motion vector field by planes and constants imply 2-D affine and 2-D translational motion models, respectively.

The experimental results indicate that if the estimated motion field is reliable, the algorithm is successful in extracting 3-D planar objects in the scene with acceptable accuracy in the boundaries. The boundaries can be improved by utilizing the spatial information, e.g., the color segmentation mask, as is done in the COST 211 AM.

The motion segmentation module of the AM, which

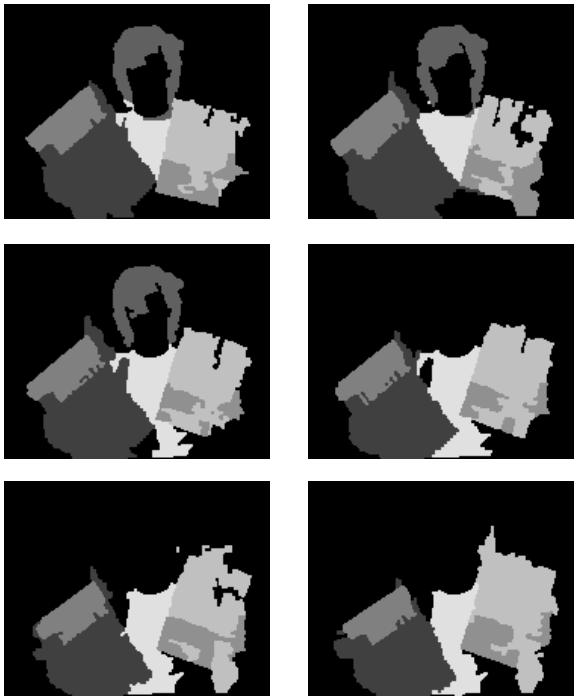

Figure 6: The segmentation result of the AM using the conventional RSST method for motion segmentation.

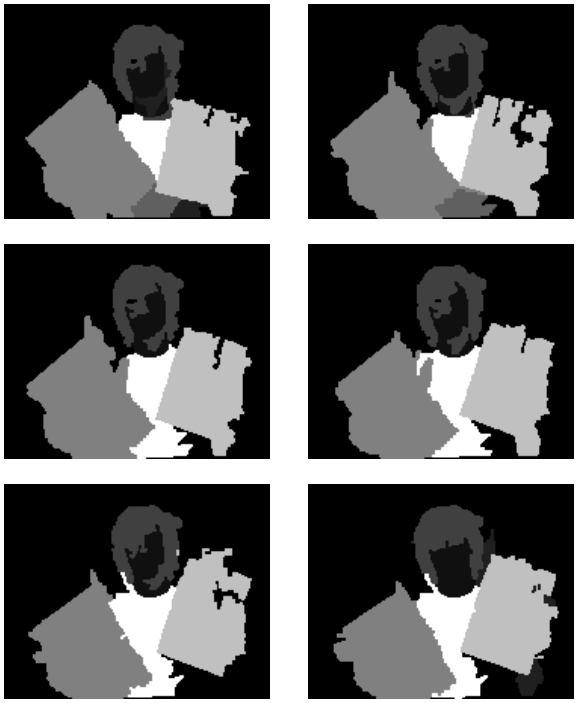

Figure 7: The segmentation result of the AM using the proposed RSST method for motion segmentation.

uses the conventional RSST, is replaced by the proposed motion segmentation method, and tested. Since 2-D affine motion model is more general than the 2-D translational model, the new structure is superior to the current AM in extracting 3-D planar objects as a whole.

The extension of RSST introduced here is fairly general. It can further be extended to cover more accurate motion models, i.e., to approximate the motion vector field components by higher order surfaces.

6. REFERENCES

[1] J. Y. A. Wang, E. Adelson, "Representing Moving Images with Layers", *IEEE Trans. on Image Processing*, vol. 3, no 9, pp 625–638, September 1994.

[2] D. W. Murray, B. F. Buxton, "Scene Segmentation from Visual Motion Using Global Optimization", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 9, no 2, pp 220–228, March 1987.

[3] M. Chang, A.M. Tekalp and M.I. Sezan, "Motion Field Segmentation using Adaptive MAP Criterion", *Proceedings of IEEE ICASSP '93*, pp 33–36, April 1993.

[4] P. B. Chou, C. M. Brown, "The Theory and Practice of Bayesian Image Labeling", *International Joint Computer Vision*, vol. 4, pp 185–210, 1990

[5] O. J. Morris, M. J. Lee, A. G. Constantinides, "Graph Theory for Image Analysis : an Approach Based on the Shortest Spanning Tree", *IEE Proceedings*, vol. 133, no. 2, pp 146–152, April 1986.

[6] A. A. Alatan, L. Onural, "Object-based 3-D Motion and Structure Estimation", *Proceedings of IEEE ICIP '95*, vol. I, pp 390–393, October 1995.

[7] A. A. Alatan, E. Tuncel and L. Onural, "Object Segmentation via Rule-Based Data Fusion", *Workshop on Image Analysis for Multimedia Interactive Services '97*, pp 51–55, June 1997.

[8] A. A. Alatan, E. Tuncel and L. Onural, "A Rule-Based Method for Object Segmentation in Video Sequences", *Proceedings of IEEE ICIP '97*, vol. II, pp 522–525, October 1997.

[9] A. A. Alatan, L. Onural, M. Wollborn, R. Mech, E. Tuncel, and T. Sikora, "Image Sequence Analysis for Emerging Interactive Multimedia Services–The European COST 211 Framework", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 8, no. 7, pp 802–813, November 1998.