Motion Segmentation Based on Perceptual Organization of Spatio-Temporal Volumes

Kishore Korimilli and Sudeep Sarkar
Computer Science and Engineering
University of South Florida, Tampa, FL 33620
Email: sarkar@csee.usf.edu

Abstract

The role of perceptual organization in motion analysis has heretofore been minimal. In this work we demonstrate that the use of perceptual organization principles of temporal coherence (common fate) and spatial proximity can result in a robust motion segmentation algorithm that is able to handle drastic illumination changes, occlusion events, and multiple moving objects, without the use of object models. The adopted algorithm does not employ the traditional frame by frame motion analysis, but rather treats the image sequence as a single 3D spatio-temporal block of data. We describe motion using spatio-temporal surfaces, which we, in turn, describe as compositions of finite planar patches. These planar patches, referred to as temporal envelopes, capture the local nature of the motions. We detect these temporal envelopes using 3D-edge detection followed by Hough transform, and represent them with convex hulls. We present a graph-based method to group these temporal envelopes arising from one object based on Gestalt organizational principles. A probabilistic Bayesian network quantifies the saliencies of the relationships between temporal envelopes. We present results on sequences with multiple moving persons, significant occlusions, and scene illumination changes.

1 Introduction

Segmenting moving objects in image sequences is among the most challenging problems in image sequence analysis and is a necessary precursor to any motion interpretation algorithm, such as gait recognition, intruder identification, or model-based tracking. The most common approach to motion segmentation relies on frame by frame image differencing [3, 4], which has been found to be sufficient in well engineered, controlled settings. However, the differencing strategy breaks down in the presence of illumination changes or noisy background motion clutter, such as that present in fluttering leaves of a tree or in rain.

Another common approach to motion segmentation uses optic flow estimates, which are usually based on local spatial and temporal information. Typical approaches aggregate the individual flow elements into regions of coherent motion [1]. Alternatively, frame based optic flow vectors are stitched together to obtain motion traces, which are then grouped based on geometric characteristics [2]. However, the local myopic nature of the information necessarily results in noisy optic flow estimates and encounters problems in the presence of occlusions. To overcome the local noisy nature of the point based flow estimates one might opt for motion estimation from extended features [5, 11, 7, 6]. However, the success of these approaches relies on the stability of detection of such features over multiple frames and the ability to effectively solve the correspondence problem.

We show that the use of perceptual organization principles of proximity and temporal coherence (common fate) to group features in the spatio-temporal volumes robustly segments moving objects in image sequences. Unlike traditional frame by frame analysis or analysis over small number (5-6) of frames, we consider a spatio-temporal block consisting of many (> 20) images that are closely sampled, temporally. Features on a moving object sweeps spatio-temporal surfaces in this volume, which exhibit significant amount of organization and structure that is very different from the surfaces due to the background features. Although we are not the first ones to suggest the use of spatio-temporal volumes for motion analysis (see Bolles and Baker [9], Jain and Liou [8], Ricquebourg and Bouthemy [12], Niyogi and Adelson [13]), this work represents one of few that exploits Gestaltic principles of perceptual organization for motion analysis. So far, the role of perceptual organization has been restricted mainly to ob-

[0]This work was supported in part by the US National Science Foundation grants IIS-9907141 and IRI-9501932.

0-7695-0750-6/00 $10.00 © 2000 IEEE

ject recognition from 2D images [1]. Shi and Malik [10] present another framework for spatio-temporal grouping using Gestalt principles, which unlike our use of temporal coherence over the whole spatio-temporal volume uses coherence over few frames at a time. The role of temporal coherence (common fate) in grouping is greater in our framework.

We show that even with fairly simple use of perceptual organizational principles we can achieve good motion segmentation in the presence of occlusion, noise, and illumination changes, without the use of object models.

2 The Approach

2.1 3D edge detection and filtering

The grouping process starts with 3D edge detection in the spatio-temporal volume, $I(x, y, t)$, based on a 3D extension of the 2D Canny edge detector. We detect single pixel width edge surfaces by using the 3D extension of the 2D non-maxima suppression scheme of Canny on the 3D gradient estimates.

Based on the edge orientation estimates, we then filter the *nearly* static background features and the features that arise due to illumination changes. Let the angle that the local 3D gradient direction makes with the time axis be denoted by θ_t. This angle is zero when there is no motion in time and is 90^o when there is scene illumination change across frames. Fig. 1(a) shows an XT slice of an image sequence with no change in the illumination and Fig. 1(b) shows the XT slice of an image sequence with changing illumination. Note that the XT slice in Fig. 1(a) contains lines predominantly parallel to the time axis whereas the XT slice in Fig. 1(b) contains lines that are parallel and perpendicular to the time axis, the perpendicular lines are due to changes in illumination.

Thus, we remove background and illumination artifacts by filtering all pixels whose value of θ_t satisfies either of the two conditions: $0 \leq \sin\theta_t \leq T_a$ or $T_p \leq \sin\theta_t \leq 1$, where T_a and T_p are the threshold values for filtering of pixels oriented along the time-axis and perpendicular to the time-axis, respectively.

2.2 Temporal envelopes

Each image feature that is undergoing motion will sweep out a surface in the spatio-temporal volume. For instance, a point on an object that is moving at constant velocity will sweep out a straight line in the spatio-temporal volume and a straight boundary of an object will sweep out a plane in the spatio-temporal space. In general, the shape of the spatio-temporal surface will be complex. Instead of a mathematical specification of this spatio-temporal surface, we opt to describe this surface as collection of *planar spatio-temporal patches*, which we call **temporal envelopes**. We detect these temporal envelopes using the Hough transform.

[1] At the recent Workshops on Perceptual Organization in Computer Vision – 1998 and 1999, participants recognized that principles of perceptual organization need not be restricted to just object recognition from 2D images, but also exploited for motion analysis.

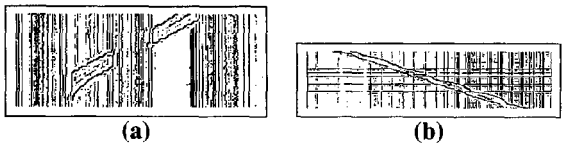

Figure 1. (a) XT slice for a normal (constant illumination) scene. (b) XT slice for a changing illumination scene. Time axis is along the vertical direction.

2.3 Detection of Temporal Envelopes

We set up the Hough transform space based on the Hessian normal form of the plane equation: $x\cos\theta_x + y\cos\theta_y + z\cos\theta_t = p$, where θ_x, θ_y, and θ_t are the angles that the surface normal makes with the three axes and p is the perpendicular distance of the plane from the origin. We estimate these angles from the computed image gradient direction, which would be along the normal to the plane. Thus, $\cos\theta_x = \frac{I_x}{|\nabla I|}$, $\cos\theta_y = \frac{I_y}{|\nabla I|}$, and $\cos\theta_t = \frac{I_t}{|\nabla I|}$, where I_x, I_y and I_t are the partial derivatives of the 3D spatio-temporal function and ∇I is the gradient. From the equation, it appears that we have four parameters for a plane. However, not all angles are independent of each other. We merge two of the angles to arrive at three independent parameters: $(\theta_t, \theta_{xy}, p)$, where $\theta_{xy} = \tan^{-1}(\frac{\cos\theta_y}{\cos\theta_x})$. Each edge pixel votes for a single point in this quantized 3D Hough space, thus eliminating the possibility of false peaks. We find local maxima by considering local peaks over a 3D-neighborhood windows. Once a local maximum is found, it is recorded and all other entries in the window are marked as ineligible to be detected as local maxima in the next iteration. This eliminates the possibility of detecting noisy peaks near the main peak.

The Hough transform fits *infinite* planes to the edge points. We arrive at *finite* planar patches – the *temporal envelopes* – by considering the 2D convex hull of the projection of edge points that voted for that plane onto the infinite plane.

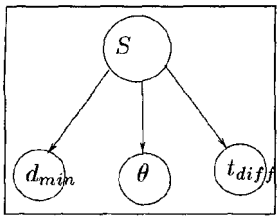

Figure 2. Bayesian network used to classify pairs of plane primitives.

2.4 Grouping of Temporal Envelopes

A moving object can result in one or more planes in the Hough space. Each detected plane in the Hough space results in a temporal envelope. A scene containing multiple moving objects would thus result in a collection of temporal envelopes in the spatio-temporal volume, with each object in motion giving rise to more one or more temporal envelopes. The problem is to group the temporal envelopes from one object, for which we use the perceptual organizational principles of proximity, continuity and parallelism, the latter being a form of the Gestalt principle of common fate. Temporal envelopes from a single moving object will tend to be close together, and mostly locally parallel and continuous.

The grouping process starts with the construction of a Gestalt relationship graph, whose nodes are the temporal envelopes and the links denote the existence of salient Gestalt relationships. We term this graph as the scene structure graph. We employ Bayesian networks to quantify and classify the relation between two temporal envelopes into two classes: salient ($S = 1$) or not salient ($S = 0$).

We classify the relation between two temporal envelopes into $S = 1$ or 0 based on the following attributes. (i) The ratio of the minimum and maximum distances between two temporal envelopes, d_{min}. (ii) The angle between the envelopes, normalized by π, which we denote by θ. (iii) The temporal intercept of the temporal envelope, normalized by the maximum distance between the envelopes, which we denote by t_{int}. All the three attributes are normalized to range from 0 to 1.

Using these attributes we construct the Bayesian Network [2] shown in Fig. 2, to classify pairs of plane primitives as being salient ($S = 1$) or not salient ($S = 0$). Apart from the structure of the Bayesian network, we need to specify the probabilities that go along with the nodes. For a node with no parents, we need to specify a prior distribution. And for variables with parents, viz. d_{min}, θ, and t_{diff}, we need conditional distributions. We assume that in the absence of contrary evidence, it is equally likely that the relation between a random pair of temporal primitives is salient or not salient. As for the conditional probabilities, we need to specify the probability of an attribute given the state of its parent. For example, for the relational attribute t_{diff} we need to specify $P(t_{diff} = t|S = s)$. Thus, the probability distribution for t_{diff} is specified by $P(t_{diff} = t|S = 1)$ and $P(t_{diff} = t|S = 0)$. For a salient relation the ideal value of t_{diff} should be zero. We can represent such a distribution using the right-triangular function, $Tn(x, b) = 2(b - x)/b^2$ over $x = (0, b)$. The probability is maximum when x is equal to zero. Since, $S = 0$ represents a completely random scenario, we choose $P(t_{diff} = t|S = 0)$ to be a uniform distribution over (0, 1). We specify all the conditional probabilities as follows:

$$\begin{array}{rcl} P(d_{min}|S=1) & = & Tn(d_{tol}) \\ P(d_{min}|S=0) & = & U(0,1) \\ P(\theta|S=1) & = & Tn(\theta_{tol}) \\ P(\theta|S=0) & = & U(0,1) \\ P(t_{diff}|S=1) & = & Tn(t_{tol}) \\ P(t_{diff}|S=0) & = & U(0,1) \end{array} \quad (1)$$

All the conditional probabilities involve three parameters: d_{tol}, θ_{tol}, and t_{tol}. These parameters represent the effective tolerances in the grouping parameters. Thus, d_{tol} is the distance tolerance, θ_{tol} is the angle tolerance, and t_{tol} is the time-intercept tolerance.

We base the grouping of the temporal envelopes on the probability that the relationship between them is salient, given the variables (evidence) d_{min}, θ and t_{diff}. We compute this probability by passing probabilistic messages in the Bayesian network.

The nodes, representing the temporal envelopes, in the Gestalt graph are connected if the value of the probability $P(S = 1|d_{min}, \theta, t_{diff})$ is greater than $P(S = 0|d_{min}, \theta, t_{diff})$. After quantifying the relationships between all pairs of temporal envelopes, we identify the groupings of temporal envelopes by the connected components of the Gestalt graph.

2.5 Spatial Envelopes

Grouping of the temporal envelopes will result in grouping of the underlying edge pixels. For a time frame t_r, we define the *spatial envelope* as the intersection of the plane

[2] Bayesian networks are directed acyclic graphs, whose nodes represent variables of interest, and edges represent dependence among these variables. They are graphical representations of joint probability distributions. To quantify the strengths of these dependencies, each node is associated with a conditional-probability that captures the relationships among that node and its parents. The most distinctive characteristic of Bayesian networks is their ability to faithfully represent causal relationships and to adapt to changing conditions by updating the probability measures attached with the nodes. The causal information encoded in the Bayesian network facilitates the process of grouping the plane primitives effectively.

$t = t_r$ with the 3D convex hull of the grouped edge pixels, which belong to one object in motion. This intersection, which is also convex shaped, represents the spatial envelope of that object in motion for that particular time frame.

3 Results

We have successfully applied the algorithm on a variety of image sequences, both synthetic and real. Here we present results on three sequences, with multiple motion, significant occlusion, and severe illumination changes. We would like to point out that motion segmentation strategies built around image differencing would fail on these sequences because of the presence of intensity changes due to other causes than just motion. Because of space limitations, we do not show results demonstrating the viability of the algorithm in the presence of noise.

3.1 Six persons

We consider a complicated scene with six moving persons: three persons walking to the left and three persons walking to the right, with different but overlapping time intervals. The entire sequence consists of 280 frames, with two sample frames shown in Figs. 3(a) and (b). We also shown an XT-slice through the 3D edges *after* the removal of the background and the illumination-change edges in Fig. 3(e). Note the complicated nature of the interaction between the spatio-temporal traces of the different persons. There are 8 occlusion events over the whole sequence. The algorithm detected 14 temporal envelopes, which were then grouped into six groups: two groups had one temporal envelope each, three groups had two temporal envelopes each, and two groups had three temporal envelopes each. Fig. 3(f) shows the spatial envelopes for the 130-th image frame, when there were four persons in the image. To show the relationships between the different groups of temporal envelopes we display one temporal envelope from each group along with some spatial envelopes in Fig. 3(g). Note how we are able to easily segment out the trajectories of the different persons.

3.2 Temporally, sparsely sampled sequence with significant occlusion

The two frames shown in Fig. 4 are from a sequence of two persons walking to the left and one person walking to the right. This sequence is challenging for three reasons: First, this is a temporally sparsely sampled sequence with just 48 frames. Second, the persons are partially occluded behind the stairs for half the number of sequences. Only parts of them were visible during these frames. Third, when

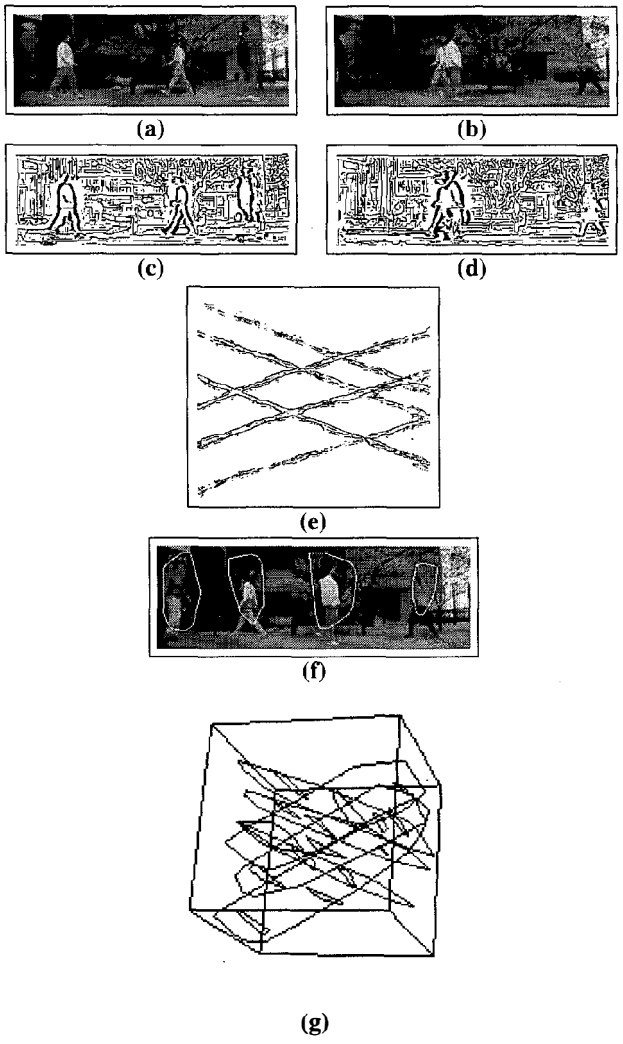

Figure 3. Results on an image sequences of 6 persons over 280 frames. The 92-nd and the 118-th image frames are shown in (a) and (b), with the XY-slices through the 3D edges corresponding to those frames shown in (c) and (d). A XT-slice through the 3D edges *after* the removal of the background and the illumination-change edges is shown in (e), with the time axis is along the vertical direction. (f) The detected spatial envelopes overlaid on the original gray scale images at the 130-th image frame. (g) One temporal envelope from each group along with some spatial envelopes embedded in the spatio-temporal volume.

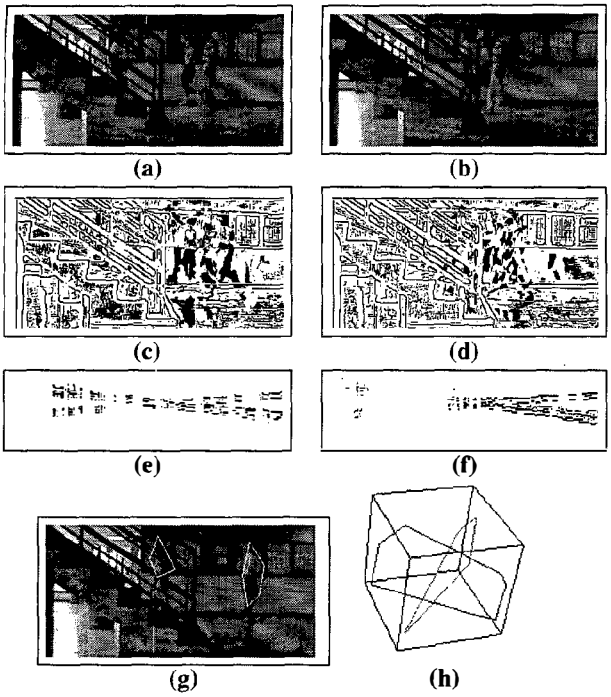

Figure 4. Results on an image sequence with severe occlusion. The 25-th and the 27=th image frames are shown in (a) and (b), which XY-slices through the 3D edges corresponding to those frame shown in (c) and (d), respectively. Two XT-slice through the 3D edges are shown in (e) and (f), with the time axis along the vertical direction. (g) Spatial envelopes overlaid on the original gray scale images at time frame=31. (h) One spatio-temporal envelope from each group.

the persons were behind the staircase, the local image structure around the railings of the stairs changed, thus changing the detected edges. This change in edge structure is not directly due to motion but due to local image structure changes.

Fig. 4(e) and (f) show the XT-slices of the 3D edge images. Note that, since the persons moving to the left were close together (walking side-by-side) in all the frames their traces were combined into a single path in the XT-slice. The algorithm detects only one plane for the two persons walking to the left and one plane for the persons walking to the right. This is because of the fact that the persons walking to the left were so close to each other that all the edge pixels corresponding to both the persons vote for a single plane; this a case when the algorithm fails to resolve the individual persons. Two spatial envelopes are shown in Fig. 4(g). The two temporal envelopes are shown in the Fig. 4(h). Inspite of severe occlusion conditions, we are able to maintain track over the whole time sequence.

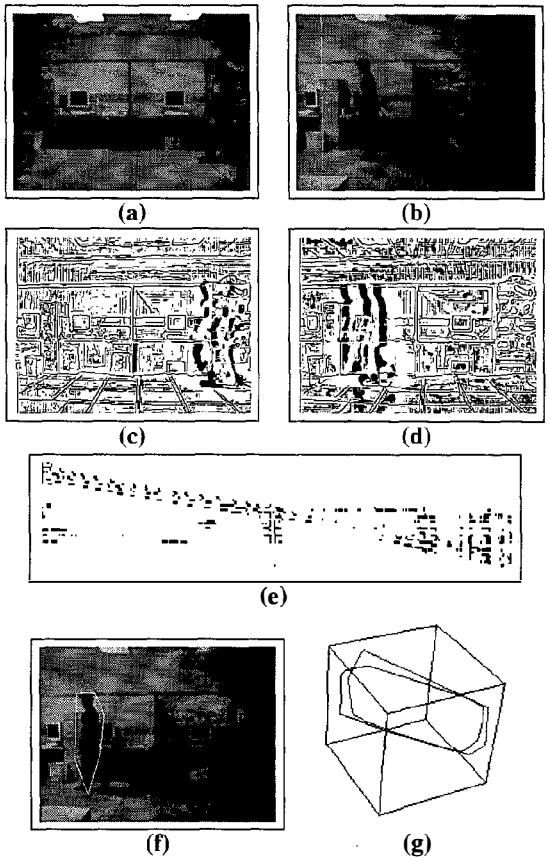

Figure 5. Results on a image sequence with scene illumination change. Two image frame are shown in (a) and (b), with the XY-slice through the 3D edges corresponding to these frame shown in (c) and (d). A XT-slice through the 3D edges is shown in (e). (f) Spatial envelopes overlaid on the original gray scale images. (g) Detected temporal envelopes.

3.3 Sequence with change in scene illumination

In the last set of experiments presented here, we consider an image sequence of 49 frames of a person walking from right to left, with changing illumination conditions. Some lights in the room were randomly turned on and off

during the motion capture. Two sample frames are shown in Figs. 5(a) and (b). It is also worth mentioning that the screen of the left computer monitor was not blank. There was a screen saver program that created changing patterns, thus representing a different kind of changing background clutter. The effect of the illumination changes is evident on the XT-slice shown in Fig. 5(e), which shows the edges after filtering the background and illumination-changes. Note that significant clutter does survive the filtering processes, however, the clutter does not form any organized structure. As a consequence, the grouping process is able to discard the clutter and we get two temporal envelopes for the person as shown in Fig. 5 (g). Fig. 5 (f) shows a spatial envelope overlaid on the original gray scale image at time frame 31. This demonstrates the ability of the algorithm to segment moving objects in the presence of illumination changes.

3.4 Effect of the Input Parameters

For all the image sequences we used an edge detector scale of 1.6. The quantization of the spatial parameter constituting the Hough space was 1 pixel and the angle quantization was 3^0. Coarse quantization of the Hough space will result in "bloating" of the spatial envelopes as more and more of background pixels participate in a Hough space maxima. The window size for peak detection in the Hough space ranged from 20 to 40 units along each dimension. The background filtering threshold ranged from 3^o to 4^o. However, the range for the illumination threshold was greater – it ranged from 5^o for sequences with no illumination change to about 16^o for sequences with drastic illumination changes. Drastic illumination changes from frame to frame, results in noisy estimates for the gradient directions, hence the need for a wider range of threshold values.

4 Conclusions

We presented a framework for segmenting moving objects in an image sequence, based on the coherence in the spatio-temporal volume. The use of perceptual organization principles renders the segmentation procedure robust and helps grouping in the presence of motion occlusions. This is one of the few applications of perceptual grouping in motion analysis. An essential property of the presented algorithm is that no a priori knowledge about the structure of the objects is required. Our framework includes a 3D edge detection algorithm, a voting based algorithm to determine the temporal envelope primitives of the moving object, and a grouping algorithm to group the temporal primitives of an object. The framework, although simple, is sufficiently powerful as demonstrated by the results on real images with severe conditions. We obtain good results on scenes with several persons, occlusions, noise, and changing illumination.

References

[1] Y. Yacob and M. J. Black, "Parameterized Modeling and Recognition of Activities," *Computer Vision and Image Understanding*, vol. 73, no. 2, pp. 232–247, Feb. 1999.

[2] M. Shah, K. Rangarajan, and P. S. Tsai, "Motion Trajectories," *IEEE Transactions on Systems Man and Cybernetics*, vol. 23, no. 4, pp. 1138–1150, Jul-Aug. 1993.

[3] R. Jain and H. H. Nagel, "On the Analysis of Accumulative Difference Pictures from Image Sequences of Real World Scenes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 1, pp. 206-214, Apr. 1979.

[4] S. N. Jayaramamurthy and R. Jain, "An Approach to the Segmentation of Textured Dynamic Scenes," *Computer Vision, Graphics and Image Processing*, vol. 21, no. 2, Feb 1983, pp. 239-261.

[5] S. T. Barnard and W. B. Thompson, "Disparity Analysis of Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, no. 4, July 1980, pp. 333-340.

[6] P. Bouthemy and M. Francois, "Tracking Complex Primitives in an Image Sequence," *Proc. of 12th IAPR International Conference on Pattern Recognition*, pp. 426-431, 1994.

[7] F. Dufaux, F. Moscheni and A. Lippman, "Spatio-Temporal Segmentation based on Motion and Static Segmentation," *Proc. of the IEEE International Conference on Image Processing*, vol. 1, pp. 306-309, 1995.

[8] R. Jain and S. P. Liou, "Qualitative Motion Analysis Using a Spatio-Temporal Approach," *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 726-727, 1991.

[9] C. R. Bolles and H. H. Baker, "Epipolar-Plane Image Analysis: A Technique for Analysing Motion Sequences," *International Journal on Computer Vision*, pp. 26-36, 1989.

[10] J. Shi and J. Malik, " Normalized Cuts and Image Segmentation," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico*, pp. 731-737, June 1997.

[11] H. Gu, Y. Shirai and M. Asada, "MDL-Based Segmentation and Motion Modeling in Long Image Sequence of Scene with Multiple Independently Moving Objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 58-64, Jan 1996.

[12] Y. Ricquebourg and P. Bouthemy, "Tracking of Articulated Structures Exploiting Spatio-Temporal Image Slices," *Proc. of the IEEE International Conference on Image Processing*, vol. 3, pp. 480-482, 1997.

[13] A. Niyogi and E. Adelson, "Analyzing and Recognizing Walking Figures in XYT," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 469-474, Dec 1997.

[14] J. L. Bentley, K. L. Clarkson and B. D. Levine, "Fast Linear Expected-Time Algorithms for Computing Maxima and Convex Hulls," *Proc. of 1st Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 56-62, 1993.

[15] J. O. Rourke, "Computational Geometry in C," Cambridge University Press, 1995.