In Proc. Int. Conf. on Image Processing, ICIP-2002
Sep. 22–25, 2002, Rochester, NY, USA

JOINT SPACE-TIME IMAGE SEQUENCE SEGMENTATION BASED ON VOLUME COMPETITION AND LEVEL SETS

Janusz Konrad and Mirko Ristivojević

Boston University, Department of Electrical and Computer Engineering
8 Saint Mary's Street, Boston, MA 02215, USA
[jkonrad,mirko]@bu.edu

ABSTRACT

In this paper, we address the issue of joint space-time segmentation of image sequences. Typical approaches to such segmentation consider two image frames at a time, and perform tracking of individual segmentations across time. We propose to perform this segmentation jointly over multiple frames. This leads to a 3-D segmentation, i.e., search for a volume "carved out" by a moving object in the (3-D) image sequence domain. We pose the problem in Bayesian framework and use the MAP criterion. Under suitable structural and segmentation/motion models we convert MAP estimation to a functional minimization. The resulting problem can be viewed as *volume competition*, a 3-D generalization of region competition. We parameterize the unknown surface to be estimated, but rather than solving for it using an active-surface approach, we embed it into a higher-dimensional function and use the level-set methodology. We show experimental results for the simpler case of object motion against still background although, given suitable models, the general formulation can handle complex motion too.

1. INTRODUCTION

In most studies to date, image sequences are primarily analyzed and processed in groups of two frames; by differentiating one frame from the other, one is able to infer the dynamics occurring in an image sequence. These short-term dynamics (such as displacement between two frames, or occlusion/exposure areas) can be linked together or temporally constrained in order to reason about longer term dynamics. Although the two-frame approach has been very successful in some applications (e.g., MPEG compression standards), it is often inadequate for the analysis of non-constant velocity motion, detection of innovation areas (occlusion and exposure), or video segmentation.

The segmentation of an image sequence into moving objects is closely related to the estimation of motion for each object. In general, accurate segmentation requires the knowledge of each object's motion parameters, whereas accurate estimation of each object's motion parameters greatly benefits from sequence segmentation. Solving both problems jointly leads to the joint motion estimation/segmentation. The early attempts to solve this problem involved simple thresholding, while later methods applied energy minimization [1] or Markov random field (MRF) models [2]. In order to automatically find the number of motion classes, K-means clustering [3] and mixture models under MDL formulation [4] have been studied.

More recently, motion segmentation methods based on active contours have been proposed. The original active-contour formulation [5] suffers from stability problems and fixed topology. These issues can be resolved by reformulating the problem *via* embedding the contour into a higher-dimensional function of which it is a zero-level set [6]. Developed originally for the modeling of flame propagation, the approach has found numerous applications in computer vision and image processing. Recently, Caselles *et al.* [7] showed that energy-minimizing active contours are related to the level set formalism by means of *geodesic active contours* (i.e., minimal distance paths) in a Riemannian space.

In the context of motion segmentation, geodesic active contours have been applied to statistically-derived motion boundary maps [8], and to tensor-derived maps [9, 10]. In each case, the curve evolution stops at large gradients of the motion-boundary map, which are closely related to the intensity gradient ∇I; the approach can be considered edge-based. An alternative approach is to consider all intensities in a region, for example by means of region competition [11], as recently proposed by Mansouri and Konrad [12, 13], Jehan-Besson *et al.* [14] and Debreuve *et al.* [15].

The above approaches perform image sequence segmentation on the basis of two image frames at a time, and therefore cannot take longer-term dynamics into account. Some early work using multiple frames includes motion detection using 3-D MRF models [16], "video-cube" segmentation based on marker selection and volume growing [17], and 3-D extension of a discretized version of the Mumford-Shah functional [18]. Also, motion boundary detection in the $x-y-t$ space, that is related to multi-frame segmentation, has been proposed [19] (see next section).

In this paper, we propose a novel framework for multiple-frame motion estimation and segmentation. The proposed framework is Bayesian, and by a suitable choice of models leads to a volume-competition formulation in space-time, where the competing volumes are described by a parametric *active surface* (3-D equivalent of active contour). The resulting cost functional is minimized using level set methodology.

2. SEGMENTATION BASED ON ACTIVE SURFACES

In order to formulate image sequence segmentation jointly over several frames, it is natural to consider an extension of active contours to active surfaces. This leads to minimal-surface formulations that have been applied to 3-D shape recovery [20, 21]. A path particularly pertinent to this paper has been undertaken by Caselles *et al.* [21]. Let $I : \Omega \times \mathcal{T} \rightarrow R^+$ be intensity and let $\vec{\varsigma}$ be a surface in 3-D space with area $\mathcal{S}$. By parameterizing the surface, $\vec{\varsigma}(p,q) : [0,1] \times [0,1] \rightarrow R^3$ with $p = (x,y,z)$ and $q = q(x,y,z)$, Caselles *et al.* have proposed to compute the min-

imal surface englobing a 3-D object as follows:

$$\min_{\vec{\varsigma}} \iint_{\mathcal{S}} g(\delta I) d\vec{\varsigma} \;\; \rightsquigarrow \;\; \frac{\partial \vec{\varsigma}}{\partial t} = [g(\delta I)\kappa_m - \nabla g(\delta I) \cdot \vec{n}]\vec{n},$$

where $g(\cdot)$ is a strictly decreasing function, δI is a measure of intensity variation, $d\vec{\varsigma}$ is a Euclidean area element, κ_m is the mean curvature and $\vec{n}$ is the inward unit normal to $\vec{\varsigma}$. The term $g(\delta I)\kappa_m\vec{n}$ smoothes out the contour by reducing the curvature, unless $g(\delta I)$ is zero which means a large intensity change (e.g., perfect edge). The term $(\nabla g(\delta I) \cdot \vec{n})\vec{n}$ "pushes" the contour towards an intensity edge as long as the orthogonal component of ∇g is non-zero. This term allows locking to edges with intensity variations or even gaps along the edge. This approach has been further extended by the same authors:

$$\min_{\vec{\varsigma}} \iiint_{\mathcal{V}} f(\delta I) d\varpi + \lambda \iint_{\mathcal{S}} g(\delta I) d\vec{\varsigma}, \tag{1}$$

where the new term is a measure of the Euclidean volume element $d\varpi$ weighted by $f(\delta I)$, and $f(\cdot)$ is another strictly decreasing function. The new term adds a constant "balloon" force $f(\delta I)\vec{n}$ to the surface evolution equation helping avoid local minima and speeding up convergence.

Depending on the definition of the measure of intensity variation δI, different applications have been devised. With $g(\delta I)$ defined as $1/(1 + |\nabla I|^p)$, and $p = 1$ or 2, the geodesic surface computation has been applied to MRI (magnetic resonance imaging) data segmentation [21]. With $\delta I = |I_t|/(I_x^2 + I_y^2)^{1/2}$, i.e., normal component of optical velocity (where I_x, I_y, I_t are horizontal, vertical and temporal intensity derivatives, respectively), (1) has been used for motion detection against static background [19] with very interesting results for synthetic data.

3. PROBLEM FORMULATION

We pose the problem in the framework of maximum *a posteriori* probability (MAP) estimation. Let again $\vec{\varsigma}$ be a parameterized surface in the $x-y-t$ space, let I_t be an image frame at time t, and $\mathcal{I}^t = \{I_\tau : t - T \leq \tau \leq t + T\}$ be a subset of image sequence based on which $\vec{\varsigma}$ is estimated. Finally, let $\boldsymbol{p}$ and $\bar{\boldsymbol{p}}$ be motion parameters (e.g., affine with constant or slowly-varying velocity [22]) for the volume inside and outside of $\vec{\varsigma}$, respectively; we assume that motion trajectory for each image point can be computed either from $\boldsymbol{p}$ or $\bar{\boldsymbol{p}}$. Following the framework for two-frame segmentation [12], the MAP-based multiple-frame segmentation can be then expressed as follows:

$$\begin{aligned} &\max_{\vec{\varsigma},\boldsymbol{p},\bar{\boldsymbol{p}}} \; p(\vec{\varsigma}, \boldsymbol{p}, \bar{\boldsymbol{p}} | \mathcal{I}^t) = \\ &\max_{\vec{\varsigma},\boldsymbol{p},\bar{\boldsymbol{p}}} \; p(I_t | \vec{\varsigma}, \boldsymbol{p}, \bar{\boldsymbol{p}}, \mathcal{I}^t \backslash \{I_t\}) p(\vec{\varsigma}, \boldsymbol{p}, \bar{\boldsymbol{p}} | \mathcal{I}^t \backslash \{I_t\}), \end{aligned} \tag{2}$$

where p denotes probability density. For the likelihood term, we propose the following spatio-temporal structural model:

$$\begin{aligned} I(\boldsymbol{x}, t) &= \mu(\boldsymbol{x}, t; \boldsymbol{r}_i) + \eta_i(\boldsymbol{x}, t), \\ \boldsymbol{r}_1 &= \boldsymbol{p}, \;\; \eta_1 \rightsquigarrow \mathcal{N}(0, \sigma_1^2) \quad \text{if } (\boldsymbol{x}, t) \text{ inside } \vec{\varsigma} \\ \boldsymbol{r}_2 &= \bar{\boldsymbol{p}}, \;\; \eta_2 \rightsquigarrow \mathcal{N}(0, \sigma_2^2) \quad \text{if } (\boldsymbol{x}, t) \text{ outside } \vec{\varsigma} \end{aligned} \tag{3}$$

where $\mu(\boldsymbol{x}, t; \cdot)$ is an average intensity along motion trajectory, and η is an independent identically-distributed zero-mean Gaussian random variable. Note that different motion parameters and noise variances are assigned to points inside and outside of $\vec{\varsigma}$. This model basically expresses intensity at $(\boldsymbol{x}, t)$ as an average value plus perturbation. As for the prior, we assume for now independence of $\vec{\varsigma}$ from $\mathcal{I}^t \backslash \{I_t\}$ thus ignoring the direct impact of spatio-temporal intensity edges on the shape of $\vec{\varsigma}$. Furthermore, we assume for now independence between $\vec{\varsigma}$, $\boldsymbol{p}$ and $\bar{\boldsymbol{p}}$, and uniform distributions for $\boldsymbol{p}$ and $\bar{\boldsymbol{p}}$. Under these assumptions we have $p(\vec{\varsigma}, \boldsymbol{p}, \bar{\boldsymbol{p}}) \propto p(\vec{\varsigma})$. Since we wish to describe the surface $\vec{\varsigma}$ most compactly (lowest bit rate), we choose the prior to be a function of the area $\mathcal{S}$ of $\vec{\varsigma}$. These assumptions lead to the following minimization:

$$\begin{aligned} \min_{\vec{\varsigma},\boldsymbol{p},\bar{\boldsymbol{p}}} \; \alpha \iiint_{\mathcal{V}} \xi(\boldsymbol{x}, t; \boldsymbol{p}) d\boldsymbol{x} dt \quad + \\ \iiint_{\bar{\mathcal{V}}} \xi(\boldsymbol{x}, t; \bar{\boldsymbol{p}}) d\boldsymbol{x} dt \quad + \quad \lambda \iint_{\mathcal{S}} d\vec{\varsigma}, \end{aligned} \tag{4}$$

where $\vec{\varsigma} = \partial\mathcal{V}$, $\mathcal{V} \cup \bar{\mathcal{V}} = \Omega \times \mathcal{T}$ ($\mathcal{V}$ is inside of $\vec{\varsigma}$ and $\bar{\mathcal{V}}$ is outside of $\vec{\varsigma}$), $\xi(\boldsymbol{x}, t; \boldsymbol{p}) = |I(\boldsymbol{x}, t) - \mu(\boldsymbol{x}, t; \boldsymbol{p})|^2$, α reflects the difference of variances between the Gaussian random variables η inside and outside of $\vec{\varsigma}$, and λ associates a cost with the Euclidean length $d\vec{\varsigma}$. Minimization (4) can be interpreted as *volume competition*: the first term measures the compatibility of image point at $(\boldsymbol{x}, t)$ with the overall intensity and motion inside of $\vec{\varsigma}$, whereas the second term measures such compatibility with the outside of $\vec{\varsigma}$. The third term assures that a minimal area (smooth) surface is sought. Thus, the minimization process seeks as smooth a surface as possible that divides $\Omega \times \mathcal{T}$ into such $\mathcal{V}$ and $\bar{\mathcal{V}}$ that each is best explained by its own motion parameters and intensity.

In order to carry out minimization (4), the problem needs to be decomposed into interleaved minimizations with respect to $\vec{\varsigma}$ and $(\boldsymbol{p}, \bar{\boldsymbol{p}})$. Since in this paper we are interested primarily in validating the joint space-time formulation of the video segmentation problem, we consider for now only the simpler case of segmenting a moving object against stationary background. We will address the general case of several moving objects and a possibly moving background, in the future. Under this assumption, we propose, after Jehan-Besson and Barlaud [14], the absolute frame difference $|I(\boldsymbol{x}, t) - I(\boldsymbol{x}, t-1)|$ as the measure of background intensity variation in time $\xi(\boldsymbol{x}, t; \bar{\boldsymbol{p}})$, and a fixed penalty α within the object ($\xi(\boldsymbol{x}, t; \boldsymbol{p})$=1). In order to attain the global minimum in (4), the surface $\vec{\varsigma}$ must assign points $(\boldsymbol{x}, t)$ with small frame difference to its outside ($\bar{\mathcal{V}}$), and those with large difference to its inside ($\mathcal{V}$). The balance between such assignments is controlled by α. Although this model may seem counterintuitive, it has been shown to work well in the 2-D case [14].

4. ESTIMATION OF THE BOUNDARY SURFACE

Since in the simplified case considered no motion parameters need to be estimated, minimization (4) reduces to:

$$\min_{\vec{\varsigma}} \iiint_{\Omega \times \mathcal{T}} h(I_t) d\boldsymbol{x} dt + \lambda \iint_{\mathcal{S}} d\vec{\varsigma}, \tag{5}$$

$$h(I_t) = \begin{cases} \alpha & \text{if } (\boldsymbol{x}, t) \in \mathcal{V}, \\ |I(\boldsymbol{x}, t) - I(\boldsymbol{x}, t-1)| & \text{if } (\boldsymbol{x}, t) \in \bar{\mathcal{V}}. \end{cases}$$

Note that this form is related to (1): the Euclidean area element weight is 1 ($g(\delta I) = 1$), whereas the weighted volume measure

is a discontinuous function $h(\cdot)$ that quantifies the competition between volumes. As we shall see, this will result in an additional force helping avoid local minima and speeding up convergence.

Although at the first glance our new formulation (5) and formulation (1) look very similar, they differ, in fact, significantly. The formulation (1) is edge-based and requires strong edges δI in the data to which the surface $\vec{\varsigma}$"locks". Our new formulation, however, is volume-based and should work well even in the presence of diffuse edges due to the inherent competition between volumes ($\mathcal{V}$ and $\bar{\mathcal{V}}$). This has been nicely illustrated on still-image (2-D) segmentation in a recent paper by Chan and Vese [23]. Clearly, the formulation (1) permits only a detection of change; it leads naturally to volume detection (e.g., detection of static 3-D objects [21], or detection of moving objects in $\boldsymbol{x}-y-t$ space [19]). However, it cannot handle volume segmentation, i.e., explicit division into sub-volumes with different characteristics such as different texture or different motion. The volume-competition formulation (5) can handle two different types of motion, and can be extended to multiple motions in similar way as proposed in [13] for 2-D case.

In order to solve for $\vec{\varsigma}$we apply the Stokes' theorem ($\int_{\mathcal{V}} d\omega = \int_{\partial\mathcal{V}} \omega$) in 3-D, and this leads to the surface evolution equation:

$$\frac{\partial\vec{\varsigma}}{\partial t} = [\alpha - |I(\boldsymbol{x},t) - I(\boldsymbol{x},t-1)| + \lambda\kappa_m]\vec{n}. \qquad (6)$$

Ignoring the curvature, $\alpha > |I(\boldsymbol{x},t) - I(\boldsymbol{x},t-1)$ will result in the surface shrinking and thus relinquishing the point $\vec{\varsigma}(\boldsymbol{x},t)$, while $\alpha < |I(\boldsymbol{x},t) - I(\boldsymbol{x},t-1)$ will cause the surface to expand thus englobing this point. Clearly, there will be a competition between two forces, one related to $\mathcal{V}$ and the other related to $\bar{\mathcal{V}}$, that will claim or relinquish 3-D points on and around the surface $\vec{\varsigma}$, thus the *volume competition* name. As for the curvature κ_m, it plays the role of a smoothing filter with respect to surface point coordinates. For sufficiently large λ, the curvature term will assure simultaneously smooth boundaries of individual-frame segmentations and temporal continuity of such boundaries (no large changes in segment shapes between consecutive frames). This can be viewed as a generalization of tracking of individual-frame segmentations; our approach is based on a sound mathematical model rather than a sequence of ad hoc steps. The inherent space-time continuity of the volume $\mathcal{V}$ is what distinguishes our approach from methods supported on 2 frames only.

The active-surface evolution equation (6) suffers from the same deficiencies as active-contour equations. Therefore, we embed the active surface $\vec{\varsigma}$into a hyper-surface u (in 4-D space) which leads to the following level-set evolution equation:

$$\frac{\partial u}{\partial t} = F\|\nabla u\| = [\alpha - |I(\boldsymbol{x},t) - I(\boldsymbol{x},t-1)| + \lambda\kappa_m]\|\nabla u\|.$$

We implement this equation iteratively using standard discretization as described in [6]. In each iteration we calculate the force F at zero level-set points, extend this force using the fast marching algorithm by solving $\nabla u \cdot \nabla F = 0$ for F, update the surface u, and re-initialize the surface using the fast marching algorithm by solving $\|\nabla u\|$=1 (signed distance).

5. EXPERIMENTAL RESULTS

We have applied the proposed algorithm to the segmentation of several image sequences. Fig. 1 shows results for two sequences: natural-texture, synthetic-motion, test sequence *Bean* in which a "bean"-shaped object undergoes accelerated zoom and rotation, and MPEG-4 test sequence *Akiyo* with small head-and-shoulders motion. In both cases, the segmentation was computed jointly over 30 frames, with α=0.5 and λ=0.01 for *Bean* and α=0.5 and λ=0.1 for *Akiyo*. The above results show that an excellent object-shape recovery and tracking between frames are possible without using intensity edges explicitly. Moreover, in case of the *Bean* sequence the results are very accurate despite significant motion (over 70 pixels, plus zoom-in). Note the clear increase of the size of the "bean" and its rotation. The shape evolves consistently over time despite no explicit tracking. Similarly accurate result has been achieved for *Akiyo* although in this case motion has been much smaller.

6. CONCLUSION

The proposed approach shows clearly promise as a tool for consistent frame-to-frame segmentation of image sequences. The current models are very simple and do not permit moving background (e.g., due to camera motion) or multiple objects. Also, the computational complexity of this approach is significant as we did not incorporate narrow-banding or hierarchical implementation. We are currently addressing these and other issues in order to improve the flexibility and performance of this approach.

7. REFERENCES

[1] E. Mémin and P. Pérez, "Dense estimation and object-based segmentation of the optical flow with robust techniques," *IEEE Trans. Image Process.*, vol. 7, no. 5, pp. 703–719, May 1998.

[2] M.M. Chang, A.M. Tekalp, and M.I. Sezan, "Simultaneous motion estimation and segmentation," *IEEE Trans. Image Process.*, vol. 6, no. 9, pp. 1326–1333, Sept. 1997.

[3] J.Y. Wang and E.H. Adelson, "Representing moving images with layers," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 625–638, 1994.

[4] H.S. Sawhney and S. Ayer, "Compact representations of videos through dominant and multiple motion estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 8, pp. 814–830, Aug. 1996.

[5] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Intern. J. Comput. Vis.*, vol. 1, pp. 321–331, 1988.

[6] J.A. Sethian, *Level Set Methods*, Cambridge University Press, 1996.

[7] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *Intern. J. Comput. Vis.*, vol. 22, no. 1, pp. 61–79, 1997.

[8] N. Paragios and R. Deriche, "Geodesic active contours and level sets for the detection and tracking of moving objects," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 3, pp. 266–280, Mar. 2000.

[9] J. Zhang, J. Gao, and W. Liu, "Image sequence segmentation using 3-D structure tensor and curve evolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 5, pp. 629–641, May 2001.

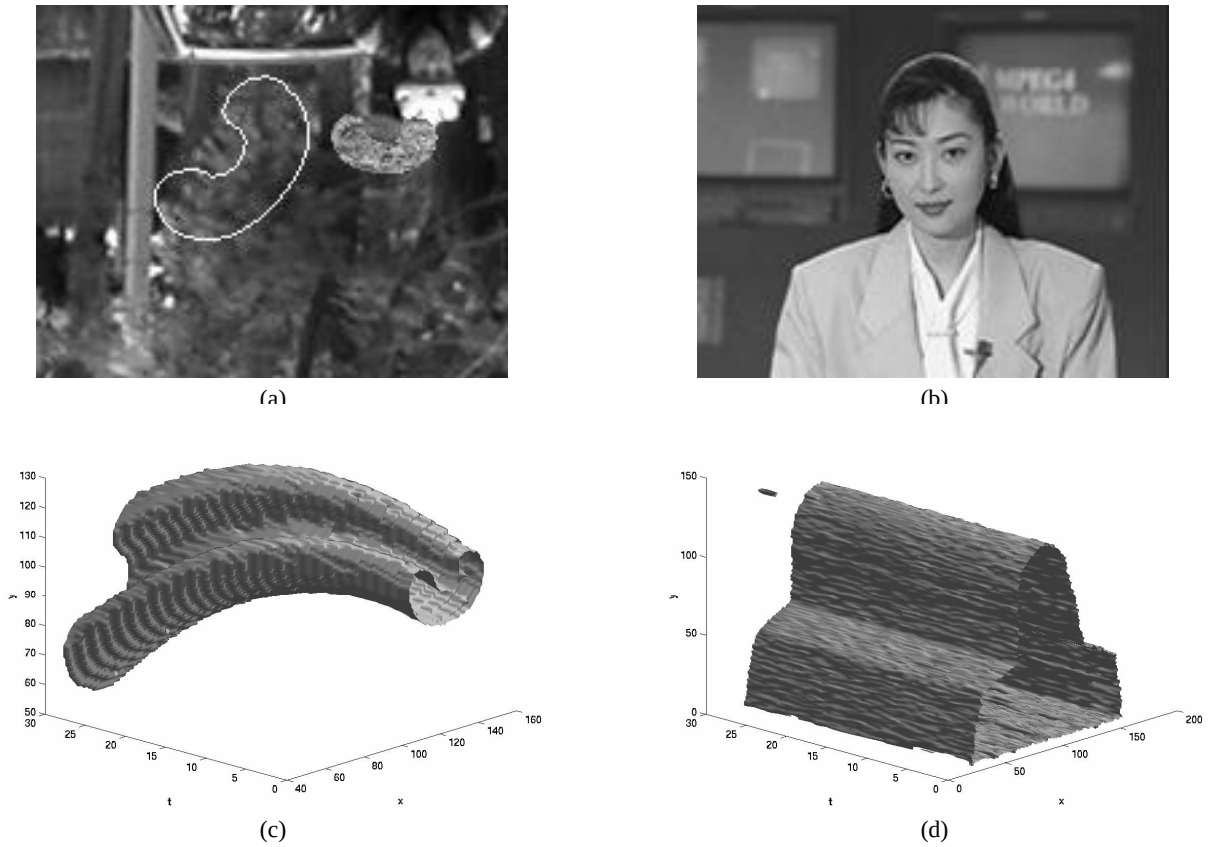

Fig. 1. First frame of (a) natural-texture, synthetic-motion test sequence *Bean* (accelerated zoom and rotation), and (b) MPEG-4 test sequence *Akiyo*, and (c-d)associated joint 30-frame segmentations estimated. The white contour shows the "bean" in frame #30.

[10] G. Kühne, J. Weickert, O. Schuster, and S. Richter, "A tensor-driven active contour model for moving object segmentation," in *Proc. IEEE Int. Conf. Image Processing*, 2001, pp. 73–76.

[11] S.C. Zhu and A. Yuille, "Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 9, pp. 884–900, Sept. 1996.

[12] A.-R. Mansouri and J. Konrad, "Motion segmentation with level sets," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 1999, vol. II, pp. 126–130.

[13] A.-R. Mansouri, B. Sirivong, and J. Konrad, "Multiple motion segmentation with level sets," in *Proc. SPIE Image and Video Communications and Process.*, Jan. 2000, vol. 3974, pp. 584–595.

[14] S. Jehan-Besson, M. Barlaud, and G. Aubert, "Detection and tracking of moving objects using a new level set based method," in *Proc. Int. Conf. Patt. Recog.*, Sept. 2000, pp. 1112–1117.

[15] E. Debreuve, M. Barlaud, G. Aubert, I. Laurette, and J. Darcourt, "Space-time segmentation using level set active contours applied to myocardial gated SPECT," *IEEE Trans. Med. Imag.*, vol. 20, no. 7, pp. 643–659, July 2001.

[16] F. Luthon, A. Caplier, and M. Liévin, "Spatiotemporal MRF approach with application to motion detection and lip segmentation in video sequences," *Signal Process.*, vol. 76, pp. 61–80, 1999.

[17] F.-M. Porikli and Y. Wang, "An unsupervised multi-resolution object extraction algorithm using video-cube," in *Proc. IEEE Int. Conf. Image Processing*, 2001, pp. 359–362.

[18] B. Parker and J. Magarey, "Three-dimensional video segmentation using a variational method," in *Proc. IEEE Int. Conf. Image Processing*, 2001, pp. 765–768.

[19] R. El-Feghali, A. Mitiche, and A.-R. Mansouri, "Tracking as motion boundary detection in spatio-temporal space," in *Int. Conf. Imaging Science, Systems, and Technology*, June 2001, pp. 600–604.

[20] R. Malladi, J.A. Sethian, and B.C. Vemuri, "Shape modeling with front propagation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 2, pp. 158–176, Feb. 1995.

[21] V. Caselles, R. Kimmel, G. Sapiro, and C. Sbert, "Minimal surfaces based object segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 4, pp. 394–398, Apr. 1997.

[22] J. Konrad and C. Stiller, "On Gibbs-Markov models for motion computation," in *Video Compression for Multimedia Computing – Statistically Based and Biologically Inspired Techniques*, H.H. Li, S. Sun, and H. Derin, Eds., chapter 4, pp. 121–154. Kluwer Academic Publishers, 1997.

[23] T.F. Chan and L.A. Vese, "Active contours without edges," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, Feb. 2001.