A Spatio-temporal Filtering Approach to Motion Segmentation

Jesús Chamorro-Martínez, J. Fdez-Valdivia, and Javier Martinez-Baena*

Department of Computer Science and Artificial Intelligence
University of Granada, Spain
{jesus,jfv,jbaena}@decsai.ugr.es

Abstract. In this paper, a new frequency-based approach to motion segmentation is presented. The proposed technique represents the sequence as a spatio-temporal volume, where a moving object corresponds to a three-dimensional object. In order to detect the "3D volumes" corresponding to significant motions, a new scheme based on a band-pass filtering with a set of logGabor spatio-temporal filters is used. It is well known that one of the main problems of these approaches is that a filter response varies with the spatial orientation of the underlying signal. To solve this spatial dependency, the proposed model allows to recombine information of motions that has been separated in several filter responses due to its spatial structure. For this purpose, motions are detected as invariance in statistical structure across a range of spatio-temporal frequency bands. This technique is illustrated on real and simulated data sets, including sequences with occlusion and transparencies.
Keywords: Motion segmentation, motion representation, motion pattern, logGabor filters, spatio-temporal models

1 Introduction

The motion segmentation, i.e. the process of dividing the scene into regions representing moving objects, is one of the most important problems in image sequence analisys. It has applications in fields such as optical flow estimartion, video coding or objects tracking.

The most common proposals to this problem relies on frame by frame analysis (for example, techniques based on optical flow estimates). Althought this kind of approaches works fine in many cases, it is well known they have problems in the presence of noise, occlusions or transparencies [1]. To overcome these problems, some authors propose to use extended features to find correspondeces beetween frames. None the less, the success of these models depends on the stability of detection of such features over multiple frames, and the way of solving the correspondence problem [2].

Unlike frame by frame analisys (or analisys over small number of frames), some approaches represents the sequence as a spatio-temporal volume. From this

* This work has been supported by the DGES (Spain) under grant PB98-1374

F.J. Perales et al. (Eds.): IbPRIA 2003, LNCS 2652, pp. 193–203, 2003.

point of view, a moving object may be observed as a three-dimensional object, where the axis x and y correspond to the spatial dimensions, and the third axis corresponds to the temporal dimension [3]. In this kind of methods, some important proposals are based on a band-pass filtering operation with a set of spatio-temporal filters [4, 5, 6, 7]. These approaches are derived by considering the motion problem in the Fourier domain, where the spectrum of a spatio-temporal translation lies in a plane whose orientation depends on the direction and velocity of the motion [8, 9]. Although these filters are a powerful tool to separate the motions presented in a sequence, it is nevertheless true that one of the main problems of these schemes is that components of the same motion with different spatial characteristics are separated in different responses. Moreover, a filter response will change if the spatial orientation or scale vary.

In this paper, we develop a methodology to motion segmentation on the basis of a spatio-temporal volumes detection. For this purpose, a new tehcnique based on a spatio-temporal filtering in the frequency domain is proposed. To solve the problems described above, we propose a new approach that groups the separated responses obtained by the filters in order to extract coherent and independent motions. Using a new distribution of 3D logGabor filters over the spatio-temporal spectrum, *a motion* is detected as an invariance in statistical structure across a range of spatio-temporal frequency bands. This new scheme recombines responses that, even with different spatial characteristics, have continuity in its motion.

2 The Proposed Method

The figure 1 shows a general diagram describing how the data flows through the proposed model. This diagram illustrates the analysis on a given sequence showing a clap of hands. The endpoint of analyzing this sequence is to separate the two hand motions. In a first stage, a three-dimensional representation is performed from the original sequence and then its Fourier transform is calculated. Given a bank of spatio-temporal logGabor filters, a subset of them is selected in order to extract significant spectral information. These selected filters are applied over the original spatio-temporal image in order to obtain a set of active responses.

In the second stage, for each pair of active filters, their responses are compared based on the distance between their statistical structure, computed over those points which form relevant points of the filters. As a result, a set of distances between active filters is obtained.

In a third stage, a clustering on the basis of the distance between the active filter responses is performed to highlight invariance of responses. Each of the cluster obtained in this stage defines a motion. In figure 1, two collections of filters have been obtained for the input sequence.

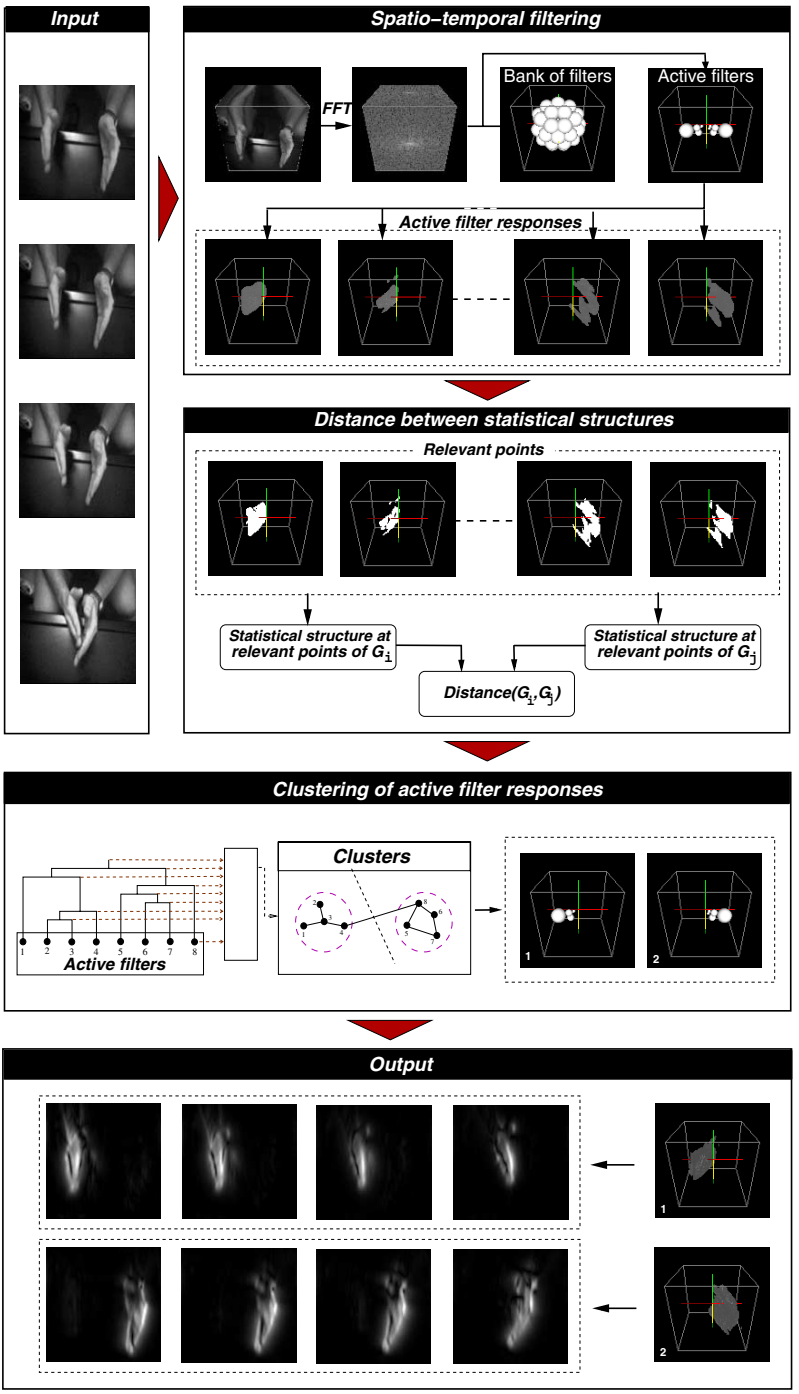

Fig. 1. A general diagram of the proposed model

(θ, φ)		
$(0.52, 0.00)$	$(-0.62, -0.53)$	$(-1.29, 0.45)$
$(0.62, 0.53)$	$(-1.08, 0.97)$	$(-1.05, 0.00)$
$(1.08, 0.97)$	$(1.08, -0.97)$	$(-1.29, -0.45)$
$(-1.08, 0.97)$	$(0.62, -0.53)$	$(1.29, -0.45)$
$(-0.62, 0.53)$	$(1.05, 0.00)$	$(1.57, 0.00)$
$(-0.52, 0.00)$	$(1.29, 0.45)$	

Table 1. Angular coordinates of the bank of filters (over an sphere of ratio 1)

2.1 Bank of Spatio-temporal Filters

To decompose the sequence, a bank of logGabor filters is used. A logGabor function can be represented in the frequency domain as:

$$\phi(\rho, \theta, \varphi) = e^{\left\{ -\frac{\left(log\left(\frac{\rho}{\rho_o}\right)\right)^2}{2\left(log\left(\frac{\sigma_\rho}{\rho_o}\right)\right)^2} \right\}} e^{\left\{ -\frac{(\theta-\theta_o)^2}{2\sigma_\theta^2} \right\}} e^{\left\{ -\frac{(\varphi-\varphi_o)^2}{2\sigma_\varphi^2} \right\}} \quad (1)$$

where σ_θ , σ_φ and σ_ρ are the angular and radial standard deviation, (θ_o, φ_o) is the orientation of the filter, and ρ_o is the central radial frequency. The bank of filters should be designed so that it tiles the frequency space uniformly. Hence we consider a bank with the following features:

1. For each radial frequency, 17 spherical orientations over dynamic planes are considered. Table 1 shows the angular coordinates used in the proposed bank.
2. The radial axis is divided into 3 equal octave bands. The wavelength in each orientation is set at 3, 6 and 12 pixels respectively.
3. The radial bandwidth is chosen as 1.2 octaves
4. The angular bandwidth is chosen as 30 degrees

The resultant filter bank is illustrated in the top of figure 1. Due to the conjugate symmetry in the Fourier domain, the filter design is only carried out on the half 3D frequency space.

Active Filters In order to reduce the number of filter responses that have to be evaluated, a selection of filters that isolate spectral information corresponding to significant motions is performed. This selection allows to reduce the computational cost and it avoids the noisy or less relevant filter responses. Given a filter ϕ_i, a measure of its relevance is defined as:

$$w_i = \frac{1}{Card[V(i)]} \sum_{(\rho,\theta,\varphi)\in V(i)} |F(\rho, \theta, \varphi)| \quad (2)$$

where $|F(\rho, \theta, \varphi)|$ is the amplitude of the Fourier spectrum at (ρ, θ, φ), and $V(i)$ represents a spectral volume associated with the filter ϕ_i. To calculate $V(i)$,

we consider that a point (ρ, θ, φ) in the spatio-temporal frequency domain will belong to $V(i)$ if

$$|\rho - \rho_o| \leq \sigma_\rho \ , \ |\theta - \theta_o| \leq \sigma_\theta \ y \ |\varphi - \varphi_o| \leq \sigma_\varphi \tag{3}$$

where σ_θ, σ_φ, σ_ρ and (θ_o, φ_o) are the logGabor filter parameters (let us remark that it is not necessary to calculate the responses of each filter to obtain these weights)

Using the filter relevance measure defined in (2), an unsupervised classification method is performed for each scale to group the filters into two classes: active ones and non-active ones. The cluster whose filters have the highest weights will determine the set of active filters (that will be noted *Active*). In our implementation, a hierarchical clustering [10] is used with a dissimilarity function between classes defined as

$$\delta(C_i, C_j) = |\mu_i - \mu_j| \tag{4}$$

where

$$\mu_k = \frac{1}{Card\,[C_k]} \sum_{r \in C_k} w_r \tag{5}$$

For each active filter, a set of 'relevant points' is computed. We calculate these points as local energy peaks on the filter responses [11]: given the response E_i of a filter ϕ_i, the maximal of E_i in the direction of the filter will determine the set of points which will focus our attention in the next stage.

2.2 Distance between Filter Responses

In this section, a distance between the statistical structures of a given pair of filter responses is proposed. To represent a statistical structure, we use the notions of *separable feature* and *integral feature* introduced in [12]. A separable feature is defined as any relevant characteristic that may be obtained for a point (phase, local contrast, energy, etc.). The combination of any subset of separable features will define an integral feature at a given point (x, y, z) . In this paper, the following five separable features proposed in [12] will be used: phase, local energy, local standard deviation, local contrast of the local energy, and local entropy.

Let $T^i(x, y, z) = \left[T^i_k(x, y, z)\right]_{k=1,2,\ldots L}$ be an integral feature at (x, y, z) which combines L separable features, noted as T^i_k, computed on the response of the filter ϕ_i. Let $\widehat{d}\left(T^i, T^j\right)$ be the distance between two integral features $T^i(x, y, z)$ and $T^j(x, y, z)$ given by the equation:

$$\widehat{d}\left(T^i, T^j\right) = \sum_{k=1}^{L} \frac{1}{Max_k} d(T^i_k, T^j_k) \tag{6}$$

with Max_k being a normalization factor [12], and $d(\cdot)$ a distance between separable features (this measure $d(\cdot)$ is defined for each separable feature in [12])

Based on the previous equation, a distance between the responses of two filters ϕ_i y ϕ_j is defined as:

$$\widehat{D}(\phi_i, \phi_j) = D\,[i,j]^2 + D\,[j,i]^2 \tag{7}$$

where

$$D\,[r,s] = \frac{1}{Card\,[P(r)]} \left(\sum_{P(r)} \left| \widehat{d}[T^r, T^s] \right|^{\beta} \right)^{\frac{1}{\beta}} \tag{8}$$

with $\widehat{d}\,[T^r, T^s]$ being the distance between integral features given by (6), and $P(r)$ the set of relevant points for the filter ϕ_r .The default value of the exponent β in (8) has been fixed to 3.

2.3 Clustering of Active Filters

In order to obtain a partition $C_1, C_2, ..., C_N$ of active filters, with C_i representing a motion, a clustering of the dataset $X = \{\phi_i \in Actives\}$ into an unknown number N of clusters is performed. For this purpose, a hierarchical clustering is used [10] with a dissimilarity function between classes defined on the basis of distances between statistical structures as

$$\delta(C_n, C_m) = min \left\{ \widehat{D}(\phi_i, \phi_j)\ ,\ \phi_i \in C_n\ ,\ \phi_j \in C_m \right\} \tag{9}$$

where $\widehat{D}(\phi_i, \phi_j)$ is given by the equation (7). Let us remark that the clustering is not performed for each point (x,y,t), but over the set of active filters X.

Selection of the Best Partition To select the level l of the hierarchy which will define the best partition $P^l = C_1, C_2, ..., C_N$, we propose the following function of goodness

$$f(P^l) = \frac{\gamma^*_{P^l}}{\varepsilon^*_{P^l}} \tag{10}$$

where $\varepsilon^*_{P^l}$ and $\gamma^*_{P^l}$ are two measures of the congruence and separation of the partition P^l respectively, given by the equations:

$$\varepsilon^*_{P^l} = max \left\{ \varepsilon_n \mid C_n \in P^l \right\} \tag{11}$$

$$\gamma^*_{P^l} = min \left\{ \gamma_n \mid C_n \in P^l \right\} \tag{12}$$

The congruence degree ε_n and separation degree γ_n of a cluster C_n are defined as

$$\varepsilon_n = max \left\{ cost(\mu^*_{i,j}) \mid \phi_i, \phi_j \in C_n \right\} \tag{13}$$

$$\gamma_n = min \left\{ \delta(C_n, C_m) \mid m = 1, ..., N\ with\ m \neq n \right\} \tag{14}$$

where $\delta(C_n, C_m)$ is defined in (9), and $cost(\mu^*_{i,j})$ is the cost of the optimal path between two elements ϕ_i and ϕ_j in C_n calculated as follow: let $\prod_{ij}$ be the set

of possibles paths linking ϕ_i and ϕ_j in C_n; given a path $\pi_{ij} \in \prod_{ij}$, its cost is defined as the greatest distance between two consecutive points on the path:

$$cost(\pi_{ij}) = max\left\{\widehat{D}(\phi_r, \phi_{r+1}) \,/\, \phi_r, \phi_{r+1} \in \pi_{ij}\right\} \tag{15}$$

where ϕ_r and ϕ_{r+1} are two consecutive elements of π_{ij}, and $\widehat{D}(\phi_r, \phi_s)$ is defined in equation (7). The optimum path $\pi^*_{ij} \in \prod_{ij}$ between ϕ_i and ϕ_j is then defined as the path that links both filters with minimum cost:

$$\pi^*_{ij} = \underset{\pi_{i,j} \in \Pi_{i,j}}{argmin} \{cost(\pi_{i,j})\} \tag{16}$$

Due to the merging process of the hierarchical clustering and the distance between classes used in this case (equation (9)), the congruence degree ε_n equals to the distance between the two cluster which were merged together to obtain C_n [12]. Thus, the calculus of ε_n do not increase the computational cost of the clustering.

3 Results

In this section, the results obtained with real and synthetic sequences are showed to prove the performance of our model. For this purpose, several cases have been tested, from simple motion to occlusions and transparencies. In all the cases, the figures show the first and the last frame of the original sequence, and the motions detected in each case. Each motion, which has associated a cluster of filters, is represented by the sum of the filters responses (energy) of its cluster. In this representation, a high level of energy (white colour) corresponds to a high presence of motion.

A synthetic case of pure translational motion with constant speed is showed in figure 2(A). Specifically, the example shows three bars with velocities of (1,0), (-1,0) and (0,-1) pixels/frame respectively. Looking at the 3D representation of the original sequence, three independent planes can be seen corresponding to the three bars in motion. Our model separates each one of these planes into three different spatio-temporal outputs corresponding to the three motions. From this 3D representation, the sequence associated to each motion is extracted.

Figure 2(B) shows another synthetic example with a moving obtect with velocity of (1,1) pixels/frame. In this case, the object has the same texture that the background, so only the motion information allows to detect the object. As figure 2(B) shows, our model generates an output corresponding to the moving object.

The figure 2(C-D) shows two synthetic sequences which have been generated with Gaussian noise of mean 1 and variance 0. The first example (figure 2(C)) shows a sequence where a background pattern with velocity (-1,0) pixels/frame is occluded by a foreground pattern with velocity (1,0). The second example (figure 2(D)) shows two motions with transparency: an opaque background pattern

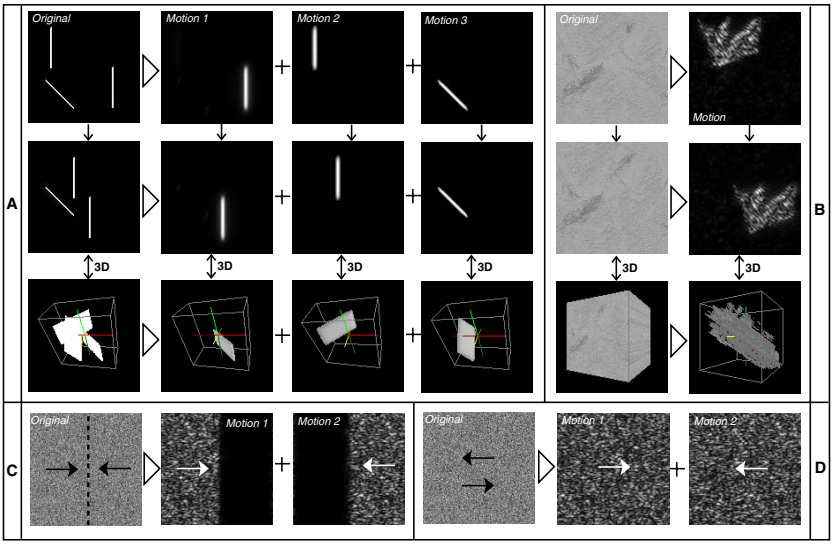

Fig. 2. Output of the model with synthetic sequences

with velocity (1,0), and a transparent foreground pattern with velocity (-1,0). In both cases, the figure shows the central frame of the sequence and the motions detected by the model (two in each case).

Figure 3(A-C) shows three examples with real sequences. In all the cases, boxes around the detected moving objects are showed over the original sequence. Each box is obtained from the energy representation (that is, the sum of the filters responses of the cluster associated to the motion) as the box which enclose the corresponding motion (to select the points with high level of energy a thresholding over the energy representation is performed). The first case corresponds to a double motion without occlusions where two hands are moving to clap. The second one shows an example of occlusion where a hand is crossing over another one. In this case, where the occlusion is almost complete in some frames, the motion combines translation and rotation without a constant velocity. The third case shows an example of transparency where a bar is occluded by a transparent object placed in the first plane. As figure 3 shows, in all the cases our model generates an output for each motion present in the sequence. Let us remark that the problem of the occlusion is solved by our model by mean of the spatio-temporal continuity of forms. Furthermore, this approach is capable of detecting motions even when different velocities and spatial orientations are present.

Figure 3(D) shows the result obtained with a noisy image sequence. This example has been generated by adding Gaussiam noise of mean 1 and variance 30 to the sequence of the figure 3(A). As figure 3(D) shows, our model segments the same two motions that were detected in the original sequence. That enlightens the consistency of the proposed algorithm in the presence of noise.

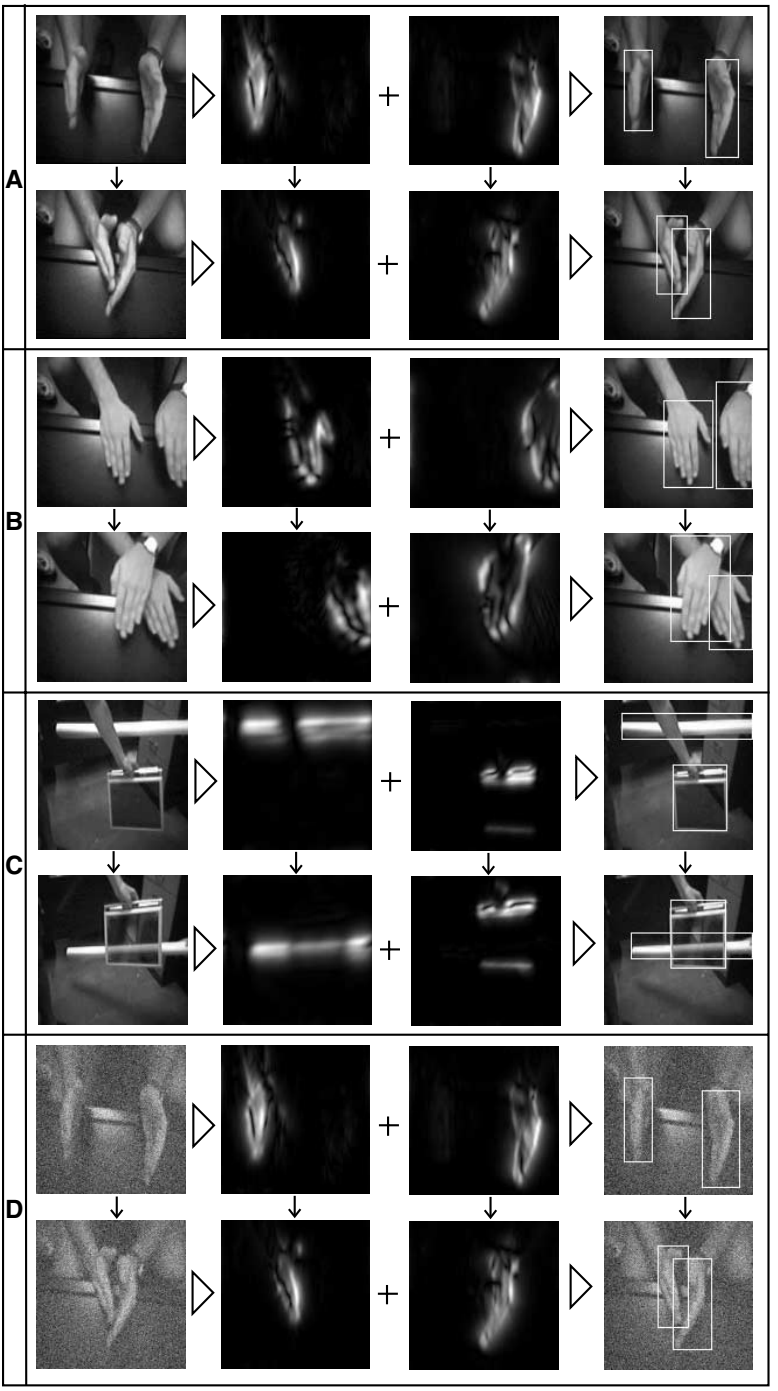

Fig. 3. Results with real sequences

4 Conclusions

In this paper, a new approach to motion segmentation in image sequences has been presented. The sequence has been represented as a spatio-temporal volume, where a moving object correspond to a three-dimensional object. Using this representation, a motion has been identified on the basis of invariance in statistical structure across a range of spatio-temporal frequency bands. To span the spatio-temporal spectrum, logGabor functions have been adopted as an appropriate method to construct filters of arbitrary bandwidth. The new approach allows to recombine information of motions that has been separated in several filter responses due to its spatial structure; as a result, the proposed model generates an output for each coherent and independent motion detected in the sequence, avoiding the classic problem associated with a representation based on spatio-temporal filters.

The technique has been illustrated on several data sets. Real and synthetic sequences combining occlusions and transparency have been tested. In all the cases, the final results enlightens the consistency of the proposed algorithm.

References

[1] F. Moscheni, S. Bhattacharjee, and M. Kunt, "Spatiotemporal segmentation based on region merging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 9, pp. 897–915, 1998. 193

[2] D. Tweed and A Calway, "Integrated segmentation and depth ordering of motion layers in image sequences," *Image and Vision Computing*, vol. 20, pp. 709–724, 2002. 193

[3] K. Korimilli and S. Sarkar, "Motion segmentation based on perceptual organization of spatio-temporal volumes," *Proceedings.15th International Conference on Pattern Recognition*, vol. 3, pp. 844–849, 2000. 194

[4] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *Journal of the Optical Society of America A.*, vol. 2, no. 2, pp. 284–299, Feb 1985. 194

[5] H. Liu, M. Hong, M. Herman, and A. Chellappa, "A general motion model and spatio-temporal filters for computing optical flow," *International Journal of Computer Vision*, vol. 22, no. 2, pp. 141–172, 1997. 194

[6] L. Wiskott, "Segmentation from motion: combining gabor and mallat wavelets to overcom de aperture and correspondence problem," *Pattern Recognition*, vol. 32, pp. 1751–1766, 1999. 194

[7] O . Nestares, C. Miravet, J. Santamaria, and R . Navarro, "Automatic segmentation of low visibility moving objects through energy analysis of the local 3d spectrum," *Proceedings of the SPIE'99*, vol. 3642, pp. 13–22, 1999. 194

[8] S.S Beauchemin and J. L. Barron, "On the fourier properties of discontinuous motion," *Journal of Mathematical Imaging and Vision*, vol. 13, pp. 155–172, 2000. 194

[9] S.S Beauchemin and J. L. Barron, "The frequency structure of 1d occluding image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 2, pp. 200–206, 2000. 194

[10] Anil K. Jain and Richard C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988. 197, 198

[11] M. C. Morrone and R. A. Owens, "Feature detection from local energy," *Pattern Recognition Letters*, vol. 6, pp. 303–313, 1987. 197

[12] R. Rodriguez-Sanchez, J. A. Garcia, J. Fdez-Valdivia, and X. R. Fdez-Vidal, "The rgff representation model: A system for the automatically learned partitioning of 'visual patterns' in digital images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 1044–1072, 1999. 197, 199