Perceptual Organization Based Computational Model for Robust Segmentation of Moving Objects¹

Sudeep Sarkar, Daniel Majchrzak, and Kishore Korimilli

Computer Science and Engineering, University of South Florida, Tampa, Florida 33620

E-mail: sarkar@csee.usf.edu, majchrza@csee.usf.edu

Received December 14, 1999; accepted February 25, 2002

The role of perceptual organization in motion analysis has heretofore been minimal. In this work we present a simple but powerful computational model and associated algorithms based on the use of perceptual organizational principles, such as temporal coherence (or common fate) and spatial proximity, for motion segmentation. The computational model does not use the traditional frame by frame motion analysis; rather it treats an image sequence as a single 3D spatio-temporal volume. It endeavors to find organizations in this volume of data over three levels—signal, primitive, and structural. The signal level is concerned with detecting individual image pixels that are probably part of a moving object. The primitive level groups these individual pixels into planar patches, which we call the temporal envelopes. Compositions of these temporal envelopes describe the spatio-temporal surfaces that result from object motion. At the structural level, we detect these compositions of temporal envelopes by utilizing the structure and organization among them. The algorithms employed to realize the computational model include 3D edge detection, Hough transformation, and graph based methods to group the temporal envelopes based on Gestalt principles. The significance of the Gestalt relationships between any two temporal envelopes is expressed in probabilistic terms. One of the attractive features of the adopted algorithm is that it does not require the detection of special 2D features or the tracking of these features across frames. We demonstrate that even with simple grouping strategies, we can easily handle drastic illumination changes, occlusion events, and multiple moving objects, without the use of training and specific object or illumination models. We present results on a large variety of motion sequences to demonstrate this robustness. © 2002 Elsevier Science (USA)

Key Words: perceptual organization; motion segmentation; Hough transform; spatio-temporal volume analysis; Bayesian networks; grouping.

¹ This work was supported by the National Science Foundation CAREER Grants IRI-9501932 and IIS-9907141. A short version of this work was presented at the International Conference on Pattern Recognition, 2000.

1. INTRODUCTION

Motion segmentation is the identification of image regions or collection of image features that exhibit independent motion. This problem is among the most challenging ones in image sequence analysis and is a precursor to most motion interpretation algorithms such as gait recognition, intruder identification, or model based tracking. Even an imperfect partial segmentation would be a useful front-end to many motion analysis strategies. It is desirable that this segmentation process be as generic as possible to be applicable across a variety of domains.

The most common approach to motion segmentation relies on image differencing, which has been found to be sufficient in well-engineered and controlled settings. However, differencing strategies, even those that rely on explicit estimation of the background, break down in the presence of (unmodeled) illumination changes, occlusions, or noisy background motion clutter. Another common approach to motion segmentation aggregates optic flow vectors into regions of coherent motion [1]. Alternatively, frame based optic flow vectors are stitched together to obtain motion traces, which are then grouped into motion patterns [32]. However, the local myopic nature of the optic flow computations renders the strategy sensitive to noise, illumination changes, and occlusion events. To overcome this local nature of point based flow estimates one might opt for motion estimation of extended features; however, the success of such approaches relies on the stability of extraction of the tracked features over frames. In addition, the existence of a good tracking strategy is required to establish feature correspondences across frames.

In this paper, we propose a computational model based on the use of perceptual organizational principles for motion segmentation. We demonstrate that the use of perceptual organizational principles, such as proximity and temporal coherence, in the spatio-temporal domain allows us to robustly segment moving objects in images sequences. Unlike traditional frame-by-frame analysis or analysis over a small number (five or six) of frames, we consider a spatio-temporal block consisting of many (> 20) images that are closely sampled, temporally (30 frames per second). Frame-by-frame analysis has been the dominant mode in motion studies. Present day desktop computers are sufficiently fast and have enough memory to undertake motion analysis based on spatio-temporal volumes, which generally exhibit more structure and organization than do individual 2D frames.

Features on a moving object sweep spatio-temporal surfaces in this volume exhibiting a significant amount of organization and structure that is very different from the spatio-temporal surfaces from background features. Although we are not the first ones to suggest the use of spatio-temporal volumes for motion analysis [12], this work demonstrates the power of the use of the Gestalt principles of perceptual organization for spatio-temporal analysis. Principles of perceptual organization have been traditionally used mostly in the context of object recognition from 2D static images. Here we extend the use of perceptual organization to motion analysis. 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.

Spatio-temporal surfaces can be as complex as the motion models considered. In this work we seek to describe the motion in terms of compositions of planar patches in the spatio-temporal domain. Of course, such descriptions would be accurate for purely constant fronto-parallel motion and would be only first order approximations for other kinds of motion models, such as accelerating or motion along curved trajectories. We, however,

show that the developed grouping strategy does not fail drastically when the motion fails to be purely fronto-parallel or if the camera moves slowly; it produces good segmentations in such cases too.

In the next section we review relevant work. In Section 3, we introduce the notions of spatial and temporal envelopes. There we also describe, in detail, the computational model and the underlying grouping algorithms. We present extensive results on challenging, real sequences in Section 4. We conclude with Section 5.

2. REVIEW OF RELATED RESEARCH

In this section we first review some of the salient approaches to the problem of motion segmentation. In particular, we consider the ones that treat a motion sequence as a spatio-temporal volume. This is not meant to be an exhaustive review but just to provide a sketch of the current major strategies toward motion segmentation. We follow this with a snapshot of the work in perceptual organization that highlights the rarity of perceptual organization work in motion analysis.

2.1. Frame-by-Frame Analysis

Frame-by-frame analysis includes those strategies that use information over a small number (<5 , typically 2) of frames. There are basically three different types of strategies for frame-by-frame analysis based motion segmentation. The first one relies on image differencing. The second relies on optic flow estimates. And the third uses extended 2D image features such as image patches or intensity edges. We next look at some of the work in these three categories.

Motion segmentation using image differencing. Jain and Nagel [4] used the interframe difference operators to estimate moving objects. Their approach exploits the systematic changes in the first- and second-order difference pictures with increasing number of frames to estimate the regions from one or more moving objects, as well as the stationary background. The technique suffers from the disadvantage that the object must be displaced completely from its position in the reference frame. Also, the occlusion problem cannot be handled. Jayaramamurthy and Jain [5] developed an algorithm to extract the masks of moving textured objects against a textured background. Their approach consists of three stages: extraction of the active regions, estimation of the motion parameters of the active regions, and estimation of the object masks. Good experimental results are obtained for random dot stereograms.

In general, image differencing techniques are good at detecting temporal changes in controlled situations where we have good control over illumination conditions so as to reduce the effect of shadow and secondary mutual illumination effects. The technique is also susceptible to noise.

In addition to techniques based on adjacent frame differences, there are algorithms that use explicit models of the background and use it to identify moving objects in each frame. Some algorithms use the mean and covariance estimates of background [64] or a mixture of a Gaussian model [65], normalized block correlation [66], temporal derivatives [67], Bayesian detection [68], or eigenbackgrounds [69]. All of these background maintenance strategies were compared recently with a sophisticated background maintenance algorithm

called Wallflower [63], which performed better. The Wallflower algorithm consists of three parts: a one-step Wiener prediction filter, difference images, and region growing are used to segment background and foreground pixels. In the Results section we show the performance of (our implementation of) the Wallflower on some of our sequences.

Motion segmentation using optic flow. Another common approach to motion segmentation aggregates optic flow vectors into regions of coherent motion [1, 3]. Alternatively, frame based optic flow vectors are stitched together to obtain motion traces, which are then grouped into motion patterns [32]. These strategies also suffer from sensitivities to noise, illumination changes, and occlusion events, although recently, there have been strategies based on recursive filtering to handle problems with occlusion and small camera motion [2]. Debrunner and Ahuja [13] present a method that uses trajectories from long image sequences to estimate 3D structure and motion for rotating objects. The trajectories are detected essentially by frame-by-frame analysis. Tomasi and Kanade [14] present a more general technique to infer motion parameters from these trajectories under the orthographic projection assumptions.

Motion segmentation using extended features. By establishing correspondences of spatial image features, such as interest points or region patches, across several frames of an image sequence, token matching techniques estimate sparse flow field. Methods differ in the techniques they employ to match the selected features. The success of a token matching approach depends on whether the approach can reliably and consistently find and match the set of feature points across multiple frames. Strategies to solve this difficult correspondence problem include relaxation labeling [6], minimum description length [15], Kalman filter based tracking [9], Markov random fields based labeling [8], and likelihood-ratio based region tracking [7]. Some of the factors that hinder generality of this approach include: (i) lack of good definitions for low level domain-independent features that can be detected with low false alarm rates, and (ii) presence of scene occlusions and the consequent uncovering of new background image patches as the occlusion boundary moves.

2.2. Spatio-temporal Volume Analysis

Bolles and Baker [11, 12] were the first to suggest using spatio-temporal volumes to extract motion parameters of a camera, moving in a straight line, using epipolar plane based analysis. They showed how motion parameters can be extracted from spatio-temporal surfaces. Motion segmentation was not the primary focus of their work.

Jain and Liou [10] extended region growing techniques to segment spatio-temporal volumes. Each segmented volume corresponded to a moving object or a stationary object. The region growing process used the concept of α -partitioning, which generated volume hypotheses by taking advantage of the locations of discontinuities in the image solid. This was followed by volume filtering that verified the smoothness variation of the volume hypothesis to reject gross errors.

Ricquebourg and Bouthemy [16] developed an approach to tackle articulated motion, especially applied to the human motion case, by exploiting the spatio-temporal image slices. They observed that articulated motions reveal significant characteristic tracks or “signatures” in appropriately selected XT slices of the image sequence volume XYT. From these appropriately selected XT slices they reconstructed the motion trajectories by fitting a Kalman filtering model. The model was made to adapt to maneuvering targets as well

TABLE 1
Classification of Prior Work in Perceptual Organization of Static
2D Images and of Motion Sequences

	2D Images	2D and time
Assembly level	[22–24, 43]	[25]
Structural level	[21, 26–31, 34–43]	[32, 33, this work]
Primitive level	[26, 44, 45]	[46–48, this work]
Signal level	[44, 49–55, 57–61]	[55, 56, this work]

as to canonical constant speed or constant acceleration motions. It also correctly handled occlusion by producing predicated states of the tracked entities. Good results were presented.

Niyogi and Adelson [17] proposed a method for analyzing walking figures in a spatio-temporal XYT volume by looking at the XT slices. They found that sequences of walking people have different signatures (patterns) in XYT and devised methods for identifying and analyzing these patterns. Snakes, which are smooth splines seeking positions on the image with high energy, were used to fit these patterns in the XT slices and the resulting fits were used to distinguish one walker from another. The approach assumes that the heights of the walkers are known and it requires individuals to walk frontoparallel to the cameras.

2.3. *Perceptual Organization of Motion Sequences*

Work in perceptual organization differ from one another with respect to the types of constituent features that are grouped and the dimensions over which the organizations are sought. In [20], Sarkar and Boyer used these two dimensions to classify work in perceptual organization and to provide a uniform nomenclature. The dimensions over which organization is sought can be 2D, 3D, 2D + time, and 3D + time. In each of the cases, the features that are classified can be of four different classes, signal, primitive, structural, and assembly, which define the different levels of organization. The signal level pertains to organizing the raw signal pixels or voxels. The primitive level deals with organizaing signal level features into extended features that are lower dimensional manifestations in the organizing field, e.g., constant curvature segments in images, and temporal envelopes in spatio-temporal analysis. At the structural level the organized features have the same dimensionality as that of the space in which they are being organized. Ribbons and closed regions are 2D manifestations in 2D and so are structural level features for 2D organization. The assembly level is concerned with further organizing the structural level features. Organizations such as parallel sets of ribbons or boxes constitute the assembly level for 2D grouping. Table 1 lists the prior work in each category. We observe that most of the work in perceptual organization has been with 2D images at the signal, primitive, and structural levels, with increased emphasis at the structural level over the past few years. Perceptual organization work in 2D and motion has not yet seriously ventured past the signal and primitive levels. The work presented here represents organizations at the signal, primitive, and structural levels.

3. THE COMPUTATIONAL MODEL AND ALGORITHMS

The computational model of the grouping process that performs motion segmentation spans three levels of abstractions: signal, primitive, and structural levels (see Fig. 1a). At

FIG. 1. Computational model of the grouping process and the associated flow of algorithms.

the signal level, we find candidate points in the spatio-temporal volume that are located on moving objects. At the primitive level, we find locally planar patches in the spatio-temporal volume, which we term the *temporal envelopes*, constituted by the signal level moving points. And at the structural level, we group these temporal envelopes to form the motion segmentations. As we shall see, this hierarchical arrangement of grouping processes allows us to handle noise in a robust manner. Higher levels capture the structure and organization in the lower levels. The presence of small errors and noise among the features at each level does not drastically change the structure exhibited by them. Thus, small errors in the detection of the signal level moving points should not drastically alter the detected temporal envelopes. Similarly, minor errors and variations in the primitive level outputs do change their overall structure amongst them, which the structural level grouping process extracts. We next describe the algorithms (see Fig. 1b) that are associated with these computational levels.

3.1. Signal Level Grouping: Candidate Points with Motion

At this level, we find points in the spatio-temporal volume that are probably part of a moving object. The process starts with the normalization of the average intensity in each frame to a fixed value of 100 so as to reduce the effect of simple overall illumination changes across frames. Of course, spatially variant illumination change manifestations remain, which will be rejected in the subsequent grouping processes. In these intensity normalized image sequence frames ($I(x, y, t)$) we find 3D edges based on a 3D extension of the 2D Canny edge detector. We detect single width edge surfaces by using the 3D extension of the 2D nonmaxima suppression scheme of Canny using the 3D gradient estimates. Figure 2b shows a slice through the 3D edge surfaces for the time instant corresponding to the frame shown in Fig. 2a. Note that some of the edges appear to be thicker than a single pixel because the 3D edge surface, which is a single pixel thick, was locally parallel to the $x - y$ image plane at those points. In other words, the frame-to-frame motion was more than a pixel.

FIG. 2. (a) Grayscale image at time t . (b) A slice through the 3D edge surface at time t .

The 3D edges due to static background features and those that arise due to illumination changes have different characteristics in the spatio-temporal volume. Let the angle that the local 3D gradient direction makes with the time axis be denoted by θ_t . This angle would be zero when there is no motion in time and would be 90° when there is a scene illumination change across frames. Figure 3a shows an XT slice of an image sequence with no change in the illumination and Fig. 3b shows the XT slice of an image sequence with changing illumination. Note that the XT slice in Fig. 3a contains lines predominantly parallel to the time axis whereas the XT slice in Fig. 3b contains lines that are parallel and perpendicular to the time axis; the perpendicular lines are due to changes in illumination.

3.2. Primitive Level Grouping: Temporal Envelopes

At this level we group the signal level 3D edges into planar patches—the temporal envelopes—that locally describe the motion in the spatio-temporal volume. The rationale for these planar patches is as follows. Each image feature in motion will sweep out a surface in the spatio-temporal volume. For instance, a point on an object that is moving at a constant velocity will sweep out a straight line in the spatio-temporal volume, a straight boundary of that object will sweep out a plane in the spatio-temporal space, and an edge that is accelerating would sweep out a curved surface. The shape of the spatio-temporal surface will, in general, be complex. *Instead of a mathematical specification of this spatio-temporal surface, we opt to describe this surface as a collection of planar spatio-temporal patches, which we call temporal envelopes.*

The first step in the detection of these temporal envelopes is the Hough transform designed to detect planes. To reduce the memory requirements, we restrict ourselves to finding vertical planes, since the motion scenarios of interest are those where the component of the motion

FIG. 3. (a) XT slice for a normal fixed illumination scene with one moving person. (b) XT slice for a changing illumination scene with one moving person. (Time axis is along the vertical direction.)

FIG. 4. (a) The spatio-temporal volume with a spatio-temporal plane parallel to the Y -axis, the vertical direction. (b) Parameters of the spatio-temporal surface.

along the vertical is significantly lower than along the other two dimensions. However, the following implementation of the Hough transforms can be easily generalized to take into account general motion, albeit at the cost of additional memory requirement associated with adding another dimension to the Hough space.

Let the X , Y , and T axes be oriented as shown in Fig. 4a; then a vertical plane can be parameterized by just the line that is the intersection of the vertical spatio-temporal plane with the XT plane. This intersecting line can, in turn, be parameterized by the intersection points with the $N_X \times N_T$ sized rectangle that bounds the motion sequence as shown in Fig. 4b. These two intersecting points can be parameterized by the locations of the points along the rectangular boundary using just two numbers, $\mathbf{p}_1$ and $\mathbf{p}_2$, that range from 0 to $2(N_X + N_T)$. Thus, the required Hough space is two dimensional, with $4(N_X + N_T)^2 / \Delta_H^2$ cells, where Δ_H is the quantization unit of each dimension. One particular advantage of this parameterization scheme is that the two dimensions represent parameters with the same units, unlike the traditional (d, θ) parameterization of straight lines. In addition, in the latter case, the fit error varies along the extracted line; the error increases as one moves away from the origin.

Each 3D edge point is associated with a local 3D gradient direction, which we use as an estimate of the normal to the plane passing through that point. Given the estimate of the surface normal and the location of the point, it is easy to compute the corresponding point $(\mathbf{p}_1, \mathbf{p}_2)$ in the Hough space. Each edge pixel votes for a single point in the Hough space, thus eliminating the possibility of false peaks. We make the vote proportional to the derivative of the intensity function along the time axis $(\frac{\partial I(x, y, t)}{\partial t})$, which would be low for 3D edge points that do not move. Strong edges that move or edges that move fast would have high weight. This weighted voting strategy automatically prunes out the background edges that do not move or those that move very slowly and biases the detection toward motion edges. We find local maxima by considering peaks in a window $N_{Peak} \times N_{Peak}$ and consider only those above V_{min} votes. Figure 5 shows a couple of actual Hough spaces from our experiments, which are described later. The peaks are pretty sharp and easy to locate.

The Hough transform results in infinite planes that fit the edge points. We arrive at a finite representation by considering all 3D edge pixels that voted for a particular plane

FIG. 5. Some sample Hough spaces with darker regions indicating regions with larger votes. (a) The Hough space for the case considered in Fig. 8, where we detected two spatial envelopes for the two persons moving in opposite directions. (b) The Hough space for the case considered in Fig. 9, where we detected six groups of temporal envelopes corresponding to the six persons in the scene.

and consider their 3D convex hull, which we term the temporal envelope—it captures the temporal coherence of a set of proximal points on the object. The intersection of a temporal envelope with a plane perpendicular to the time axis is a *spatial envelope*.

3.3. Structural Level Grouping: Spatial Envelopes

A moving object in a scene 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 volumes, with each object in motion giving rise to more one or more temporal envelopes. The problem is to group the temporal envelopes that are from a single object. For this grouping process, we use the perceptual organizational principles of proximity, continuity, and parallelism. Temporal envelopes from a single moving object will tend to be close together, mostly locally parallel, and continuous.

The grouping process is based on the existence of salient relationships which form the links of a Gestalt relationship graph whose nodes are the temporal envelopes. We employ Bayesian networks to quantify the saliency of the relationships between any two temporal envelopes. We classify the relation between two temporal envelopes into two classes: “SALIENT” ($S = 1$) or “-SALIENT” ($S = 0$), $S = 1$ being the case when both the primitives should be grouped together and $S = 0$ being the other case.

The classification of a relationship into $S = 0$ and $S = 1$ is based on the following two attributes between two temporal envelopes. For every pair of temporal envelopes, one attribute is the angle, θ , between the surface normals, normalized by $\pi/2$. The second attribute, D_{diff} , is the absolute difference in the perpendicular distances of the planes from the origin normalized by the maximum possible distance; i.e., $\sqrt{N_X^2 + N_T^2}$, where N_X is the number of columns in each image frame and N_T is the number of frames. Note that both the attributes are normalized and range from 0 to 1.

FIG. 6. Bayesian network used to classify pairs of plane primitives.

Based on the values of these three attributes, we classify the relationship using the Bayesian network² shown in Fig. 6. The network classifies pairs of plane primitives as SALIENT or $\neg$ SALIENT. To fully specify a Bayesian network we have to specify the priors of the root nodes and the conditional probabilities of the other nonroot nodes given the states of their respective parents. The only root node in the Bayesian network is the classificatory node with two states $S = 1$ and $S = 0$. In the absence of contrary evidence, we assume equal prior, i.e., it is equally likely for the SALIENT and $\neg$ SALIENT relations to exist between a random pair of temporal primitives. As for the conditional probabilities, we need to specify the probability of an attribute node given the state of its parent in the Bayesian network. For example, for the relational attribute θ we need to specify $P(\theta = t|S = s)$, where s is zero or one. Thus the probability distribution for θ is specified by $P(\theta = t|S = 1)$, and $P(\theta = t|S = 0)$, respectively. For a salient relation the value of θ should be ideally zero. We can represent such a distribution using the right angled triangular function, $Tn(x, b)$, which is equal to $2(b - x)/(b^2)$ over $x \in (0, b)$ and is zero elsewhere. The probability is maximum when x is equal to zero. The state $S = 0$ represents a completely random scenario, which we capture by choosing $P(\theta = t|S = 0)$ to be a uniform distribution over $(0, 1)$, denoted by $U(0, 1)$. Note that Bayesian networks easily facilitate classification using mixtures of different distributions. We choose all the conditional probabilities of the Bayesian Networks to be as follows:

$$\begin{aligned}
 P(D_{diff}|S = 1) &= Tn(D_{diff}, D_{tol}) \\
 P(D_{diff}|S = 0) &= U(0, 1) \\
 P(\theta|S = 1) &= Tn(\theta, \theta_{tol}) \\
 P(\theta|S = 0) &= U(0, 1).
 \end{aligned} \tag{1}$$

Note that all these conditional probabilities are expressed in terms of D_{tol} and θ_{tol} . These parameters represent the effective tolerances in the grouping parameters. Thus, D_{tol} is the distance tolerance and θ_{tol} is the angle tolerance.

Grouping of the temporal envelopes is based on the probability that the relationship between them is salient given the variables (evidence) D_{diff} and θ . This probability is computed by probabilistic message passing in the Bayesian network [62]. The nodes representing the

² Bayesian networks [62] are directed acyclic graphs whose nodes represent variables of interest and whose 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 conditional dependencies and to adapt to changing conditions by updating the probability measures attached with the nodes. The dependency information encoded in the Bayesian networks facilitates the grouping of the plane primitives effectively.

temporal envelopes in the scene Gestalt graph are connected if the value of the probability $P(S=1|D_{diff}, \theta)$ is greater than $P(S=0|D_{diff}, \theta)$. The connected components of this Gestalt graph determine the group of temporal envelopes from one object.

The reader might wonder why we invoked the use of the Bayesian network for the simple grouping process outlined here. It is certainly true that our present grouping strategy at the structural level, which is fairly simple, could have been cast in terms of simple probabilistic language without the use of the powerful Bayesian network formalism. Our choice was motivated by the realization that for motion scenarios that are more complex than those considered in this paper, we would have to consider a more substantial structural level grouping process than that considered here. And from our prior experience with grouping in the spatial domain [29], we believe that Bayesian networks offer a powerful language to capture the semantics of complex groupings. Hence, with the potential future need in mind, we have formulated this structural grouping level in terms of Bayesian networks.

4. EXPERIMENTAL VALIDATION

We have successfully tested the grouping process on a variety of challenging image sequences. We show results on (i) sequences in outdoor settings, (ii) sequences with multiple moving persons, (iii) sequences of an indoor scene with significant occlusions, (iv) sequences with severe simulated noise conditions, (v) sequences with changing illumination conditions, and (vi) a sequence taken with a slowly moving camera. The results will demonstrate not only the viability of a computational model based on perceptual organization for motion segmentation, but also that such a framework also renders the segmentation robust to many image artifacts such as noise, occlusions, and illumination changes. Many existing motion segmentation strategies would not be able to handle many of the severe imaging conditions that we have used.

To give some comparison context to the state of the art, we also present the results of processing the same image sequences by the recently proposed background maintenance algorithm Wallflower [63]. A brief overview of our implementation of this algorithm and the results on several test sequences are given in Section 4.12.

We also note that all the frames in all the sequences used were gray-level images; color was not used.

4.1. Two Persons Walking in the Same Direction

Figure 7 displays the results on a sequence in which two persons are walking in the same direction, left to right, but with some delay. Figures 7a and 7b show two image frames from the sequence of 153 frames. One of the persons in this sequence has light colored clothing whose intensity is very similar to the background. Many of the conventional motion tracking algorithms that rely on difference techniques would fail to handle such a situation. Figures 7c and 7d show XY slices through the 3D edges detected in the images in Figs. 7a and 7b, respectively. Notice how the edges from the light-color clothed person do not organize very well spatially. However, there is significant organization in the temporal dimension, as is shown in Fig. 7e, which shows an XT slice through the 3D edges.

The grouping algorithm detected three planes, which were grouped into two groups, corresponding to each of the two persons. Figure 7f shows the temporal envelopes in the

FIG. 7. Segmentation of an image sequence of two persons moving in the same direction. (a) Image frame 70. (b) Image frame 130. XY slices of the 3D edges corresponding to frames 70 and 130 are shown in (c) and (d), respectively. (e) XT slice of the 3D edges at $Y = 44$. (Axis T is along the vertical direction). (f) The temporal envelopes in the spatio-temporal volume. The spatial envelopes for frames 70 and 130 are shown overlaid on the original image in Figs. 20(a) and (b).

spatio-temporal volume. Each group of temporal envelopes has a unique color. Figures 7g and 7h show the spatial envelopes, which are cross-sections of the temporal envelopes, overlaid on the grayscale images. It should be mentioned that even motion segmentations that do not result in perfect outlines of the moving object (e.g., the feet are outside the envelopes) are useful. These motion segmentation outputs can be used, for instance, to decide where to apply and to guide many gait analysis trackers.

4.2. Two Persons Walking in Opposite Directions

Figure 8 displays the results on a sequence in which two persons from the previous experiment are walking in opposite directions. In this sequence of 141 frames, there were four to five frames during which one person significantly occluded the other person. Also, the background had an undulating shading pattern across the sequence, which is evident in the XT slice shown in Fig. 8e.

FIG. 8. Segmentation of an image sequence of two persons moving in opposite directions. (a) Image frame 45. (b) Image frame 90. XY slices of the 3D edges corresponding to frames 45 and 90 are shown in (c) and (d), respectively. (e) XT slice of the 3D edges at $Y = 43$. (The T-axis is along the vertical direction.) (f) The temporal envelopes in the spatio-temporal volume. The spatial envelopes for frames 45 and 90 are shown overlaid on the original image in Figs. 20(c) and (d).

The algorithm detected two groupings of the three temporal envelopes, corresponding to the persons, which are shown in Fig. 8f. The spatial envelopes corresponding to the frames in Figs. 8a and 8b are shown in Figs. 8g and 8h.

4.3. Six Persons

Next, we consider a more complicated scenario of six persons: three persons walking to the left and three persons walking to the right, during different but overlapping time intervals. The entire sequence consists of 233 frames, with two sample frames shown in Figs. 9a and 9b. In this sequence too, the shading pattern across the background changed over time, which also shows up in the XT slice shown in Fig. 9e, especially toward the left and the right sides of the image. Note the complicated nature of the interactions between the spatio-temporal traces of the different persons. There are five occlusion events over the whole sequence. The algorithm detected 10 temporal envelopes, which were then grouped

FIG. 9. Segmentation of an image sequence of six moving persons. (a) Image frame 78. (b) Image frame 168. XY slices of the 3D edges corresponding to frames 78 and 168 are shown in (c) and (d), respectively. (e) XT slice of the 3D edges at $Y = 41$. (The T -axis is along the vertical.) (f) The temporal envelopes in the spatio-temporal volume. The spatial envelopes for frames 78 and 168 are shown overlaid on the original image in Figs. 20(e) and (f).

into six groups. These grouped temporal envelopes are shown in different colors in Fig. 9f. Figs. 9g and 9h show the spatial envelopes overlaid on the original gray-level image. Note how we are able to easily segment out the trajectories of the different persons.

4.4. Indoor Sequence with Significant Occlusion

The two frames shown in Figs. 10a and 10b are from a 154-frame sequence of two persons walking to the left and one person walking to the right. This sequence is challenging for two reasons. First, the persons walking were occluded partially behind the stairs for half the sequence. Only parts of them were visible during these frames. Second, as the persons cross behind the staircase, the image intensities of the neighborhood around the thin stair railings change, thus changing the differential image structure at the railings, resulting in noisy edges. Many frame-by-frame analysis strategies would have a tough time on this sequence.

FIG. 10. Segmentation of an image sequence of three moving persons with occluding scene elements. (a) Image frame 50. (b) Image frame 120. XY slices of the 3D edges corresponding to frames 50 and 120 are shown in (c) and (d), respectively. (e) XT slice of the 3D edges at $Y = 49$. (f) The temporal envelopes in the spatio-temporal volume. The spatial envelopes for frames 50 and 120 are shown overlaid on the original image in Figs. 20 (g) and (h).

Figure 10e shows 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 are close together. The four groups of temporal envelopes found are shown in Fig. 10f. Three of the groups correspond to the three persons and one group is a false alarm inferred from partial trace data from each of the two persons moving to the left. The spatial envelopes are shown overlaid on the images in Figs. 10g and 10h. Notice that the integrity of the envelopes is maintained across the staircase occlusions.

4.5. Outdoor Running Sequence Taken in the Late Evening

To test whether the speed of motion relative to the rate of frame captures, which was 30 frames per second, affect motion segmentation, we consider a sequence of four running persons, whose two frames are shown in Figs. 11a and 11b. Another aspect in which this sequence is different from the earlier ones is that it was taken during the evening hours,

FIG. 11. Segmentation of an image sequence of four persons running one after another. (a) Image frame 16. (b) Image frame 19. XY slices of the 3D edges corresponding to frames 16 and 19 are shown in (c) and (d), respectively. (e) XT slice of the 3D edges at $Y = 68$. (f) The temporal envelopes in the spatio-temporal volume. The spatial envelopes for frames 16 and 19 are shown overlaid on the original image in Figs. 21(a) and (b).

in partial darkness. The slices through the 3D edges, shown in Figs. 11c and 11d, show decreased spatial coherence; however, the XT slice, in Fig. 11e, shows significant temporal coherence, which helps in detecting the temporal envelopes shown in Fig. 11f. There are four temporal envelopes corresponding to the four persons who ran by the camera, one after the other: two ran from the right to the left followed by two who ran in the other direction. The detected spatial envelopes, two of which are shown in Figs. 11g and 11h, are, as expected, larger in size than for walking sequences.

4.6. Sequences with Simulated Severe Noise and Illumination Conditions

To study the behavior of the segmentation strategy in the presence of noise we considered the six person outdoor sequence in Fig. 9 and added artificially generated noise and illumination effects.

Noise effects. We added uniformly distributed noise so that the resulting signal to noise ratio, defined as the ratio of the average image intensity to the standard deviation of the noise, was 0.5. Two noisy frames are shown in Figs. 12a and 12b. We added noise in two different manners. First, we added noise to all the frames. In the second set of experiments we added noise to frames 40 to 50, 100 to 120, and 180 to 200 to simulate the effect to burst noise. Figures 12c and 12d show the XT slices through 3D edges found in the spatio-temporal

FIG. 12. Results on an image sequence with six moving persons with (case i) simulated uniform noise added to all the frames, and (case ii) uniform noise added to frames 40 to 50, 100 to 120, 180 to 200. Two noisy frames are shown in (a) and (b). XT slices through the 3D edges for cases (i) and (ii) are shown in (c) and (d), respectively. The detected temporal envelopes for the two cases are shown in (e) and (f), respectively.

volume after adding noise in these two manners, respectively. Increased noise is evident. Figures 12e and 12f show the six groups of temporal envelopes that were detected for the two noise cases; the output is similar to the noiseless case seen before. The algorithm is able to segment the six persons, in spite of the severe noise.

Illumination changes. We considered two different types of illumination changes: one where there was random change in the overall intensity from frame to frame for short periods of time and the other where there was gradual change, both across each frame and over time. To simulate the first effect we added (after intensity normalization) to frames 40 to 50, 100 to 120, and 180 to 200 randomly generated intensity values, in the range $[-I_{avg}/2, I_{avg}/2]$, where I_{avg} is the average intensity of the image. All pixels in the same frame were modified by one randomly generated intensity. This condition simulates the effect of flickering lights. To simulate the effect of gradual illumination change we added a sinusoidal function to

FIG. 13. Results on an image sequence with six moving persons with drastic (i) spatially invariant and (ii) spatially variant illumination change. Two noisy frames from the spatially variant case are shown in (a) and (b). XT slices through the 3D edges for cases (i) and (ii) are shown in (c) and (d), respectively. The detected temporal envelopes for the two cases are shown in (e) and (f), respectively.

the intensities, whose phase changed over time. Two frames from this modified sequence are shown in Figs. 13a and 13b, where we can see the banded light and dark pattern. Note that location of these illumination bands moves over time.

Figures 13c and 13d show the XT slices through 3D edges found in the spatio-temporal volume for the burst flickering and the gradual sinusoidally varying illumination noise cases, respectively. Note that the nature of the noise in the 3D edges is different for the two illumination variations. For one it gives rise to noise edges that are organized perpendicular to the time axis and for the other, the extraneous edges are somewhat organized along the time axis. Figure 13e shows the temporal envelopes detected in the burst flicker illumination effect. Note that the envelopes for the six persons are detected along with some envelopes that are perpendicular to the time axis, which can be easily separated out based on their orientations. Figure 13d shows the temporal envelopes detected for the gradual sinusoidally varying illumination case; the output is similar to the noiseless case seen before. The algorithm is able to segment the six persons, in spite of the noise. It is also worth noting

that parameters used to generate these results were the same as those for the noiseless case, except for the $V_{\min}$ parameter used to filter the peaks in the Hough space.

4.7. Indoor Sequence with Real Change in Illumination

In this experiment, we consider an image sequence of 105 frames of a person walking from left to right, with drastic changes in illumination conditions. Lights in the room were randomly turned on and off, multiple times, during the motion capture. Two sample frames are shown in Figs. 14a and 14b. It is also worth mentioning that the white board in the background has a significant specular component to it and so it created additional illumination artifacts that change with time. The illumination change is not just a change in the overall intensity.

The effect of the illumination changes is evident in the XT slice through the 3D edges shown in Fig. 14e. Note the highly organized illumination edges perpendicular to the time axis. The algorithm detects temporal envelopes, which are shown in Fig. 14f, representing both the illumination artifacts and the moving person; however, these two groups are easily

FIG. 14. Segmentation of an image sequence of one person moving on an elliptical trajectory. (a) Image frame 18. (b) Image frame 35. XY slices of the 3D edges corresponding to frames 18 and 35 are shown in (c) and (d), respectively. (e) XT slice of the 3D edges at $Y = 49$. (f) The temporal envelopes in the spatio-temporal volume. The spatial envelopes for frames 18 and 35 are shown overlaid on the original image in Figs. 21(c) and (d), respectively.

separable from each other. The spatial envelopes corresponding to the person are shown overlaid on two image frames in Figs. 14g and 14h.

4.8. Outdoor Sequence with Elliptical Motion Trajectory

The sequences in the experiments presented so far had fronto-parallel motion. In this experiment we wanted to study how the algorithm will behave with motion that is not strictly fronto-parallel. For non-fronto-parallel motion the spatio-temporal traces will not be planar; however, the algorithm should be able to produce planar approximations of them. We considered a 659-frame sequence of a person walking in an elliptical path in front of the camera. The camera axis was not perpendicular to the major axis and was placed off center, more in line with the left focus of the ellipse. So, the captured motion is not strictly fronto-parallel. Two frames from this sequence are shown in Figs. 15a and 15b. The slices

FIG. 15. Segmentation of an image sequence of one person moving on an elliptical trajectory. (a) Image frame 180. (b) Image frame 470. XY slices of the 3D edges corresponding to frames 180 and 470 are shown in (c) and (d), respectively. (e) XT slice of the 3D edges at $Y = 57$. (f) The temporal envelopes in the spatio-temporal volume. The spatial envelopes for frames 180 and 470 are shown overlaid on the original image in Figs. 21(e) and (f), respectively.

through the 3D edges are shown in Figs. 15c and 15d, which show significant amount of background clutter and the edges for the moving person are a small fraction of the total number of edges. However, the moving edges are easily separated out in the temporal domain, as is evident from the XT slice shown in Fig. 15e. Note that the temporal trace is not linear—it is almost parabolic. The algorithm finds two groups of temporal envelopes, which are shown in Fig. 15f, approximating the almost planar portions of the spatio-temporal surface. One group corresponds to the person moving to the left and the other set is created when the person moves back to the right. The spatial envelopes, overlaid on two images, are shown in Figs. 15g and 15h. We would like to draw attention to the segmentation of the person when she is in front of the checkered calibration object. From our experience, we have found that the strong edges in the checked box create artifacts in algorithms that rely on image differencing, either computed between consecutive frames or computed using an estimated background.

4.9. Outdoor Short Sequence with Slowly Moving Camera

The sequence studied in this section differs from the others in three respects: (i) it considers a nonhuman moving subject, namely a lion, (ii) it does not consider a static camera, but a slowly moving one, i.e., motion toward the left, and (iii) it is very short sequence of 22 frames. Two frames from this sequence are shown in Figs. 16a and 16b, with the corresponding slices through the 3D edges shown in Figs. 16c and 16d. As the XT slice through the 3D edges in Fig. 16e shows, the motion is toward the camera and not fronto-parallel. Even with a short sequence, we were able to detect the temporal envelopes corresponding to the motion of the lion, which are shown in Fig. 16f. As the spatial envelopes overlaid on the original images in Figs. 16g and 16h show, the envelopes correspond to the different parts of the lions that are in motion.

4.10. Algorithm Parameter Choices

The algorithm parameter choices that were used to generate the results shown in this paper are listed in Table 2. Notice that for the most part the same parameters were used across a variety of imaging and scene conditions, which attests to the robustness of the algorithm. For the few cases when we had to change the parameters, they were mostly at the structural and the primitive levels to control the number of false alarm temporal envelopes detected due to severe noise or illumination conditions.

4.11. Computational Resources

Most of the computational resources are used at the signal level to generate the 3D edges. In terms of memory use, all the frames of the sequences need not be loaded onto the memory simultaneously, but can be cached in as they are needed for edge detection. The 3D edges for a particular frame are dependent on a finite number of neighboring frames, depending on the temporal smoothing parameter, σ_t . If we truncate the smoothing mask at $4\sigma_t$, a liberal limit, we need just $4\sigma_t$ frames before and after a particular frame to compute the 3D edges in that frame. Thus, in our implementation we keep $8\sigma_t + 2$ image frames in the memory at any time. As 3D edges data are computed frame by frame, new frames are cached in and old ones cached out at each iteration. Thus, the 3D edge data are generated

FIG. 16. Segmentation of an image sequence of a lion moving toward a camera that itself is moving slowly to the left. (a) Image frame 13. (b) Image frame 16. XY slices of the 3D edges corresponding to frames 13 and 16 are shown in (c) and (d), respectively. (e) XT slice of the 3D edges at $Y = 109$. (f) The temporal envelopes in the spatio-temporal volume. The spatial envelopes for frames 13 and 16 are shown overlaid on the original image in Figs. 21(g) and (h), respectively.

frame by frame, sequentially, at a rate taken to smooth and process one frame of image data. The Hough transform space is two dimensional and the memory required is not a concern.

The times taken to process each of the sequences shown in this paper are listed in Table 3. Note that the time taken for the primitive and the structural level groupings is small compared to the signal level. The times are on a 440-MHz Pentium PC with Windows98 and with 512 MByte main memory and include CPU and I/O times. With a faster processor and better computing architectures, these times can be further reduced. Also, in the present software configuration, the primitive level grouping process follows the completion of the signal-level grouping, which need not be the case; 3D edge points can vote into the Hough space as they are detected, frame by frame. This interlaced strategy will further reduce the memory and time needed to save and access the 3D edge information. In the future, it might be possible to implement a real time version of the algorithm, with faster CPUs and some amount of parallelism.

TABLE 2

Algorithm Parameters Values, Grouped Together According to the Level of Grouping, That Were Used to Generate the Results Shown in This Paper

Sequence (Figure #)	Parameters							
	Signal level			Primitive level			Structural level	
	σ_{xy}	σ_t	T_{mag}	Δ_H	N_{peak}	V_{min}	θ_{tol}	D_{tol}
Fig. 7 (Two persons, same direction)	3.0	2.0	1.0	3	1	10,000	0.05	0.04
Fig. 8 (Two persons, different directions)	3.0	2.0	1.0	3	1	10,000	0.05	0.04
Fig. 9 (Six persons)	3.0	2.0	1.0	3	1	10,000	0.05	0.04
Fig. 10 (Three persons, staircase)	3.0	2.0	1.0	3	1	10,000	0.01	0.04
Fig. 11 (Four persons, running)	3.0	2.0	1.0	3	1	10,000	0.05	0.04
Fig. 12e (Noise in all frames)	3.0	2.0	1.0	3	1	8,000	0.05	0.04
Fig. 12f (Burst noise frames)	3.0	2.0	1.0	3	1	8,000	0.05	0.04
Fig. 13e (Burst flickering illumination)	3.0	2.0	1.0	3	1	10,000	0.05	0.04
Fig. 13f (Gradual sinusoidal illumination)	3.0	2.0	1.0	3	1	8,000	0.05	0.04
Fig. 14 (Flashing lights)	3.0	2.0	1.0	3	1	10,000	0.05	1.00
Fig. 15 (Elliptical motion path)	3.0	2.0	8.0	3	1	20,000	0.05	0.04
Fig. 16 (Motion toward camera, lion)	3.0	2.0	1.0	5	2	1,500	0.05	1.00

Any real time version of the algorithm will, however, have a lag proportional to the minimum number of frames needed to generate the temporal envelopes. This minimum number would depend on the amount of occlusion and illumination change one would like to overcome. Without further detailed experimentation with a real-time version of the algorithm in the domain of interest it is difficult to pin down this number. However, we

TABLE 3

User Times (CPU + I/O) to Generate the Results in This Paper on a 440-MHz Pentium PC with 512M Main Memory

Sequence (Figure #)	Frame Size	Number of frame	User time – CPU + I/O (seconds)			
			Signal level	Primitive & structural	Total	Avg. time per frame
Fig. 7 (Two persons, same direction)	151 × 320	153	187.24s	2.20s	189.44s	1.23s
Fig. 8 (Two persons, different directions)	151 × 320	141	177.03s	1.76s	178.79s	1.26s
Fig. 9 (Six persons)	195 × 320	233	368.28s	9.34s	377.62s	1.62s
Fig. 10 (Three persons, staircase)	151 × 320	154	238.71s	4.55s	243.26s	1.57s
Fig. 11 (Four persons, running)	140 × 320	141	206.46s	3.90s	210.36s	1.49s
Fig. 12e (Noise in all frames)	195 × 320	233	378.68s	10.00s	388.68s	1.66s
Fig. 12f (Burst noise frames)	195 × 320	233	375.30s	10.71s	386.01s	1.66s
Fig. 13e (Burst flickering illumination)	195 × 320	233	374.93s	23.01s	397.94s	1.71s
Fig. 13f (Gradual sinusoidal illumination)	195 × 320	233	395.29s	19.17s	414.46s	1.78s
Fig. 14 (Flashing lights)	215 × 320	105	196.57s	5.50s	202.07s	1.92s
Fig. 15 (Elliptical motion path)	240 × 720	659	2875.79s	937.46s	3813.25s	5.79s
Fig. 16 (Motion toward camera, lion)	235 × 310	22	50.14s	3.57s	53.71s	2.44s

can hazard some guesses for the motion domains studied. The sequences in this papers typically had motion occlusions over five to six frames: to bridge such occlusions would require about 10 to 15 frames. In the lion sequence we have demonstrated that the algorithm can generate segmentations in short sequences of 22 frames. Thus at least a 15- to 20-frame lag seems to be necessary for the kinds of motion scenarios studied.

4.12. Wallflower

To put the results in the performance context of a background subtraction based strategy that is better than just plain differencing, we consider the recent Toyama–Krumm–Brumitt–Meyers Wallflower algorithm, which is a combination of several techniques. It consists of pixel- region- and frame-level processing steps [63]. The pixel-level processing consists of a one-step predictive Wiener filter. The region-level processing uses the results of the pixel-level processing in combination with image differencing and region growing. The frame-level processing is an attempt to deal with changes in illumination. This module requires a separate background model (i.e., training sequence) for each illumination scenario encountered. Since our sequences lacked the required illumination training data, which should include all possible illumination scenarios, and due to the lack of implementation detail of this module in [63], we could not implement the frame-level processing. We, however, implemented the pixel- and the region-level processing. Various parameters were tuned to the test sequences and then those settings were used through out the trials discussed below.

Figure 17 displays the results on a sequence in which two people are walking in opposite directions. It should be noted that due to the nature of the filter used during the pixel processing that the first 25 frames had no foreground pixels marked. This is a serious drawback in the background maintenance algorithms; they usually have to train on a relatively static background before they can be effective. If this training is absent, motion in the early frames of a sequence is missed. As can be seen, the algorithm did poorly in marking the foreground pixels in this sequence. This can be attributed to several factors: sensitivity to noise, too much sensitivity in region growing, and lack of robustness to the fluctuating background. Also, due to the use of two sequential difference images in the region level processing, an object is included in the foreground only if it has a cross-section parallel to the direction of motion that is smaller than the magnitude of the motion. This algorithm also does badly in handling the occlusion of two foreground objects as can be seen in Fig. 17b.

FIG. 17. Wallflower segmentation of an image sequence of two people moving in opposite directions. (a) A frame with no occlusion and (b) a frame with occlusion.

FIG. 18. Results of Wallflower processing for indoor images of three people occluded by a staircase. (a) Segmentation of the three people and (b) segmentation fails.

The Wallflower algorithm can also have problems separating foreground and background in a cluttered environment. In Fig. 18b, which corresponds to the staircase sequence from Fig. 10, the convergence of the three people causes the algorithm to mistakenly mark not only the three people as foreground, but also some static parts of the scene. It should be noted that there was no long sequence of background-only training frames. This may have changed the accuracy of the algorithm, but as noted above we consider the need for such training to be a problem. In addition, the algorithm seems quite sensitive to small changes in several parameters. It seems as if the algorithm must be fine-tuned for each sequence to be effective.

Similar types of problems are evident in Fig. 19, a sequence of six people moving three in one direction and three in the other with multiple occlusions. The combination of changes in the background and convergence of many moving objects causes the accuracy of the algorithm to vary from one frame to another. In Fig. 19a the algorithm does a reasonable job of segmenting the people that are visible in that frame. In Fig. 19b the background fluctuations have caused most of the scene to be mislabeled.

While these problems may be seen as particular to the Wallflower algorithm, they are representative of the problems seen in most of the pixel-level processing algorithms, particularly lack of robustness, sensitivity to noise, and the need for a long training sequence.

FIG. 19. Results of Wallflower processing on image sequences of six people, three walking in one direction and three in the other. (a) In a frame with three people visible the segmentation works, (b) in a similar frame the segmentation fails.

FIG. 20. Spatial envelopes for some sample frames of the sequences in presented in this paper. (a) and (b) are for frames 70 and 130 of the sequence in Fig. 7(c) and (d) are for frames 45 and 90 of the sequence in Fig. 8(e) and (f) are for frames 78 and 168 of the sequence in Fig. 9(g) and (h) are for frames 50 and 120 of the sequence in Fig. 10.

5. CONCLUSIONS

This work demonstrates the viability of a perceptual organization based computational framework for motion segmentation. The use of perceptual organization principles on the spatio-temporal volume of data renders the segmentation procedure robust with respect to occlusions, noise, and illumination changes. 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, nor are any training data required. The algorithms used for the computational model are simple ones, based on

FIG. 21. Spatial envelopes for some sample frames of the sequences in presented in this paper. (a) and (b) are for frames 16 and 19 of the sequence in Fig. 11(c) and (d) are for frames 18 and 35 of the sequence in Fig. 14(e) and (f) are for frames 180 and 470 of the sequence in Fig. 15(g) and (h) are for frames 13 and 16 of the sequence in Fig. 16.

well-known tools in computer vision, namely, 3D edge detection, Hough transform, convex hulls, probabilistic reasoning, and graph operations; however, the results are dramatic. We are able to handle image conditions that most current motion segmentation algorithms will not be able to or will require extensive modification. Thus, our conclusion: perceptual organization principles are powerful constraints and should be more extensively used in motion analysis than is presently done.

The present strategy relies on extracting planar representations in the spatio-temporal domain, which are accurate for constant fronto-parallel motion but are only first-order approximations for more complex motion scenarios, such as accelerating motion or non-fronto-parallel motion, which are associated with curved but smooth spatio-temporal surfaces. Thus, in the future we plan to use a richer primitive level description in terms of curved surface patches and a more substantial structural-level grouping.

REFERENCES

1. Y. Yacob and M. J. Black, Parameterized modeling and recognition of activities, *Comput. Vision Image Understand.* **73**, 1999, 232–247.
2. S. Fejes and L. S. Davis, Detection of independent motion using directional motion estimation, *Comput. Vision Image Understand.* **74**, 1999, 101–120.
3. T. Y. Tian and M. Shah, Motion estimation and segmentation, *Mach. Vision Appl.* **9**, 1996, 32–42.
4. R. Jain and H. H. Nagel, On the analysis of accumulative difference pictures from image sequences of real world scenes, *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, 1979, 206–214.
5. S. N. Jayaramamurthy and R. Jain, An approach to the segmentation of textured dynamic scenes, *Comput. Vision, Graphics Image Processing* **21**, 1983, 239–261.
6. S. T. Barnard and W. B. Thompson, Disparity analysis of images, *IEEE Trans. Pattern Anal. Mach. Intell.* **2**, 1980, 333–340.
7. P. Bouthemy and M. Francois, Tracking complex primitives in an image sequence, in *Proc. of 12th IAPR International Conference on Pattern Recognition*, pp. 426–431, 1994.
8. J. M. Odobez and P. Bouthemy, Direct incremental model-based image motion segmentation for video analysis, *Signal Process.* **66**, 1998, 143–155.
9. F. Dufaux, F. Moscheni, and A. Lippman, Spatio-temporal segmentation based on motion and static segmentation, in *Proc. of the IEEE International Conference on Image Processing*, Vol. 1, pp. 306–309, 1995.
10. R. Jain and S. P. Liou, Qualitative motion analysis using a spatio-temporal approach, in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 726–727, 1991.
11. C. R. Bolles, H. H. Baker, and D. H. Marion, Epipolar-plane image analysis: An approach to determining structure from motion, *Internat. J. Comput. Vision* **1**, 1987, 7–55.
12. C. R. Bolles and H. H. Baker, Generalizing epipolar-plane image analysis on the spatio-temporal surface, *Internat. J. Comput. Vision* **3**, 1989, 33–49.
13. C. Debrunner and N. Ahuja, Segmentation and factorization-based motion and structure estimation for long image sequences, *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1998, 202–211.
14. C. Tomasi and T. Kanade, Shape and motion from image streams under orthography—a factorization method, *Internat. J. Comput. Vision* **9**, 1992, 137–154.
15. 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. Pattern Anal. Mach. Intell.* **18**, 1996, 58–64.
16. Y. Ricquebourg and P. Bouthemy, Tracking of articulated structures exploiting spatio-temporal image slices, in *Proc. of the IEEE International Conference on Image Processing*, Vol. 3, pp. 480–482, 1997.
17. A. Niyogi and E. Adelson, Analyzing and recognizing walking figures in XYT, *IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 469–474.
18. J. L. Bentley, K. L. Clarkson, and B. D. Levine, Fast linear expected-time algorithms for computing maxima and convex hulls, in *Proc. of 1st Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 56–62, 1993.
19. J. O. Rourke, *Computational Geometry in C*, Cambridge University Press, Cambridge, UK, 1995.
20. S. Sarkar and K. L. Boyer, Perceptual organization in computer vision: A review and a proposal for a classificatory structure, *IEEE Trans. Systems, Man, Cybernet.* **23**, 1993, 382–399.
21. R. Mohan and R. Nevatia, Using perceptual organization to extract 3-D structures, *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 1989, 1121–1139.
22. K. Rao and R. Nevatia, Space description from imperfect and incomplete data, in *International Conference on Pattern Recognition*, pp. 414–426, 1990.
23. S. Sarkar, Learning to form large groupings of salient image features, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 780–786, 1998.
24. P. L. Rosin and G. A. W. West, Extracting surfaces of revolution by perceptual grouping of ellipses, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 677–678, 1991.
25. D. D. Hoffman and B. E. Flinchbaugh, The interpretation of biological motion, *Biol. Cybernet.* **42**, 1982, 195–204.

26. D. G. Lowe, Three-dimensional object recognition from single two-dimensional images, *Artificial Intell.* **31**, 1987, 355–395.
27. J. D. McCafferty, *Human and Machine Vision: Computing Perceptual Organization*, Ellis Horwood, West Sussex, England, 1990.
28. A. Etemadi, J.-P. Schmidt, G. Matas, J. Illingworth, and J. Kittler, Low-level grouping of straight line segments, in *Proceedings of the British Machine Vision Conference*, pp. 119–126, Springer-Verlag, Berlin/New York, 1991.
29. S. Sarkar and K. L. Boyer, Integration, inference, and management of spatial information using Bayesian networks: Perceptual organization, *IEEE Trans. Pattern Anal. Mach. Intell.* **15**, 1993, 256–274. [Special Section on Probabilistic Reasoning]
30. S. Sarkar and K. L. Boyer, A computational structure for preattentive perceptual organization: Graphical enumeration and voting methods, *IEEE Trans. Systems, Man, Cybernet.* **24**, 1994, 246–267.
31. C. Lin, A. Huertas, and R. Nevatia, Detection of building using perceptual grouping and shadows, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 62–69, 1994.
32. M. Shah, K. Rangarajan, and P. S. Tsai, Motion trajectories, *IEEE Trans. Systems, Man, Cybernet.* **23**, 1993, 1138–1149.
33. K. Gould and M. Shah, The trajectory primal sketch: A multi-scale scheme for representing motion characteristics, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 79–85, 1989.
34. G. Roth and M. D. Levine, Geometric primitive extraction using a genetic algorithm, *IEEE Trans. Pattern Anal. Mach. Intell.* **1**, 1994, 901–905.
35. R. L. Castano and S. Hutchinson, A probabilistic approach to perceptual grouping, *Comput. Vision Image Understand.* **64**, 1996, 399–419.
36. L. Williams and A. R. Hanson, Perceptual completion of occluded surfaces, *Comput. Vision Image Understand.* **64**, 1996, 1–20.
37. D. W. Jacobs, Robust and efficient detection of salient convex groups, *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, 1996, 23–37.
38. D. Geiger, K. Kumaran, and L. Parida, Visual organization for figure ground separation, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 156–160, 1996.
39. J. August, K. Siddiqi, and S. W. Zucker, Fragment grouping via the principle of perceptual occlusion, in *Proceedings of the International Conference on Pattern Recognition*, Vol. 1, pp. 3–8, 1996.
40. P. Havaladar, G. Medioni, and F. Stein, Perceptual grouping for generic recognition, *Internat. J. Comput. Vision* **20**, 1996, 59–80.
41. T. F. Syeda-Mahmood, Data- and model-driven selection using parallel line groups, *Comput. Vision Image Understand.* **67**, 1997, 205–222.
42. A. Jascot-Descombes and T. Pun, A synchronous perceptual grouping: From contours to relevant 2d structures, *Comput. Vision Image Understand.* **66**, 1997, 1–24.
43. R. C. Nelson and A. Selinger, A perceptual grouping hierarchy for 3d object recognition and representation, in *DARPA Image Understanding Workshop*, pp. 157–163, 1998.
44. A. Amir and M. Lindenbaum, A generic grouping algorithm and its quantitative analysis, *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1998, 168–185.
45. P. L. Rosin and G. A. W. West, Segmenting curves into elliptic arcs and straight lines, in *International Conference on Computer Vision*, pp. 75–78, 1990.
46. J. Zhong, T. S. Huang, and R. J. Adrian, Salient structure analysis of fluid flow, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 310–315, 1994.
47. A. Verri and T. Poggio, Against quantitative optic flow, in *International Conference on Computer Vision*, pp. 171–180, 1987.
48. C. F. Shu and R. C. Jain, “Vector field analysis for oriented patterns,” in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, p. 673, 1992.
49. S. W. Zucker, Early orientation selection: Tangent fields and the dimensionality of their support, in *Human and Machine Vision* (A. Rosenfeld, Ed.), pp. 335–364, Academic Press, San Diego, 1986.

50. S. W. Zucker, Computational and psychophysical experiments in grouping: Early orientation selection, in *Human and Machine Vision* (J. Beck, B. Hope, and A. Rosefeld, Eds.), pp. 545–567, Academic Press, San Diego, 1983.
51. S. W. Zucker, C. David, A. Dobbins, and L. Iverson, The organization of curve detection: Coarse tangent fields and fine spline coverings, in *Proceedings on 2nd International Conference on Computer Vision*, pp. 568–577, Dec. 1988.
52. A. Sha'ashua and S. Ullman, Structural saliency: The detection of globally salient structures using a locally connected network, in *International Conference on Computer Vision*, pp. 321–327, 1988.
53. N. Ahuja and M. Tuceryan, Extraction of early perceptual structure in dot patterns: Integrating region, boundary, and component gestalt, *Comput. Vision, Graphics, Image Process.* **48**, 1989, 304–356.
54. L. Herault and R. Horaud, Figure ground discrimination: A combinatorial optimization method, *IEEE Trans. Pattern Anal. Mach. Intell.* **15**, 1993, 899–914.
55. J. Shi and J. Malik, Normalized cuts and image segmentation, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 1997.
56. A. Shio and J. Sklansky, Segmentation of people in motion, in *Proceedings of the IEEE Workshop on Visual Motion*, pp. 325–332, 1991.
57. J. Elder and S. Zucker, Computing contour closure, in *European Conference on Computer Vision*, pp. 399–412, 1996.
58. G. Guy and G. Medioni, Inferring global perceptual contours from local features, *Internat. J. Comput. Vision* **20**, 1996, 113–133.
59. H. Ishikawa and D. Geiger, Segmentation by grouping junctions, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 125–131, 1998.
60. S. Casadei and S. K. Mitter, Hierarchical image segmentation: Detection of regular curves in a vector graph, *Internat. J. Comput. Vision* **27**, 1998, 71–100.
61. L. R. Williams and D. W. Jacobs, Stochastic completion fields: A neural model of illusory contour shape and salience, in *International Conference on Computer Vision*, pp. 408–415, 1995.
62. J. Pearl, "Probabilistic Reasoning in Intelligent Systems," Morgan Kaufman, San Mateo, California, 1988.
63. K. Toyama, J. Krumm, B. Brumitt, and J. Meyers, Wallflower: Principles and practice of background maintenance, in *Proc. Internal Conference on Computer Vision*, pp. 255–261, 1999.
64. C. R. Wren, A. Azarbayejani, T. J. Darrell, and A. P. Pentland, Pfunder: Real-time tracking of the human body, in *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 19, no. 7, pp. 780–785, 1997.
65. W. E. L. Grimson, L. Lee, R. Romano, and C. Stauffer, Using adaptive tracking to classify and monitor activities in a site, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 22–31, 1998.
66. T. Matsuyama, T. Ohya, and H. Habe, "Background Subtraction for Non-stationary Scenes," Dept. of Electronics and Communications, Graduate School of Engineering, Kyoto University, Sakyo, Kyoto, Japan, 1999.
67. I. Haritaoglu, D. Harwood, and L. S. Davis, W4s: Real time surveillance of people and their activities, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, no. 8, pp. 809–830, 2000.
68. H. Nakai, Non-parameterized Bayes decision method for moving object detection, in *Asian Conference on Computer Vision*, pp. 447–451, 1995.
69. N. Oliver, B. Rosario, and A. Pentland, A Bayesian computer vision system for modeling human interactions, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, no. 8, pp. 831–843, 2000.