Utilization of the Recursive Shortest Spanning Tree Algorithm for Video-Object Segmentation by 2-D Affine Motion Modeling

Ertem Tuncel, *Student Member, IEEE,* and Levent Onural, *Senior Member, IEEE*

***Abstract*—A novel video-object segmentation algorithm is proposed, which takes the previously estimated 2-D dense motion vector field as input and uses the generalized recursive shortest spanning tree method to approximate each component of the motion vector field as a piecewise planar function. The algorithm is successful in capturing 3-D planar objects in the scene correctly, with acceptable accuracy at the boundaries. The proposed algorithm is fast and requires no initial guess about the segmentation mask. Moreover, it is a hierarchical scheme which gives finest to coarsest segmentation results. The only external parameter needed by the algorithm is the number of segmented regions that essentially control the level at which the coarseness the algorithm would stop. The proposed algorithm improves the "analysis model" developed in the European COST211 framework.**

***Index Terms*—Multimedia, recursive shortest spanning tree method, video processing, video-object segmentation.**

I. Introduction

VIDEO-OBJECT segmentation refers to defining a partition on the frames of a video sequence. The regions of that partition should correspond to *semantically* meaningful objects in the scene. It is logical to utilize the *temporal* as well as *spatial* information in the segmentation process.

Temporal information is usually utilized by estimating the motion field and then searching for regions with *coherent* motion [1]–[4]. However, because of the possible errors in the motion-estimation step, regions with coherent motion usually turn out to have inaccurate boundaries. So, spatial information is also necessary in order to find exact boundaries of objects [5]–[7].

During the search for regions with coherent motion, it is helpful to assume that objects in the scene make *rigid* motion in the 3-D world. The projection of rigid motion onto the 2-D image plane constitutes a *parametric model* throughout the 2-D projection of the object. Once a parametric model is assumed, a good strategy is to search for regions on the 2-D image plane for which a good parameter set that explains the observed motion successfully exists. This search is known as the strategy of *segmentation through surface fitting* and can also be applied to still-image segmentation [8]–[10].

In this paper, a novel video-object segmentation algorithm based on the generalization of recursive shortest spanning tree (RSST) method [11] is introduced. The RSST method is a powerful method in the sense that it is fast and requires no initial segmentation masks. Furthermore, it is a hierarchical segmentation scheme, i.e., yielding segmentation masks of various scales, from the *finest* to the *coarsest*; as the algorithm evolves from the finest level to the coarser levels, it may be stopped when the number of regions is reduced to the desired number, K, which should be specified externally. This novel algorithm is also plugged in the emerging "analysis model" (AM) developed in the European Cost211 framework [5]–[7].

Organization of this paper is as follows. Section II introduces the model and the cost function for the search of the best segmentation mask, and then after briefly discussing the disadvantages of the existing methods, describes the generalized RSST method which aims to minimize the cost function. In Section III, the experimental work is described and the results are given. Section IV briefly describes the AM, and shows some results when its motion segmentation module is replaced by the proposed one. Finally, Section V concludes the paper.

II. Segmentation Algorithm

A. The Model and the Cost Function

An affine (or six-parameter) model is assumed for the dense motion vector field $\vec{\mathbf{v}}(x,y)$. Given an estimate for $\vec{\mathbf{v}}(x,y)$, the algorithm should try to extract regions for which a good parameter set exists, i.e., explaining the estimated motion field successfully. Extracting parameters for a fixed region is equivalent to fitting surfaces for each motion field component $\mathbf{v}_x(x,y)$ and $\mathbf{v}_y(x,y)$ in that region. The approximated surfaces constitute the *synthesized vector field* in terms of the extracted parameters and are denoted as $\vec{\mathbf{w}}^R(x,y)$ for each region R. For affine motion model, the relation between the parameters and the approximated surfaces is

$$\begin{aligned}\mathbf{w}_x^R(x,y) &= a_1^R x + a_2^R y + a_3^R \\ \mathbf{w}_y^R(x,y) &= a_4^R x + a_5^R y + a_6^R.\end{aligned} \tag{1}$$

The cost function for a fixed segmentation mask is

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

where K is the number of regions and $\{R_i\}$ are the nonoverlapping regions. It is obvious that, for a fixed set of regions $\{R_i\}$, the surface fitting or equivalently the extraction of the optimal parameters $\{a^{R_i}\}$ must be done in the *least squares* sense to

Manuscript received December 15, 1998; revised January 26, 2000. This paper was recommended by Associate Editor R. Koenen.

E. Tuncel was with Bilkent University, TR-06533 Ankara, Turkey. He is now with the University of California, Santa Barbara, CA 93105 USA.

L. Onural is with Bilkent University, TR-06533 Ankara, Turkey.

Publisher Item Identifier S 1051-8215(00)06561-7.

minimize D. The problem is to find an optimal set of R_i to minimize D.

B. Existing Methods

A modified K-means algorithm, where optimal parameters for clusters are stored instead of the cluster means, is used in [1]. The method iterates between assignment of pixels to clusters and reoptimization of cluster parameters. The performance depends on a good initialization and a resultant region might have disconnected areas.

Bayesian approaches [2]–[4] can be a remedy; spatial connectivity is supported by adding to D a new term, which penalizes discontinuities in the segmentation field, and local minima is avoided by utilizing the *simulated annealing* method or similar methods to minimize the overall cost function. However, even a single iteration, i.e., for a fixed temperature, is computationally very intensive. Furthermore, these methods require *ad hoc* determination of some algorithmic parameters, e.g., a Lagrangian weight for the additional term in D, or a parameter for stopping criterion for each iteration.

Our proposal, the generalization of the RSST method, described in Section II-C, is free of iterations, *ad hoc* parameters (except for the number of regions K, see Section II-D), and the need for an initial guess for the segmentation field.

C. Generalized RSST Method

The original version of RSST is introduced in [11]. The RSST method looks only one step ahead to minimize D, i.e., it minimizes a certain ΔD at each step.

The 2-D discrete image domain is converted into a graph, where each node represents a region, and each link represents the 4-adjacency between regions. This is the initialization step, achieved by dividing the image domain into $N \times N$ blocks and assuming each block to be a region. Associated with each node R_i, there is a set of optimal parameters $\{a_k^{R_i}\}$ to construct $\vec{\mathbf{w}}^{R_i}(x,y)$. Associated with each link $L(i,j)$, there is a *distance* $d(R_i,R_j)$. At each step, the link $L(i^*,j^*)$, where

$$(i^*,j^*) = \arg\min_{i,j} d(R_i,R_j) \tag{3}$$

is removed from the graph. This removal corresponds to merging of adjacent regions R_{i^*} and R_{j^*} to form a new region. The optimal parameters of the merged region, and the distance values assigned to the links departing from that region, are to be calculated. Repeating this procedure, the number of regions can be reduced down to one. The removed links construct a so-called *spanning tree* of the initial graph. If the order of removal of links is recorded, the segmentation mask for an arbitrary number of regions, K, can be found by unremoving the last $K-1$ links.

For the minimization of D, it is logical to set

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

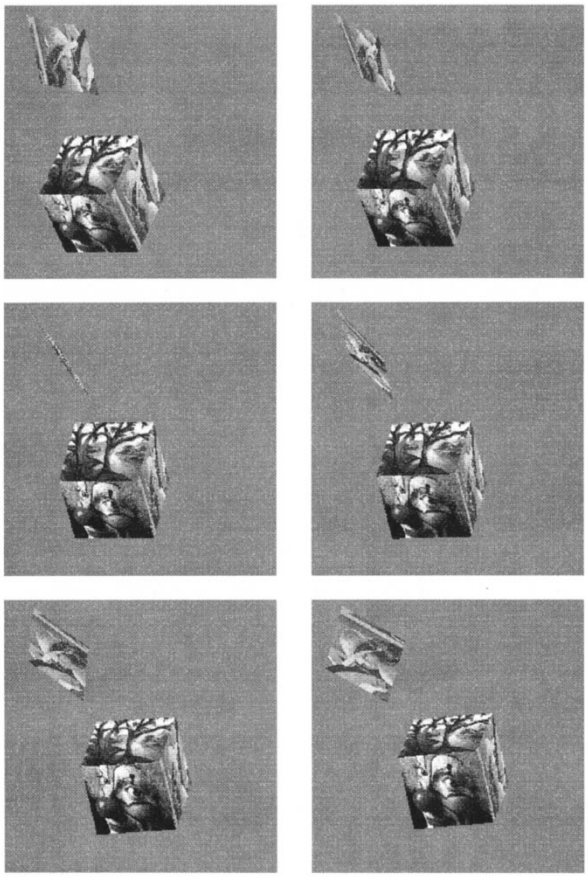

Fig. 1. Samples from the artificially generated sequence.

where $R_{ij} = R_i \cup R_j$. In other words, $d(R_i,R_j)$ is the change in D when regions R_i and R_j are merged.

This distance measure involves the knowledge of the optimal surface approximation $\vec{\mathbf{w}}^{R_{ij}}(x,y)$ for every $L(i,j)$ on the graph, at each intermediate step. However, since $d(R_i,R_j)$ is calculated for each link $L(i,j)$ at the initialization phase, there is only the need to recalculate $L(i,j^*)$ and $L(i^*,j)$ at intermediate steps where R_{i^*} and R_{j^*} are merged. We adopt this distance measure in our algorithm with the block size $N=2$. This is the smallest block size possible for the six-parameter motion field model because infinitely many least-squares solutions exist for a_k^R unless region R contains at least three pixels that are not collinear.

In the original RSST described in [11], the objective of the algorithm is to find a good piecewise-constant approximation to the given field; this corresponds to an implicit assumption of a 2-D translational motion model when the motion field is the input to the algorithm. Our modified RSST is more general because we are able to assume higher order models, such as the affine motion model, and search for a good piecewise-smooth approximation to the given motion field.

D. How to Choose K

One of the most challenging problems of unsupervised segmentation algorithms is to determine the number of regions, K. Obviously, the performance depends on the specified K. However, there is no single "true" K. On the contrary, what K

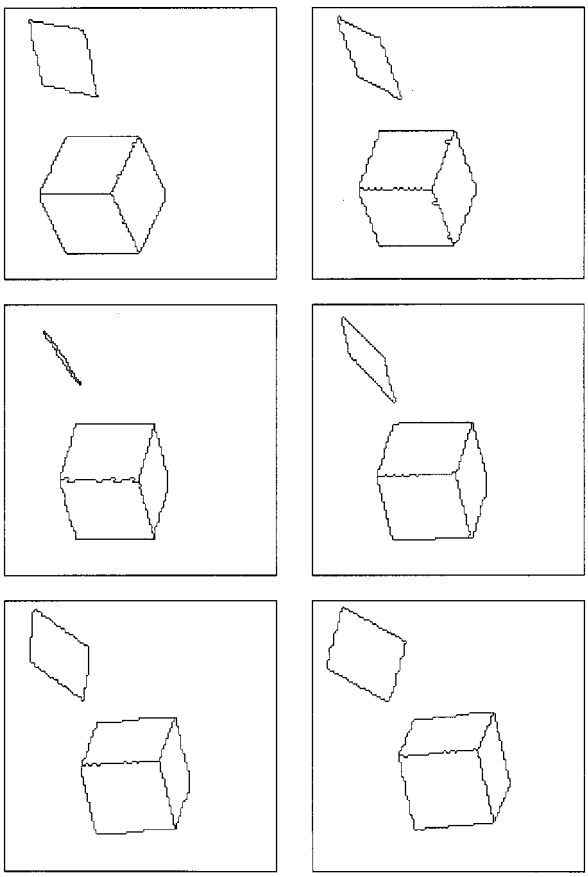

Fig. 2. Segmentation of the artificially generated sequence with the generalized RSST algorithm ($K = 5$).

should be depends on the application. For example, for an object-based coding application, the optimum value for K is the one that achieves the minimum overall bit rate. This principle is known as the *minimum description length*. Another application (like MPEG-4 video-object plane generation) may prefer segmentation to semantically meaningful objects. In such a case, the automated objective selection of K may be impossible; user interaction might be needed.

Being a hierarchical scheme, the described RSST method has the advantage of giving the user a chance to pick a segmentation result from a chain of segmentation results from finest ($K =$ number of pixels$/4$) to coarsest ($K = 1$). On the other hand, in the existing methods, one has to rerun the algorithm to change K.

III. Experimental Work and Results

The experiments are performed on two sequences of different kinds. The first is a 256×256 artificial sequence, a short part of which is shown in Fig. 1, consisting of pure 3-D planar objects which are orthographically projected onto the 2-D image plane. The motion-estimation step is bypassed, since there is the *a priori* knowledge of the motion vector field. This is to guarantee that the performance of the algorithm is not deteriorated by the errors introduced during the estimation of motion field. Furthermore, since rigid motions of 3-D planar objects constitute an affine motion field when orthographically projected onto

Fig. 3. Samples from the natural sequence.

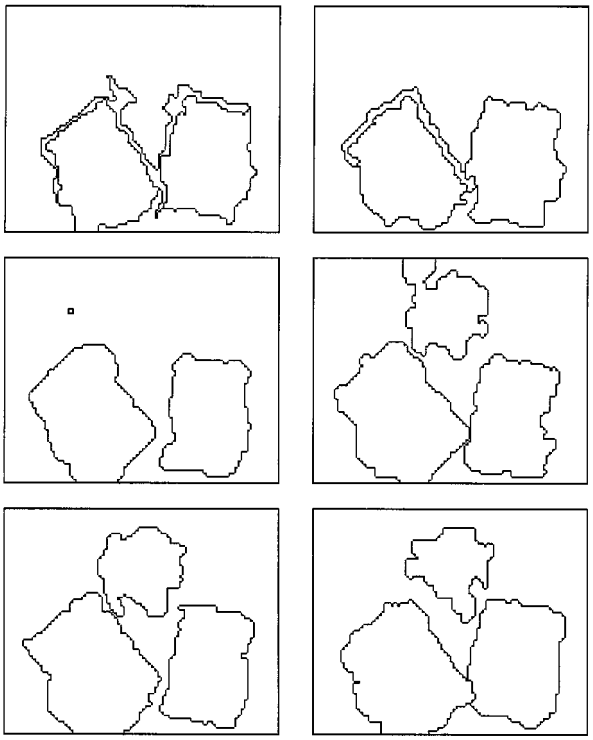

Fig. 4. Segmentation of the natural sequence with the generalized RSST algorithm ($K = 4$).

2-D image plane, the algorithm is being tested under optimal conditions. The algorithm is executed with number of regions K, set to five. As seen from Fig. 2, the plane-fitting strategy is successful in extracting all the meaningful *parts* of objects. Although, this is only to show that the algorithm does a good job

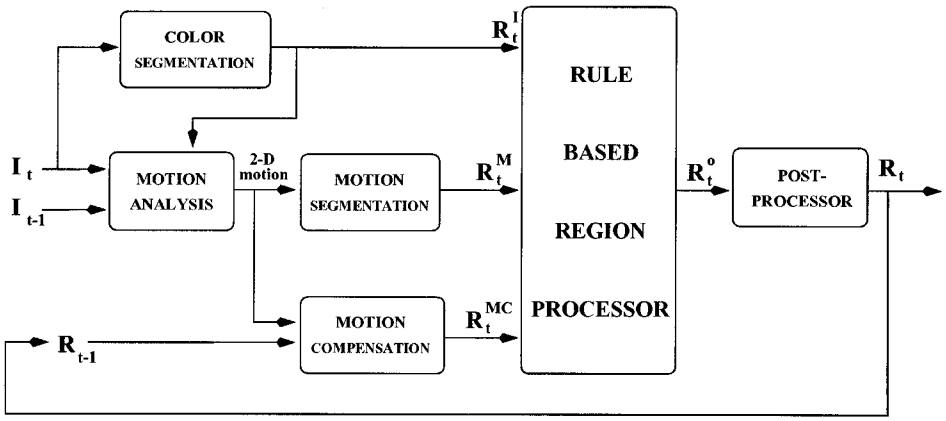

Fig. 5. Block diagram of the AM.

under optimal conditions, it is a promising step toward the application on realistic cases.

The second sequence, which is in QCIF size 176×144, is a natural one; some of its frames are shown in Fig. 3. The surfaces of primary interest are again of 3-D planar objects. The motion-estimation step is necessarily executed and this is to test the segmentation algorithm under more realistic cases where motion field is not known *a priori*. Especially near the object boundaries and untextured areas, estimation of the motion involves some errors. In this case, first the motion field $\vec{\mathbf{v}}(x,y)$ is estimated using the Gibbs-based algorithm in [12], then the segmentation algorithm is executed with $K = 4$. The results are as shown in Fig. 4. Although there is some inaccuracy in the estimated motion field because of either the covered/uncovered background problems, or the areas lacking enough texture, the performance of the segmentation algorithm is still good. The planar objects in the scene are captured precisely. The inaccuracy in the detection of object boundaries come from the fact that the most unreliably estimated motion vectors are the ones in the vicinity of object boundaries. When the indicated number of regions K (four, in this example) fits the number of semantically meaningful objects as in the bottom row of Fig. 4, the segmentation results are better. When K does not fit the number of semantically meaningful objects, as in the top row of Fig. 4 where the head is interpreted as a part of background since it is also still at that moment, undesired regions as seen at the borders of the books are inevitable.

IV. Improvements on the Analysis Model of COST211ter

The first version of the emerging AM of the European COST211 project, which aims to fulfill the *object definition* and *object tracking* functionalities, is described in detail in [5] and [6]. The latest version is described in [7]. The AM offers a novel approach for object segmentation and tracking, where motion, color, and accumulated segmentation information can be fused at the "region level" by the help of some predefined rules.

The block diagram of the model is shown in Fig. 5. Various types of segmentation masks, such as color- and motion-based segmentation results and the segmentation result of the previous frame are given as inputs to the rule-based decision box which determines the segmentation mask for the current frame. Fusion of segmentation results via a rule-based decision process leads to good segmentation results by utilizing the motion-based regions to *capture* the objects in the scene, and the color-based regions to extract the *true boundaries* of these objects. The segmentation result of the previous frame serves as a temporally accumulated segmentation information, which is essential for *tracking*.

The major disadvantage of this model is the implicit 2-D translational motion field model assumption it imposes. This assumption comes from the fact that the motion segmentation module in the AM executes the conventional RSST, in which the motion vector field $\vec{\mathbf{v}}(x,y)$ is approximated by *constants* instead of planes.

However, the algorithm described in the previous section can readily be plugged in the AM. Fig. 6 shows the results of the AM proposed in [5] with the natural sequence as its input. This time, the regions are painted with distinct gray levels in order to show the achievement of object tracking functionality promised by AM. The 3-D planar surfaces in the scene are hardly captured as *single* objects. The algorithm splits them into several parts because a *piecewise constant* approximation cannot explain the observed motion field $\vec{\mathbf{v}}(x,y)$ without assuming such split objects.

The results with the described motion segmentation tool plugged in is shown in Fig. 7. Now the algorithm is much more successful in capturing the books in the scene as single objects. Note that in both Figs. 6 and 7, the boundaries of the objects are much better. This is because in the AM, the color segmentation results are utilized to estimate the boundaries of the captured objects. This is the most advantageous merit of the AM. Note also that the number of regions is not imposed on the resultant segmentation field. Instead, it is determined by the region merging rules in the rule processing module.

V. Conclusion

The main work done in this paper is the development of a novel video-object segmentation algorithm, based on the generalization of a conventional image-segmentation tool, namely the RSST method. The original RSST method applied to the es-

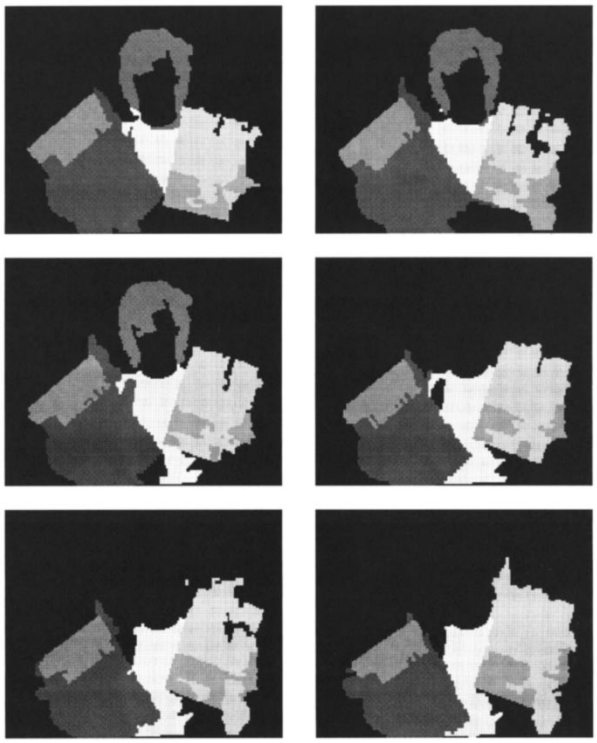

Fig. 6. Segmentation result of the AM using the conventional RSST method for the motion segmentation block.

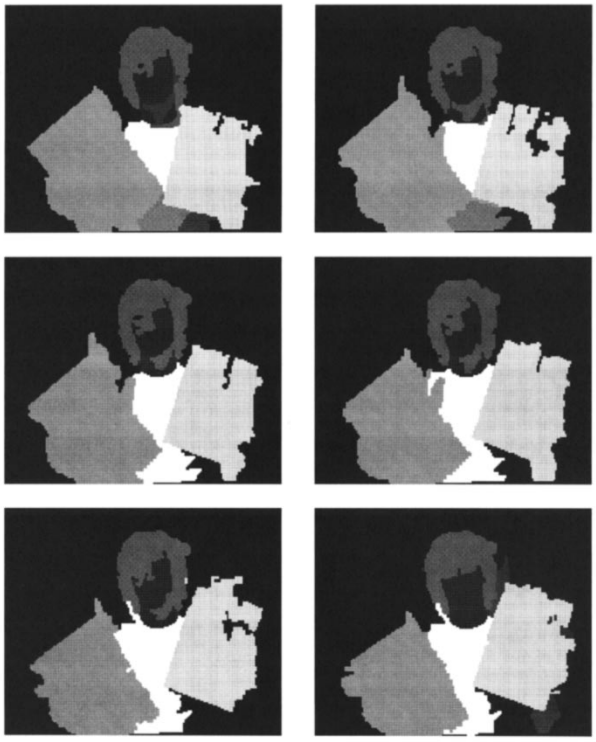

Fig. 7. Segmentation result of the AM using the proposed RSST method for the motion segmentation block.

timated motion field approximates the motion field as a piecewise constant function. In this paper, a more general approach, where the motion field is approximated by planes, is presented.

If the motion vectors are estimated reliably, the resultant segmentation is successful in capturing the 3-D planar objects in the scene; the accuracy at the boundaries of the objects in the real-life video is acceptable. RSST has some advantages over existing algorithms in the literature, such as being free from the determination of "initial" parameters and from presetting "weights" of different parts of the involved cost function. Furthermore, being a hierarchical algorithm, by a single run, the RSST method gives a chain of segmentation results, varying from the finest to the coarsest.

Another major advantage is that, without sacrificing from the performance, the RSST method achieves a much lower computational cost compared to existing motion-segmentation algorithms. For example, in [2], the simulated annealing algorithm, which is famous for its computational burden, is used. In [3], the computational cost is still high although it is significantly lowered by using iterative conditional modes (ICM), instead of simulated annealing, for the minimization. Perhaps the most efficient all the existing algorithms is presented in [1], where a modified K-means clustering algorithm is used. However, it is an iterative algorithm and the computational cost heavily depends on the stopping criterion, which imposes a tradeoff between computational performance and accuracy. On the other hand, the proposed RSST algorithm is not iterative. At each hierarchical level, after the merging of the best pair of regions, only a few links are updated by solving the corresponding 3×3 linear systems. Moreover, the "best" pair of regions is always kept track of in a very efficient way.

The emerging AM of the European COST211$^{\text{ter}}$ project aims to achieve the object detection and tracking functionalities in an unsupervised way. The proposed video-object segmentation tool can readily be plugged into the AM, whose motion-segmentation module uses the conventional RSST. The replacement of the conventional RSST in the AM by the generalized RSST results in a better segmentation, as expected.

REFERENCES

[1] J. Y. A. Wang and E. Adelson, "Representing moving images with layers," *IEEE Trans. Image Processing*, vol. 3, pp. 625–638, Sept. 1994.

[2] D. W. Murray and B. F. Buxton, "Scene segmentation from visual motion using global optimization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, pp. 220–228, Mar. 1987.

[3] M. Chang, A. M. Tekalp, and M. I. Sezan, "Motion field segmentation using an adaptive MAP criterion," in *Proc. IEEE ICASSP '93*, vol. 5, Apr. 1993, pp. 33–36.

[4] P. B. Chou and C. M. Brown, "The theory and practice of Bayesian image labeling," *Int. J. Comput. Vis.*, vol. 4, pp. 185–210, 1990 .

[5] A. A. Alatan, E. Tuncel, and L. Onural, "Object segmentation via rule-based data fusion," in *Workshop Image Analysis for Multimedia Interactive Services (WIAMIS'97)*, June 1997, pp. 51–55.

[6] ——, "A rule-based method for object segmentation in video sequences," in *Proc. IEEE ICIP'97*, vol. 2, Oct. 1997, pp. 522–525.

[7] A. A. Alatan, L. Onural, M. Wollborn, R. Mech, E. Tuncel, and T. Sikora, "Image sequence analysis for emerging interactive multimedia services—The European COST211 framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 802–813, Nov. 1998.

[8] P. J. Besl and R. C. Jain, "Segmentation through variable-order surface fitting," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, pp. 167–192, Mar. 1988.

[9] A. Leonardis, A. Gupta, and R. Bajcsy, "Segmentation as the search for the best description of the image in terms of primitives," in *Proc. 3rd Int. Conf. Computer Vision*, 1990, pp. 121–125.

[10] A. Ackah-Miezan and A. Gagalowicz, "Discrete models for energy-minimizing segmentation," in *Proc. 4th Int. Conf. Computer Vision*, 1993, pp. 200–207.
[11] O. J. Morris, M. J. Lee, and A. G. Constantinides, "Graph theory for image analysis: An approach based on the shortest spanning tree," in *Proc. Inst. Elect. Eng.*, vol. 133, Apr. 1986, pp. 146–152.
[12] A. A. Alatan and L. Onural, "Object-based 3-D motion and structure estimation," in *Proc. IEEE ICIP'95*, vol. 1, Oct. 1995, pp. 390–393.

Ertem Tuncel (S'99) was born in Antalya, Turkey, in 1974. He received the B.S. degree from Middle East Technical University, Ankara, Turkey, in 1995, and the M.S. degree from Bilkent University, Ankara, Turkey, in 1997, both in electrical engineering.

After 1997, he joined Ph.D. program in the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he is currently a Research Assistant. His research interests include pattern recognition, optimization with deterministic annealing methods, and rate-distortion theory with emphasis on scalable source coding.

Levent Onural (S'82–M'85–SM'91) was born in Izmir, Turkey, in 1957. He received the B.S. and M.S. degrees in electrical engineering from Middle East Technical University, Ankara, Turkey, in 1979 and 1981, respectively, and the Ph.D. degree in electrical and computer engineering from State University of New York at Buffalo in 1985. He was a Fulbright scholar between 1981 and 1985.

After a Research Assistant Professor position at the Electrical and Computer Engineering Department, State University of New York at Buffalo, he joined the Electrical and Electronics Engineering Department, Bilkent University, Ankara, Turkey, where he is currently a Professor. He visited the Electrical and Computer Engineering Department, University of Toronto, on a sabbatical leave between September 1994 and February 1995. His current research interests are in the area of image and video processing, with emphasis on very low bit-rate video coding and video-object segmentation, texture modeling, nonlinear filtering, holographic TV, and signal-processing aspects of optical wave propagation.

Dr. Onural was the Organizer and first Chair of the IEEE Turkey Section (1989–1991). He also served as the Chair of the IEEE Circuits and Systems Chapter in Turkey (1994–1996). He was the Chair of the IEEE Region 8 (Europe, Africa, and Middle East) Student Activities Committee (1995–1998), the Vice Chair of the IEEE Regional Activities Board (1998–1999), and Chair of the IEEE Student Activities Committee (1998–1999). He has also served in many other IEEE committees. Recently, he was elected to the IEEE Board of Directors (2001–2002) as the Director of IEEE Region 8. Additionally, he is the General Co-Chair of the IEEE ICASSP 2000 Conference. In 1995, he received the Young Investigator Award from TÜBITAK in 1995.