# Motion Representation using Composite Energy Features

Raquel Dosil

Corresponding Author

Dep. de Electrónica e Computación, Univ. de Santiago de Compostela,

Campus Universitario Sur, s/n, 15782, Santiago de Compostela, Spain

e-mail address: rdosil@usc.es

Phone: +34 981 563 100

Fax: +34 981 528 012

Xosé R. Fdez-Vidal

Escola Politécnica Superior, Univ. de Santiago de Compostela,

Campus Universitario, s/n, 27002, Lugo, Spain

E-mail address: faxose@usc.es

Xosé M. Pardo

Dep. de Electrónica e Computación, Univ. de Santiago de Compostela,

Campus Universitario Sur, s/n, 15782, Santiago de Compostela, Spain

E-mail address: pardo@dec.usc.es

**Abstract**.

This work tackles the segmentation of apparent-motion from a bottom-up perspective. When no information is available to build prior high-level models, the only alternative are bottom-up techniques. Hence, the whole segmentation process relies on the suitability of the low-level features selected to describe motion. A wide variety of low-level spatio-temporal features have been proposed so far. However, all of them suffer from diverse drawbacks. Here, we propose the use of composite energy features in bottom-up motion segmentation to solve several of these problems.

Composite energy features are clusters of energy filters –pairs of band-pass filters in quadrature–, each one sensitive to a different set of scale, orientation, direction of motion and speed. They are grouped in order to reconstruct independent motion patterns in a video sequence. A composite energy feature, this is, the response of one of these clusters of filters, can be built as a combination of the responses of the individual filters. Therefore, it inherits the desirable properties of energy filters but providing a more complete representation of motion patterns.

In this paper, we will present our approach for integration of composite features based on the concept of Phase Congruence. We will show some results that illustrate the capabilities of this low-level motion representation and its usefulness in bottom-up motion segmentation and tracking.

**Keywords**: Spatio-temporal energy filtering; Feature integration; Composite energy features; Apparent-motion segmentation and tracking.

## 1. Introduction

Many recent motion segmentation techniques have been very successful due to the use of top-down approaches [1][2]. These techniques can be applied as long as there is some available prior model or template of the target object. If this is not the case, the performance of segmentation depends on the rules applied to infer high level representations from low level data –bottom-up process–, but also on the quality of the low level representation itself. A wide variety of low-level spatio-temporal features have been proposed in literature so far [3]. However, all of them suffer from diverese drawbacks. They can be classified into motion detection and motion estimation techniques.

Motion detection methods classify points into static or mobile. As a consequence, they are only valid for static background scenes, and can not classify motion patterns according to their velocity and direction of motion. A very popular class of approaches for motion detection is background subtraction. Moving objects are identified as groups of connected pixels whose

features differ from those of the background model. Background modelling is usually accomplished using Gaussian mixture models [4]. Although these techniques are the most widely used, they are not free of limitations. They are very sensitive to scene changes in illumination and motion. Another solution for background modelling is PCA analysis within spatio-temporal pixel blocks, which reduces uncertainty and provided more stability to noise and illumination changes [5]. Finally, temporal differencing approaches, detect moving pixels by differencing between two or three adjacent frames [6]. They are highly adaptive to scene changes but fail to extract full foreground regions when foreground objects have uniform texture or move.

On the other hand, motion estimation computes the parameters of some motion model based on some low-level motion feature. The most popular low-level motion feature is optical flow. Many techniques have been proposed for the estimation of optical flow, but they all present diverse kinds of problems [7][8]. Allowed motions are usually restricted to some specific model, such as translational or affine motion. Moreover, motion in homogeneous regions is not detected by optical flow. In general, most optical flow estimation techniques assume brightness constancy along frames, which in real situations does not always hold. Particularly, differential methods for optical flow estimation that are consistent with the brightness constancy assumption are not very robust to noise and aliasing. Finally, most optical flow estimation techniques present strong limitations in the allowed displacements of the objects from one frame to the next one. Differential methods try to find the position of a pixel in the next frame, but search is restricted to a small neighbourhood. This limitation can be overcome by coarse-to-fine analysis or by imposing smoothness constraints [7]. Still, large displacements are usually problematic, as well as occlusions.

Sato and Aggarwal [9] have proposed an original approach for motion estimation based on the representation of linear motion patterns using the so-called Temporal Spatio-Velocity (TSV) transform. It consists of a Hough transform evaluated over windowed spatio-temporal images. Segmentation is accomplished by simple thresholding of the TSV image. Each resulting blob represents a local linear motion pattern. Conversely, the TSV transform has proved to be very

robust to noise and to easily deal with large displacements and occlusions. Its main drawback is that it is limited to translational motion with constant velocity.

Here, we have paid attention to an early approach to low level motion representation based on the concept of energy filtering [10][11][12][13][14]. It consists in the estimation of motion from the amplitude of the responses of spatio-temporal filter pairs in quadrature, tuned to different scales and orientations. Spatio-temporal orientation sensitivity is translated into sensitivity to spatial orientation, speed, and direction of motion. These techniques are known to be robust to noise and aliasing, to give confident measurements of velocity, and to allow an easy treatment of the aperture problem, i.e., the reliable estimation of the direction of motion. However, to the best of our knowledge there is no motion segmentation method based on energy filtering.

Energy filtering performs a detection of planar structures in a spatio-temporal domain. Hence, it is straightforward to correlate pixels of a moving object from different frames, avoiding problems with occlusions and large displacements. Moreover, energy filtering is robust to noise and aliasing. Besides, energy features are suitable for texture description in MPEG-7 video sequences for content-based retrieval applications [15]. The main limitation of energy features for motion representation is being restricted to translational motion with constant velocity. In this paper, we propose the use of composite energy features in bottom-up motion segmentation to cope with this inconvenience.

Composite energy features are clusters of energy features, this is, the responses to band-pass pairs of filters in quadrature, each of them sensitive to a different set of scale, orientation, direction of motion, and speed. Energy features are grouped together in order to identify and reconstruct independent motion patterns in a video sequence. This means that the elements to be clustered are whole features, not pixels. A composite energy feature is built as a combination individual energy features in a cluster. Therefore, it inherits the desirable properties of energy features but providing a more complete representation of motion patterns. Composite energy features have proved to be a powerful tool for the representation of visually independent spatial patterns in 2D data [16], volumetric data [17][18], and video sequences [19].

To identify relevant composite features in a sequence, it is necessary to define an integration

criterion able to relate band-pass energy features contributing to the same motion pattern. Earlier approaches [16][19] use complex statistical measures with high computational cost. In previous works [17][18][20], we have introduced an integration criterion that improves computational cost and performance. Instead of applying arbitrary statistical rules, our criterion is inspired in biological vision. It is based on the hypothesis of Morrone and Owens [21] that the Human Visual System (HVS) perceives features at points of locally maximal Phase Congruence (PC). PC is the measure of the local degree of alignment of the local phase of Fourier components of a signal. The sensitivity of the HVS to PC has been studied by other authors as well [14][22][23][24]. Here, we have extended this concept to spatio-temporal signals to define our criterion for clustering of spatio-temporal energy features.

We will show that composite features clustered under this criterion represent visually independent motion patterns in a video sequence, with different velocity, direction, and/or scale content, and that it performs motion estimation without the imposition of any motion model. We will see that, unlike many of the previously cited method, this representation is robust to noise and illumination changes. Besides, it naturally correlates information from different frames, so that it easily deals with occlusions and large inter-frame displacements, while common low level features need the aid of tracking techniques to find the correspondence of the position of an object from different frames.

We have applied the composite feature representation as the basis of a method for segmentation and tracking. Typical bottom-up methods for motion tracking include active models [25][26], Bayesian region classification [27][28], and Kalman filtering [29]. A comprehensive survey can be found in [3]. We have chosen a geodesic active model [25] for tracking. We apply our representation for both the initialization of the model at each frame and as a low level feature to guide the evolution of the model.

The outline of this chapter is as follows. In section 2, we describe in detail the composite feature detection process. Section 3 explains how the geodesic active model uses the low level motion representation to guide segmentation. In section 4, we illustrate the behaviour of the composite feature representation in different problematic situations, including some standard video

sequences. We will also show the results of segmentation and tracking. In section 5, we expound the conclusions of the work.

## 2. Composite-Feature Detector Synthesis

As aforementioned, composite energy features are clusters of energy features. Then, in the first place, we must perform a multiresolution decomposition to obtain the individual energy features. To this end, we apply a bank of non-causal spatio-temporal energy filters to the video sequence. The complex-valued volume generated as the response of a spatio-temporal energy filter to a given video sequence is here called a *band-pass energy feature*. We will call *composite energy features* to motion patterns with multiple speed, direction and scale contents generated as the combination of band-pass energy features in a cluster. The set of filters associated to an energy feature cluster are referred to as *composite-feature detector*.

Feature grouping is accomplished by applying cluster analysis to the set of band-pass features of the video sequence. We perform clustering of the features as a whole, not point wise, since a point in space can be occupied by several visual patterns with different frequency content. Hence, more that one composite feature can have large response in the same location in space and time. This is not allowed by crisp clustering of pixels. Besides, pixel clustering is more sensitive to local variations and noise.

Clustering is accomplished using a hierarchical algorithm, which is based on a dissimilarity matrix reflecting the distances among band-pass features. The dissimilarity measure is defined in order to identify band-pass features contributing to the same local maxima of Phase Congruence (PC). We will see that the correlation coefficient can provide a good global measure of the PC between band-pass features [17][20].

Finally, each composite feature is reconstructed as a combination of the responses of the filters in a given cluster. Instead of a simple linear combination, we propose a more sophisticated pooling expression, in order to enhance the representation. The next subsections present the details of the process.

### 2.1 Bank of Spatio-Temporal Filters

The basis function of the bank of spatio-temporal filters applied here [17][20] is an extension to 3D of the log Gabor function [30]. The filter is designed in the frequency domain, since it has no analytical expression in the spatial domain. Filtering is realized as the inner product between the transfer function of the filter and the Fourier transform of the sequence. Filtering in the Fourier domain is very fast when using Fast Fourier Transform and Inverse Fast Fourier Transform algorithms.

The filters' transfer function $T$ is designed in spherical frequency coordinates as the product of separable factors, $R$ and $S$, in the radial and angular components respectively, such that $T = R \cdot S$. The radial term $R$ is given by the log Gabor function [30]

$$R(\rho;\rho_i) = \exp\left( -\frac{\left(\log\left(\rho/\rho_i\right)\right)^2}{2\left(\log\left(\sigma_{\rho i}/\rho_i\right)\right)^2} \right), \tag{1}$$

where $\sigma_{\rho i}$ is the standard deviation and $\rho_i$ is the central radial frequency of the filter.

The angular component is designed to achieve orientation selectivity in both the azimuthal component $\phi_i$ of the filter, which reflects the spatial orientation of the pattern in a frame and the direction of movement, and the elevation component $\theta_i$, related to the speed and direction of motion. For static patterns, $\theta_i = 0$. To achieve rotational symmetry, $S$ is defined as a Gaussian on the angular distance $\alpha$ between the position vector $\mathbf{f}$ of a given point in the spectral domain and the direction of the filter $\mathbf{v} = (\cos\phi_i \cdot \cos\theta_i, \cos\phi_i \cdot \sin\theta_i, \sin\phi_i)$ [31]

$$S(\phi,\theta;\phi_i,\theta_i) = S(\alpha) = \exp\left( -\frac{\alpha^2}{2\sigma_\alpha^2} \right), \quad \text{with } \alpha(\phi_i,\theta_i) = \operatorname{acos}\left(\mathbf{f} \cdot \mathbf{v} / \| \mathbf{f} \|\right), \tag{2}$$

where $\mathbf{f}$ is expressed in Cartesian coordinates and $\sigma_{\alpha i}$ is the angular standard deviation.

Active filters are selected from a predefined band partitioning of the 3D spectral domain. Frequency bands are described by the central frequency ($\rho_i, \phi_i, \theta_i$) of the filters and their width parameters ($\sigma_{\rho i}, \sigma_{\alpha i}$). The selection of these parameters determines the kind of energy features yield by the multiresolution decomposition. In this application we desire to get wide coverage of

the scale space, uniform sampling of the orientation space, and high orientation sensitivity. To this end, the parameters of the bank are established as follows.

Frequency is sampled so that $\rho_i$={1/2, 1/4, 1/8, 1/16}, in pixels$^{-1}$. Parameter $\sigma_{\rho i}$ is determined for each band in order to obtain a 2 octave bandwidth. $\theta_i$ is sampled uniformly while the number of $\phi_i$ samples decreases with elevation. $\phi_i$ are determined in order to keep a constant "density" of filters, by maintaining equal arc-length between adjacent $\phi_i$ samples over the unit radius sphere. $\sigma_{\alpha i}$ is set to 25º for all orientations. Following this criterion, the filter bank has been designed using 23 orientations, i.e. ($\phi_i$,$\theta_i$) pairs. In total, we have 23 orientation × 4 scales = 92 bands with a certain overlapping between neighbouring bands, yielding a redundant decomposition and a wide coverage of the spectrum.

**2.2 Selection of Active Bands**

To achieve improved performance, it is convenient to reduce the number of bands involved in cluster analysis. The exclusion of frequency channels that are not likely to contribute to motion patterns facilitates the identification of clusters associated to composite motion features. Furthermore, it reduces computational cost. Features selected for clustering are called *active*. A previous solution for active band selection, applied in [17], consisted in the detection of bands enclosing spectral amplitude values over a maximum noise level. This implies the setting of a threshold parameter and the use of a radial median filter, which is highly computationally expensive. Here, we have introduced a channel selection stage based on a statistical analysis of the amplitude responses of the band-pass features. This solution outperforms that in [17] in both computational cost and efficiency.

Our method for the selection of active channels is based on the works of Field [32] and Nestares et al. [33]. Field has studied the statistics of the responses of a multiresolution decomposition based on log-Gabor wavelets that resembles the coding in the visual system of mammalians. He has observed that the histograms of the filter responses are not Gaussian, but leptokurtic distributions –pointed distributions with long tails–, revealing the sparse nature of both the

sensory coding and the features from natural images. According to Field, when the parameters of the wavelet codification fit those in the mammalian visual system, the histogram of the responses is highly leptokurtic. This is reflected in the fourth cumulant of the distribution. Namely, he uses the kurtosis to characterize the sparseness of the response.

Regarding spatio-temporal analysis, Nestares et al. [33] applied channel selection to a bank of third order Gaussian derivative filters based on the statistics of filters responses. They have observed that features corresponding to mobile targets present sparser responses than those associated to background –weather static or moving. This fact is illustrated in Fig. 1. They measure different statistical magnitudes reflecting sparseness of the amplitude response, realize a ranking of the channels based on such measures, and perform channel selection by taking the $n$ first channels in the ranking, where $n$ is a prefixed number.

Based on these two works, we have designed our filter selection method. The statistical measure employed to characterize each channel is the kurtosis excess $\gamma_2$

$$\gamma_2 = k_4 / k_2^2 - 3, \tag{3}$$

where $k_4$ and $k_2$ are respectively the fourth and second cumulants of a histogram. If the kurtosis excess takes a positive value, the distribution is called leptokurtic and presents a narrow peak and long tails. If it is negative, the distribution is called platykurtic and presents a broad central lobe and short tails. Distributions with zero kurtosis excess, like the Gaussian distribution, are called mesokurtic.

We measure $\gamma_2$ for both the real and imaginary components of each feature $\psi_i$ and then we compose a single measure $\delta$

$$\delta_i = \gamma_2\left(\mathrm{Re}(\psi_i)\right) + \gamma_2\left(\mathrm{Im}(\psi_i)\right) \tag{4}$$

Instead of selecting a fixed number of channels with the larges values of $\delta$, we perform cluster analysis to identify two clusters, one for active channels, with large values of $\delta$, and another for non active channels. Here, we have applied a k-means algorithm. The cluster of active channels is identified as the one with higher $\delta$ on average.

### 2.3 Energy Feature Clustering

As aforementioned, it seems plausible that the visual system of humans perceives features where Fourier components are locally in phase [21]. Morrone and Owens [21] defined the concept of Phase Congruence (PC) of a signal as a measure of the local degree of alignment of the local phase of Fourier components. Venkatesh and Owens [34] demonstrated that points of locally maximal PC present maxima in the local energy. Hence, local energy can be used as a bio-inspired low level feature in vision tasks as well as PC.

When applying multiresolution decomposition, the frequency content of a composite feature is distributed in a subset of frequency bands, i.e., a maximum in the energy of a signal is decomposed in smaller maxima of a subset of band-pass features. This implies that, band-pass features contributing to the same visual pattern should present a large degree of alignment in their local energy maxima –see Fig. 2. Our goal is to identify and reconstruct a specific composite feature by identifying the set of band-pass features that carry part of the energy of its maxima. Band-pass features are grouped if their energy maxima are coincident in space and time. To this end, we will employ a global measure that, given a pair of band-pass energy features, provides an estimation of the overall degree of coincidence of their energy maxima. Based on a study over several possible measures [17][20], we have chosen the correlation coefficient $\rho$, because of its simplicity and good performance. If we denote the amplitude response of a band-pass feature $\psi_i$ as $A_i=A(\psi_i)=\|\psi_i\|=(\mathrm{Im}(\psi_i)^2+\mathrm{Re}(\psi_i)^2)^{1/2}$, then the actual distance between two band-pass features $\{\psi_i, \psi_j\}$ is calculated from $\rho(A_i,A_j)$ as follows

$$D_\rho(\psi_i,\psi_j) = \left(1-\sqrt{\left(1+\rho(A_i,A_j)\right)/2}\right)^2 . \tag{5}$$

This distance function takes values in the range [0,1]. Zero distance corresponds to exact match among maxima, i.e., a linear dependence with positive slope. Maximal distance corresponds to linear dependence with negative slope, like in the case of a image and its inverse. This measure does not depend on the selection of any parameter, and does not involve the discrete estimation of joint and/or marginal probabilities –histograms.

The computational cost of the whole composite feature computation is dominated by the

estimation of distances between pairs of energy features. Using the distance in equation (5), and supposing an input sequence of size $N$ and a number $F$ of energy features –or number of active filters–, the asymptotical computational cost of this stage is $O(N \cdot F^2)$ –see references [17] or [20].

Our technique detects visual patterns by clustering of active bands. To this end, dissimilarities between each pair of energy features are computed to build a dissimilarity matrix. To determine the clusters from the dissimilarity matrix, a hierarchical clustering method is applied. Among the different algorithms for hierarchical clustering [35], we have chosen the Ward's algorithm, since it which has proved to stand out from other metrics [35]. It is characterized by employing an inter-cluster distance measure that produces minimum variance partitions. The number of clusters $N_c$ produced is an input parameter for the hierarchical algorithm. The usual strategy to determine the $N_c$ is to run the algorithm for each possible $N_c$ and to evaluate the quality of each resulting partition according to a given validity index. Our validity index is a modification of the Davies-Boulding index [36], which is a graph-theory based index that measures the compactness of the clusters in relation to their separation.

A stage of cluster merging follows cluster analysis. Clusters with average inter-cluster correlation over a given threshold are merged to form a single cluster. This is made because hierarchical algorithms can only analyse the magnitude of a distance in relation to others, not in an absolute fashion. This fact is often the cause of a wrong classification, due to the splitting of clusters into smaller subgroups. We have defined a correlation threshold of 0.75 to merge clusters.

### 2.4 Composite Feature Reconstruction

The response $\psi$ of an energy filter is a complex-valued sequence, where the real and imaginary components account for even and odd symmetric features respectively. In this section we describe how band-pass complex-valued features in a cluster are combined to produce a composite-feature $\Psi$. We define three different representations, obtained from real, imaginary or amplitude components, which are used in different situations. For simple visualization

purposes, we reconstruct motion patters using the real components. If we are only interested on odd-symmetric components to represent mobile contours, we will use only the imaginary component of band-pass features. To estimate the energy of the composite feature, we use the amplitudes of the band-pass features $||\psi||$.

Here we define the general rule for the reconstruction of $\Psi$ based on a given representation $E$ of the responses of the filters, which can be either $\mathrm{Re}(\psi)$, $\mathrm{Im}(\psi)$ or the amplitude $A(\psi)=||\psi||$ $=(\mathrm{Im}(\psi)^2+\mathrm{Re}(\psi)^2)^{1/2}$. The easiest way of constructing the response $\Psi$ of a set $\Omega_j$ of filters in a cluster $j$ is linear summation

$$\Psi^j(x,y,t)=\sum_{i\in\Omega_j} E_i(x,y,t) \quad (6)$$

However, simple summation presents one important problem. There might be features in the cluster that contribute, not only to the corresponding motion pattern, but also to other patterns or static structures in the sequence. Only points with contributions from all features in the cluster should have a non null response to the composite feature detector. To avoid this effect, linear summation is modulated by a mask that weights locations with contributions from all features in the clustering more heavily. The mask represents the portion of features contributing to energy at a given location. It is computed as the number of features over a given threshold in relation to the total number o features in the cluster. If $\widetilde{E}_i \in [0,1]$ is the result of thresholding $E_i$ using a sigmoid function, then

$$\Psi^j(x,y,t)=\frac{\sum_{i\in\Omega_j}\widetilde{E}_i}{Card(\Omega_j)}\sum_{i\in\Omega_j} E_i\,, \quad (7)$$

where $\Omega_j$ is the set of all bands in cluster $j$. The effect of masking is illustrated in Fig. 3. Reconstruction using different representations for $E$ is illustrated in Fig. 4. For visualization purposes, we employ the real component $\Psi^j_{even}=\Psi^j(E_i=\mathrm{Re}(\psi_i))$. The odd-symmetric representation of $\Psi$ is constructed by full-wave rectification of the expression in equation (7), so that $\Psi^j_{odd}=\left|\Psi^j(E_i=\mathrm{Im}(\psi_i))\right|$ does not have into account the sign of the contour. Energy

estimation can be accomplished using the amplitude representation $\Psi^{j}_{amp} = \Psi^{j}\left(E_i = \|\psi_i\|\right)$.

## 3. Motion segmentation using composite features

In a bottom-up strategy, high level modelling usually implies constraints over continuity, smoothing or shape parameters. One of the most popular modelling techniques able to introduce such constraints is the active model [37]. In this work we have chosen a geodesic active model [38] to obtain a high level representation from our composite features. Each frame of a sequence is segmented independently. We perform a segmentation process for each composite feature, which we denote be $\Psi$, omitting the superindex. Fig. 5 presents a scheme of the segmentation method for a given pattern and a given frame. First of all, an initial model and an image potential have to be defined. Initialization involves sigmoid thresholding of the amplitude representation of the pattern $\Psi_{amp}$. Image potential is defined from a combination of the spatio-temporal feature $\Psi_{odd}$ and pure spatial contours. After evolving a geodesic model in each frame, the segmented sequence is generated by stacking the segmented frames.

The geodesic active model represents an object as a contour in a 2D+*t* domain. The evolution of the contour is determined from the evolution of the zero-level set of an implicit function representing the distance *u* to the contour. The equations governing the evolution of the implicit function are the following. If $\Omega := [0, a_x] \times [0, a_y]$ is the frame domain and $u_0(x,y)$ is a scalar image on $\Omega$ representing the initial state of the model, then

$$\begin{aligned} u\left(x,y,t=t_k,\tau=0\right) &= u_0(x,y,t=t_k) \quad \text{on } \Omega \\ \frac{\partial u}{\partial \tau} &= g\left(s\right)\left|\nabla u\right|\kappa + \nabla g\left(s\right)\nabla u \quad \text{on } \Omega\times(0,\infty) \end{aligned}, \tag{8}$$

where symbol $\tau$ stands for time in the evolution equations of *u* to distinguish it from the frame index *t*, *g* is a function with values in the interval [0, 1] that decreases in the presence of image features, *s* is the selected image feature and $\kappa$ is the curvature of the model. The actual implementation of the geodesic active model used here is the one described in [39].

According to these equations, the zero-level set contour is attracted to minima of the image potential function *g*(*s*). In order that the contour evolves to the target object, one of the

conditions that must be fulfilled is that the image feature $s$ properly represents the target, showing local maxima in the location of the target. The other condition must be either that $s$ does not have maxima for non target structures or that the initial state $u_0$ is closer to the target object than to any other.

As has been mentioned, one of the characteristics of the composite energy feature representation is that it allows for the isolation of the motion patterns that are visually independent. This makes it suitable as an image feature $s$ for the geodesic active model. Besides, it can be used for the definition of the initial state of the model. Since the composite feature is already a good approach to the shape of the target, the correct evolution is further ensured, and moreover, convergence is accelerated. In the next subsections we describe how we apply the composite feature representation to define the terms $s$ and $u_0$ in equation (8).

### 3.2 Image Potential

The expression for the image potential function is that in [39]

$$g(s) = \frac{1}{1 + (s/s_{\min})^2}, \tag{9}$$

with $s_{\min}$ being a real constant. The image feature $s$ is here defined from the spatio-temporal composite feature $\Psi_{odd}$, built as described in section 2.4. This motion pattern may present artefacts, due to the diffusion of patterns from neighbouring frames produced when applying energy filtering. This situation is illustrated in Fig. 6.a and b. To minimize the influence of these artefacts, this motion pattern is modulated by a factor representing the localization of spatial contours. It is calculated from the 2D contour detector response by thresholding using a sigmoid function. Hence, $s$ is defined from the following spatio-temporal image feature

$$C_m(x, y, t_k) = \frac{1}{1 + \exp(-K(C_s(x, y, t_k) - C_0))} \frac{\Psi_{odd}(x, y, t_k)}{\max(\Psi_{odd}(x, y, t_k))}, \tag{10}$$

where $C_s$ is a spatial contour detector based on the frame gradient, $C_0$ is the gradient threshold and $K$ is a positive real constant. The effect of this modulation can be observed in Fig. 6.c and d.

Although here we are interested in segmenting objects based on their motion features, it is convenient to include a spatial term in the image feature. This is necessary to close a contour when part of the boundary of the moving object remains static –when there is a partial occlusion by a static object or scene boundary or when part of the moving contour is parallel to the direction of motion. Therefore, the image feature $s$ is the weighted sum of two terms, $C_m$ and $C_s$, respectively related to spatio-temporal and pure spatial information.

$$s = w_S C_S + w_m C_m \text{, with } w_S + w_m = 1 \text{ and } w_S, w_m > 0 \tag{11}$$

The weight of the spatial term $w_s$ must be much smaller than the motion term weight $w_m$, so that the active model does not get "hooked" on static contours not belonging to the target object. The spatial feature employed to define the spatial feature is the regularized image gradient. Regularization of a frame is accomplished here by feature-preserving 2D anisotropic diffusion filtering, which brakes diffusion in the presence of contours and corners. The 3D version of the filter is described in [40]. If $I^*(x,y,t_k)$ is the smoothed version of the $k^{\text{th}}$ frame, then

$$C_S(x,y,t_k) = \left\|\nabla I^*(x,y,t_k)\right\| \Big/ \max\left(\left\|\nabla I^*(x,y,t_k)\right\|\right) \tag{12}$$

The selection of parameters in previous equations and in section 3.3 is not critical for the efficiency of the method. The values chosen here have been selected *ad hoc* and have proved to produce satisfactory results. The specific values are $C_0 = 0.1$, $K = 20$, $w_s = 0.1$, and $w_m = 0.9$. In equation (9), $s_{\min}$ is calculated so that, on average, $g(s(x,y)) = 0.001$, $\forall x,y : C_m(x,y) > 0.1$. Considering the geodesic active model in a front propagation framework, $g = 0.001$ means a sufficiently slow speed of the propagating front to produce stopping in practical situations.

### 3.3 Initialization

The initial state of the geodesic active model is defined, in general, from the amplitude representation $\Psi_{amp}$ of the selected motion pattern. The even-symmetric representation can be used for initialization of objects with uniform contrast and is defined by applying a half-wave rectification, replacing $\Psi_{amp}$ with $\max(\pm\Psi^j_{even}, 0)$. Sign will depend on the specific contrast. To enhance the response of the cluster, we apply a sigmoid thresholding to $\Psi_{amp}$. The result is

remapped to the interval [–1,1]. The zero-level of the resulting image is the initial state of the contour.

$$u_0(x, y, t_k) = \frac{2}{1 + \exp\left(-K\left(\Psi_{amp}(x, y, t_k) - \Psi_0\right)\right)} - 1 \tag{13}$$

When the object remains static during a number of frames the visual pattern has a null response. For this reason, the initial model is defined as the weighted sum of two terms, respectively associated to the current and previous frames. The contribution from the previous frame must be very small.

$$u_0(x, y, t_k) = w_k \left( \frac{2}{1 + \exp\left(-K\left(\Psi_{amp}(x, y, t_k) - \Psi_0\right)\right)} - 1 \right) + w_{k-1}\, u_{\tau=\tau_{\max}}(x, y, t_{k-1}) \tag{14}$$

with $w_k$ and $w_{k-1}$ being positive real constants that verify $w_k + w_{k-1} = 1$. In the experiments presented in next section, $w_k = 0.9$, $w_{k-1} = 0.1$, $K = 20$ and $\Psi_0 = 0.1$.

## 4 Results

In this section, some results are presented to show the behaviour of the method in problematic situations. The complete video sequences with the original data and the segmentation results are available at http://www-gva.dec.usc.es/~rdosil/motion_segmentation_examples.htm.They are summarized in the next subsections.

**Example #1**

The following example shows the ability of the method to deal with complex motion patterns with variations in illumination, non linear motion and deformations. In particular, the following sequence, a fragment of 27 frames of the standard movie know as “silent” – top row of Fig. 7–, presents a mobile object, with variable shape, speed and direction. As can be appreciated, the motion pattern of the hand can not be properly described by an affine transformation. Moreover, the brightness constancy assumption is not verified in this case.

The middle row of Fig. 7 shows one of the two identified composite patterns, representing the

moving hand. It can be seen that, despite the complexity of the image, the composite-feature representation model is able to isolate the hand and properly represent its changing shape in different frames. Images in the bottom row of Fig. 7 present the segmentation results.

**Example #2**

In this example, we use a fragment of 27 frames of the well-known sequence "flower garden" – see Fig. 9. It is a static scene recorded by a moving camera, so that all the pixels in the scene are moving. The composite energy feature representation identifies two motion patterns, one for the moving tree and one for the moving background. This is possible because the tree is placed in a foreground plane, so that its speed is larger. When visualizing a cut in the *x*-*t* plane on both the image and the motion patterns –Fig. 8–, it can be seen that different speeds are translated into different orientations in the spatio-temporal domain. The composite energy feature representation has been able to cluster energy features with similar frequencies –speeds, orientations and sizes.

The result of segmentation using one of the detected motion patterns is shown in Fig. 9. The image potential in that example is computed from the composite energy feature corresponding to the tree. Hence, there are not deep minima of the potential caused by moving background contours, which leads to a correct segmentation.

**Example #3**

In this example we use a fragment of 22 frames of the well-known "table tennis" video sequence, shown in first row of Fig. 10, which presents non linear motion patterns and large inter-frame displacements, due to a small sampling rate. This makes it difficult to find the correspondence between the positions of the object in two consecutive frames.

The composite feature representation is able to identify the energy components of the motion pattern, and to generate a composite feature in which the motion pattern is isolated from the static background –see Fig. 10, middle row. The result of segmentation is shown in bottom row of Fig. 10. Since the active model is initialized from the motion pattern at each frame, the large

inter-frame displacements do not influence tracking.

**Example #4**

Occlusions give rise to the same problem as with fast objects. In the next example, in Fig. 11, a sequence of 31 frames is showing a cylinder rolling behind another object, completely disappearing from the scene during some frames. The middle row in Fig. 11 shows one of the composite features identified by our representation method, corresponding to the moving cylinder, in its amplitude representation. The other composite feature detected, not shown here, corresponds the static objects.

The result of segmentation using this composite feature is shown in last row of Fig. 11. As can be seen, initialization with the composite energy feature leads to a correct segmentation, even when the object disappears from the scene during several frames. The model collapses when the cylinder disappears behind a static object and is reinitialized automatically when it reappears, without the need of a prior model.

**Example #5**

The next example, in Fig. 12, with 39 frames, presents occlusions as well, but now the occluding object is also moving, which complicates tracking. The scene shows two people walking in opposite directions. The second and third rows in Fig. 12 show two of the composite features identified by our representation method, corresponding to the two moving persons. The remainder composite features detected, not shown here, correspond to static objects. Again, the representation model is capable of decomposing the scene into visually independent motion patterns by integrating band-pass energy features.

Two active models have been optimized, each guided by one of these composite features. Segmentation results are presented in the fourth and fifth rows of Fig. 12. The use of composite features for both initialization and definition of image potential avoids interferences between different motion patterns during model optimization. Besides, tracking is straightforward since there is only one model for each pattern in each frame, although it may be split.

**Example #6**

The following example is a fragment of the standard video sequence “coast guard”, comprising 48 frames. This sequence shows two mobile objects, one of them being followed by the camera. The results for this sequence are presented in Fig. 13. The model classifies the filters corresponding to the moving background as non active, excluding them from the subsequent analysis. The model is able to identify two motion patterns, one for each of the moving objects. Segmentation of the static object is not interfered by background texture or background mobile contours.

**Example #7**

The last example, in Fig. 14.a and e, is a sequence of 126 frames showing two motion patterns with different scales, speeds and directions of motion. Both patterns are occluded in different parts of the sequence. The background is not completely static, but there are also local motion patterns –branches in the trees are moving due to the wind. It is a low resolution video, with a high noise level.

The proposed model classifies the filters contributing to the background as non active, including local motion in the branches of the trees –see Fig. 15.b– and active filters give place to two motion patterns, one for each moving object –Fig. 15.c, d, f and g. Segmentation results for both composite features are shown in Fig. 15. Initialization with composite features solves the problem of occlusions in both situations. Image potentials from composite features avoid interferences between the two motion patterns and interferences with motion in background pixels.

## 5 Discussion and conclusions

Results show that the proposed model for the representation of motion is able to group band-pass features associated to different motion patterns in a scene without the use of prior

information, and to isolate visually independent motion patterns. The studied examples cover some of the most common problems in motion estimation, namely, presence of noise, moving background, variations in illumination, non affine motion patterns, scenes with multiple motion patterns, occlusions, and large displacements between neighbouring frames. As summarized in the introduction, other common low level motion representations present difficulties in some of these situations.

The key characteristic of the composite feature representation is that integration is accomplished by clustering of spatio-temporal energy features as a whole, not by point-wise region classification. This fact yields a representation that intrinsically correlates information from different frames. This property is responsible for the robustness to partial and total occlusions and large inter-frame displacements or deformations. This property allows the direct use of this representation in guiding a high level segmentation technique like the active model, without any intermediate step of region classification or object search from one frame to the next. Segmentation results prove the success of the combination of composite feature representation and high level modelling by a level-set approach. The problem of the initialization of active models is easily solved using the amplitude representation of composite features. Furthermore, convergence of the model to the target object is ensured by building the image potential from features that discriminate the target object from other structures and moving objects in the scene. In future, we plan to tackle more complex cases by applying the composite feature representation to aid top-down motion analysis.

## 6. Acknowledgements

This work has been financially supported by the Ministry of Education and Science of the Spanish Government, through the research project TIN2006-08447.

## 7. References

[1]. H.T. Nguyen, A.W.M. Smeulders, Fast Occluded Object Tracking by a Robust

Appearance Filter, IEEE Trans. Pattern Anal. Mach. Intell., 26 (2004) 1099-1104.

[2]. C. Kervrann, F. Heitz, A Hierarchical Markov Modeling Approach for the Segmentation and Tracking of Deformable Shapes, Graphical Models and Image Processing, 60 (1995) 173-195.

[3]. W. Hu, T. Tan, L. Wang, S. Maybank. A Survey on Visual Surveillance of Object Motion and Behaviors, IEEE Trans on Systems, Man and Cybernetics –Part C: Applications and Reviews, 34 (2004) 334-351.

[4]. C. Stauffer, W. Grimson, Adaptive Background Mixture Models for Real-Time Tracking, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, Vol 2., 1999, 246-254.

[5]. L.J. Latecki, V. Megalooikonomou, R. Miezianko, D. Pokrajac, Using Spatiotemporal Blocks to Reduce the Uncertainty in Detecting and Tracking Moving Objects in Video, Intelligent Systems Technologies and Applications, 1 (2006) 376-392.

[6]. A.J. Lipton, H. Fujiyoshi, R.S. Patil, Moving Target Classification and Tracking from Real-Time Video, in Proc. IEEE Workshop Applications of Computer Vision, 1998, 8-14.

[7]. J.L. Barron, D.J. Fleet, S.S. Beauchemin, Performance of Optical Flow Techniques, Int J Comput Vis, 12 (1994) 43-77.

[8]. C. Stiller, J. Konrad, Estimating Motion in Image Sequences: A Tutorial on Modeling and Computation of 2D Motion, IEEE Signal Processing Magazine, 16 (1999) July 71-91.

[9]. K. Sato, J.K. Aggarwal, Temporal Spatio-Velocity Transform and its Application to Tracking and Interaction, Computer Vision and Image Understanding, 96 (2004) 100-128.

[10]. D.J. Heeger, Model for the Extraction of Image Flow, J. Opt. Soc. Am. A, 4 (1987) 1555-1471.

[11]. E.P. Simoncelli, E.H. Adelson, Computing Optical Flow Distributions using Spatio-Temporal Filters, MIT Media Lab. Vision and Modeling, Tech. Report 165, 1991. URL: http://web.mit.edu/persci/people/adelson/pub_pdfs/simoncelli_comput.pdf

[12]. A.B. Watson, A.J. Ahumada Jr., Model for Human Visual-Motion Sensing, J. Opt. Soc. Am. A, 2 (1985) 322-342.

[13]. E.H. Adelson, J.R. Bergen, Spatiotemporal Energy Models for the Perception of Motion,

J. Opt. Soc. Am. A,2 (1985) 284-299.

[14]. D. Fleet, Measurement of Image Velocity, Kluwer Academic Publishers, Massachusetts, 1992.

[15]. Y.M. Ro, M. Kim, H.K. Kang, B.S. Manjunath, J. Kim, MPEG-7 Homogeneous Texture Descriptor, ETRI Journal, 23 (2001) 41-51.

[16]. R. Rodríguez-Sánchez, J.A. García, J. Fdez-Valdivia, X.R. Fdez-Vidal, The RGFF Representational Model: A System for the Automatically Learned Partition of "Visual Patterns" in Digital Images, IEEE Trans. Pattern Anal. Mach. Intell, 21(1999) 1044-1073.

[17]. R. Dosil, Data Driven Detection of Composite Feature Detectors for 3D Image Analysis, PhD Thesis, Universidade de Santiago de Compostela (Spain), 2005. URL: http://www-gva.dec.usc.es/~rdosil/ficheiros/thesis_dosil.pdf

[18]. R. Dosil, X.M. Pardo, X.R. Fdez-Vidal, Decomposition of 3D Medical Images into Visual Patterns, IEEE Trans. on Biomedical Engineering, 52 (2005) 2115-2118.

[19]. J. Chamorro-Martínez, J. Fdez-Valdivia, J.A. García, J. Martínez-Baena, A Frequency Domain Approach for the Extraction of Motion Patterns, IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 3, Hong Kong, 2003, pp. 165-168.

[20]. R. Dosil, X.R. Fdez-Vidal, X.M. Pardo, Dissimilarity Measures for Visual Pattern Partitioning, in: J. Marques, N. Pérez de la Blanca (Eds.), LNCS: Pattern Recognition and Image Analysis, Vol. 3523, Springer-Verlag Berlin Heidelberg, 2005, pp. 287-294.

[21]. M.C. Morrone, R.A. Owens, Feature Detection from Local Energy, Pattern Recognition Letters, 6 (1987) 303-313.

[22]. A. Oppenheim, J. Lim, The Importance of Phase in Signals, Proc. of the IEEE, 69 (1981) 529–541.

[23]. J. Ross, M.C. Morrone, D. Burr, The Conditions under which Mach Bands are Visible, Vision Research, 29 (1989) 699–715.

[24]. J. du Buf, Ramp Edges, Mach Bands and the Functional Significance of the Simple Cell Assembly, Biological Cybernetics, 70 (1994) 449–461.

[25]. N. Paragios, R. Deriche, Geodesic Active Contours and Level Sets for the Detection and

Tracking of Moving Objects, IEEE Trans. Pattern Anal. Mach. Intell., 22 (2000) 266-279.

[26]. A.-R. Mansouri, J. Konrad, Multiple Motion Segmentation with Level Sets, IEEE Trans. on Image Processing, 12 (2003) 201-220.

[27]. M.M. Chang, A.M. Tekalp, M.I. Sezan, Simultaneous Motion Estimation and Segmentation, IEEE Trans. on Image Processing, 6 (1997) 1326-1333.

[28]. R. Montoliu, F. Pla, An Iterative Region-Growing Algorithm for Motion Segmentation and Estimation, Int. Journal of Intelligent Systems, 20 (2005) 577-590.

[29]. Y. Boykov, D.P. Huttenlocher, Adaptive Bayesian Recognition in Tracking Rigid Objects, IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), Vol. 2, 2000, pp. 697-704.

[30]. D.J. Field, What is the Goal of Sensory Coding. Neural Computation, 6 (1994) 559-601.

[31]. F.G.A. Faas, L.J. van Vliet, 3D-Orientation Space; Filters and Sampling, in: J. Bigun, T. Gustavsson (Eds.), LNCS: Scandinavian Conference on Image Analysis, Vol. 2749, Springer-Verlag Berlin Heidelberg, 2003, pp.36-42.

[32]. D.J. Field, Scale–Invariance and Self-Similar "Wavelet" Transforms: An Analysis of Natural Scenes and Mammalian Visual Systems, in: M. Farge, J.C.R. Hunt, J.C. Vassilicos, (Eds.), Wavelets, fractals and Fourier Transforms, Clarendon Press, Oxford, 1993, pp. 151-193.

[33]. O. Nestares, C. Miravet, J. Santamaria, R. Navarro, Automatic Enhancement of Noisy Image Sequences Through Local Spatiotemporal Spectrum Analysis, Optical Engineering, 39 (2000) 1457-1469.

[34]. S. Venkatesh, R. Owens, On the Classification of Image Features, Pattern Recognition Letters, 11 (1990) 339-349.

[35]. A. Jain, R. Dubes, Algorithms for Clustering Data, Prentice Hall, New Jersey, 1988.

[36]. N.R. Pal, J. Biswas, Cluster Validation Using graph Theoretic Concepts, Pattern Recognition, 30 (1996) 847-857.

[37]. M. Kass, A. Witkin, D. Terzopoulos, Snakes: Active Contour Models, Int. Journal of Computer Vision, 55 (1988) 321-331.

[38]. V. Caselles, R. Kimmel, G. Sapiro. Geodesic Active Contours, Int J Comput Vis, 22, (1997) 61-79.

[39]. J. Weickert, G. Kühne, Fast Methods for Implicit Active Contour Models, in: S. Osher, N. Paragios (Eds.), Geometric Level Set Methods in Imaging, Vision and Graphics, Springer, New York, 2003, pp. 43-58.

[40]. R. Dosil, X.M. Pardo, Generalized Ellipsoids and Anisotropic Filtering for Segmentation Improvement in 3D Medical Imaging, Image and Vision Computing, 21 (2003) 325-343.

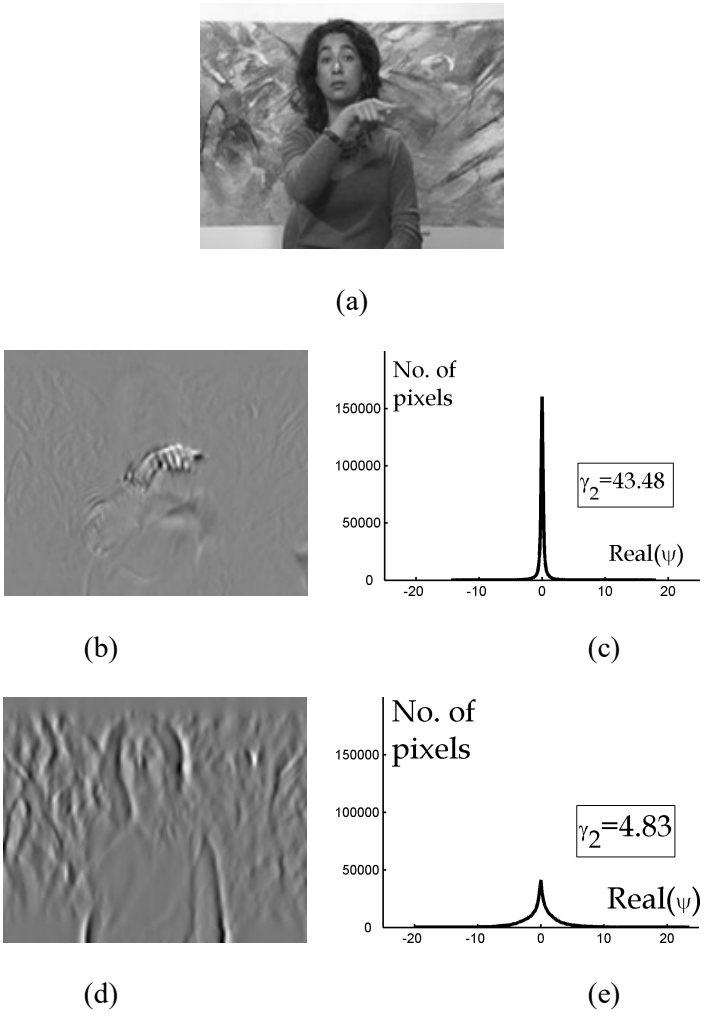
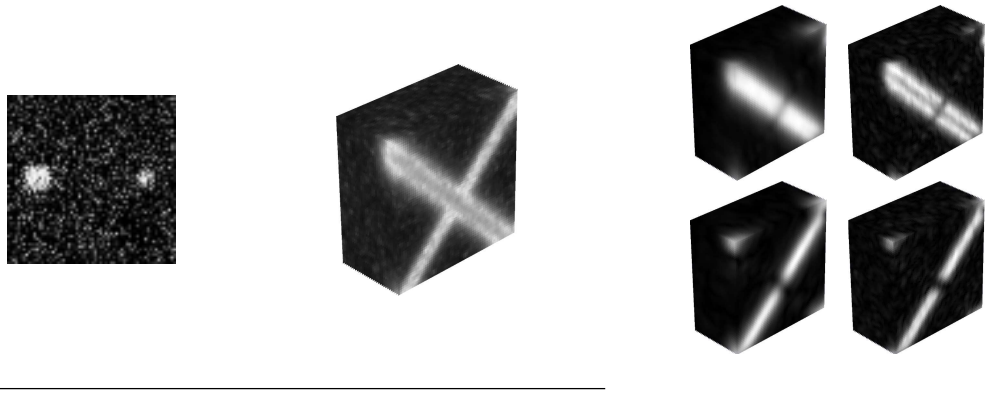



Fig. 1. (a) A frame of the "Silent" standard sequence, showing a moving hand. (b) and (d) A frame of the real part of two band-pass features of the *Silent* video sequence. (c) and (e) Histograms corresponding to band-pass features in (b) and (d) respectively.



(a) (b) (c)

Fig. 2. (a) A frame of a synthetic video sequence, where two light spots move from side to side with opposite direction. (b) A cut perpendicular to the temporal axis of the total energy of the sequence. (c) Energy of some band-pass versions of the sequence. Those on top row correspond to one of the spots and they present some degree of concurrence on their local energy maxima. Bottom row shows two band-pass features contributing to the other motion pattern.

Fig. 3. A frame of the “Silent” video sequence: *Left*: Input data. *Centre*: Even-symmetric representation of the response of one of the detected composite features, corresponding to the moving hand, computed using equation (6) and, *Right*: using equation (7).

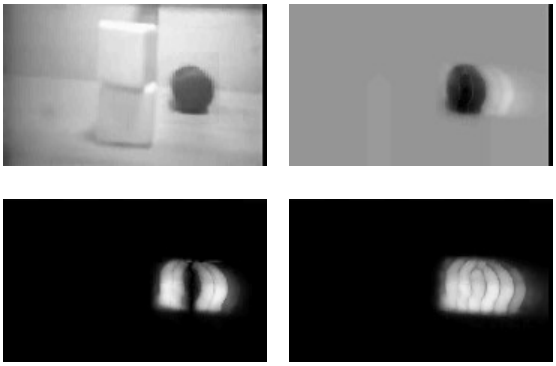


Fig. 4. *Top left*: One frame of an example sequence with a moving dark cylinder. The remainder images show different representations for one of the identified composite features. *Top right*: Even representation. *Bottom left*: Odd representation. *Bottom right*: Amplitude representation.

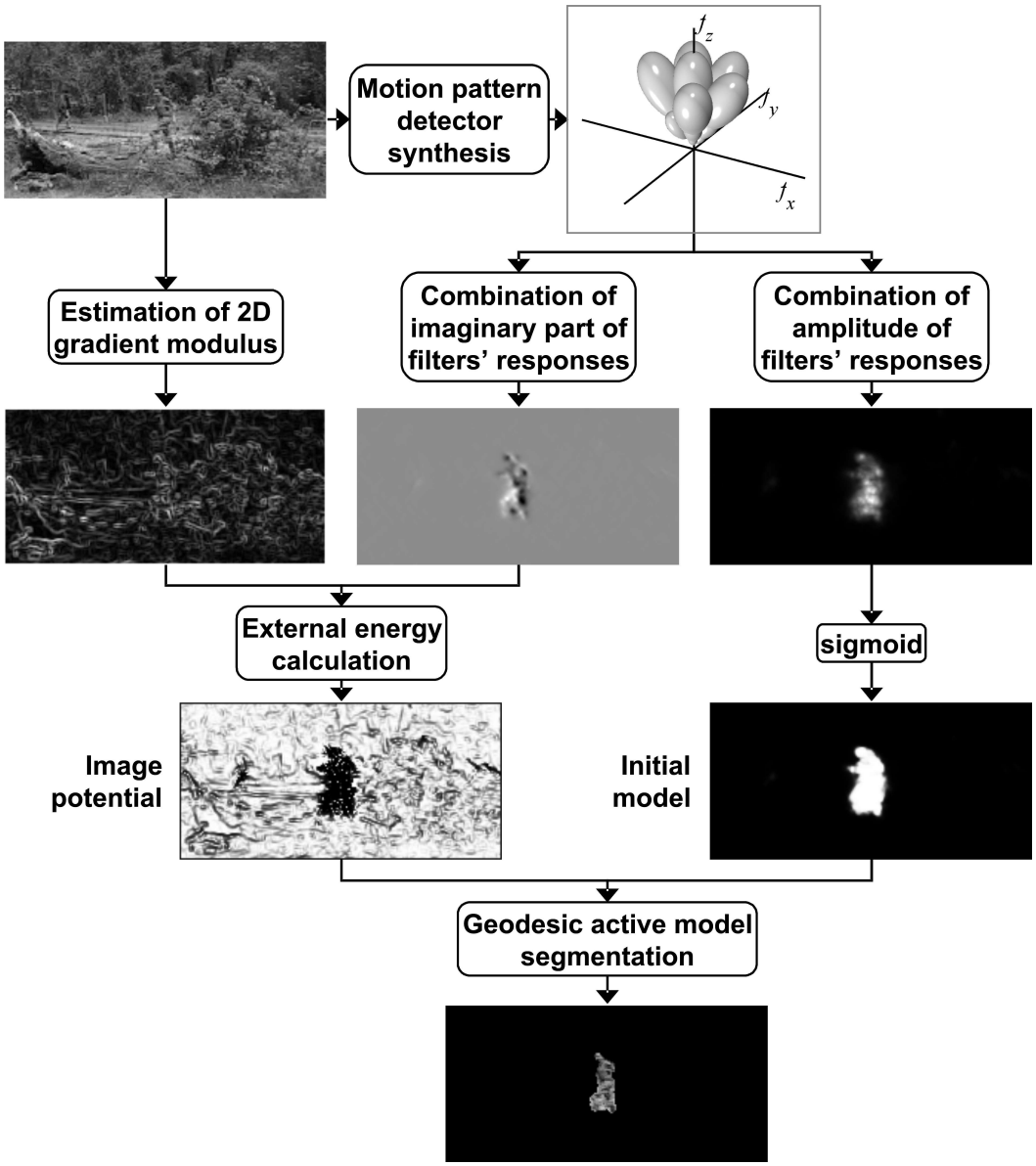
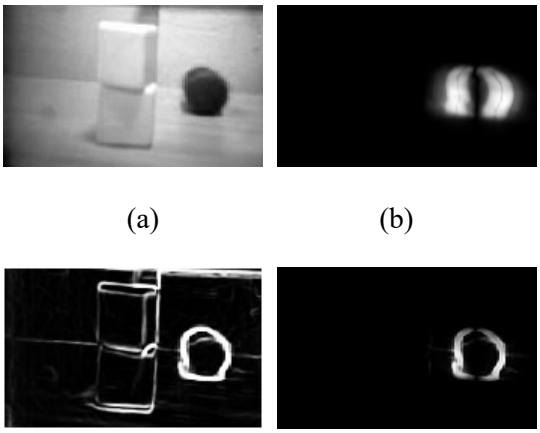



Fig. 5. Scheme of the segmentation technique. Filters in the composite feature detector are represented by its constituent filters, and these are represented as isosurfaces of their transfer functions, with level exp(−1/2).



(a) (b)

(c) (d)

Fig. 6. (a) One frame of an example sequence where a dark cylinder is moving from left to right. For one of the composite features detected: (b) $\Psi_{odd}$ representation. (c) Gradient after sigmoid thresholding. (d) Motion feature $C_m$ from equation (10) as the product of images (b) and (c).

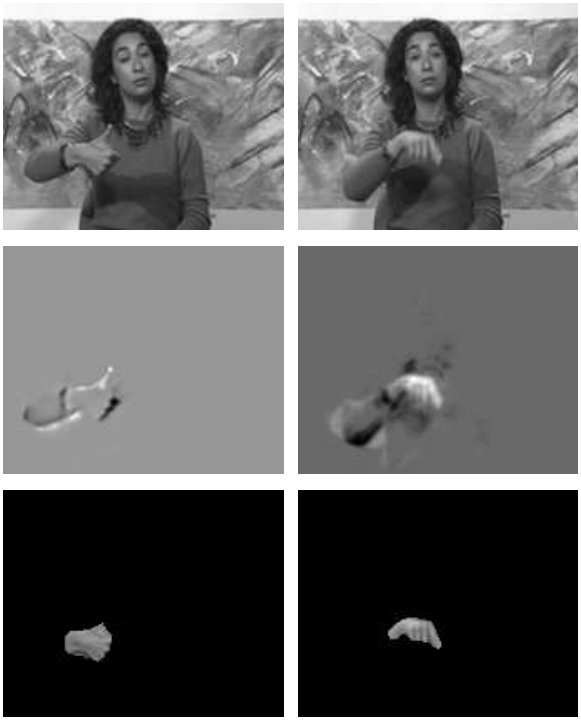


Fig. 7. Two frames of the “Silent” video sequence: *Top Row*: Input data. *Middle Row*: $\Psi_{even}$ of the motion pattern corresponding to the moving hand. *Bottom Row*: Segmentation using the active model based on the composite-feature.

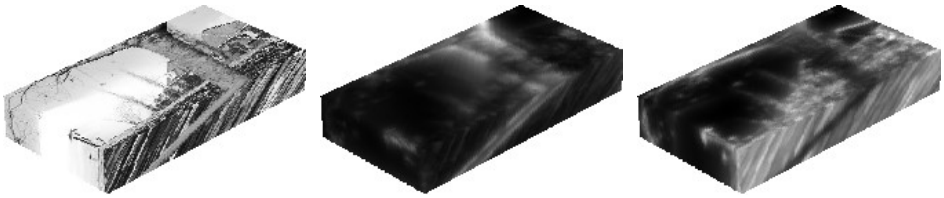


Fig. 8. A transversal cut of the "Flower Garden" video sequence: *Left*: Input data. *Centre* and *Right*: $\Psi_{amp}$ of the two motion patterns isolated by the composite-feature representation model.

Fig. 9. *Left*: A frame of the original sequence, and *Right*: the resulting segmentation.

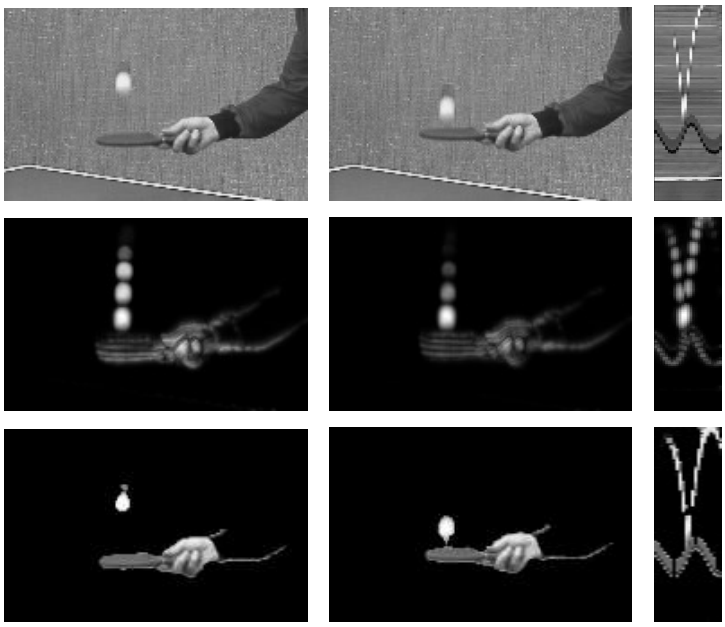


Fig. 10. *Top*: Left and centre images correspond to two consecutive frames of the "table tennis" video sequence. The image at the left is a traversal cut in the *y-t* plane. For frames on top row, *Middle Row*: $\Psi_{amp}$ of the selected composite-feature, and *Bottom*: Segmentation obtained with one of the detected composite-features.

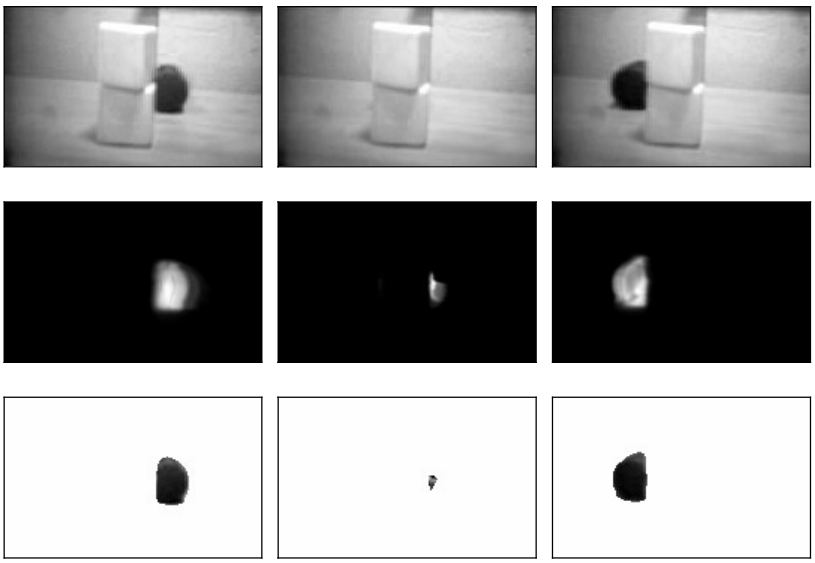


Fig. 11. *Top:* Three frames of a video sequence where a moving object is totally occluded during several frames. *Middle Row*: Initialization of the frames using the $\Psi_{amp}$ representation of one of the detected composite-feature. *Bottom*: Segmentation using initialization with the composite feature

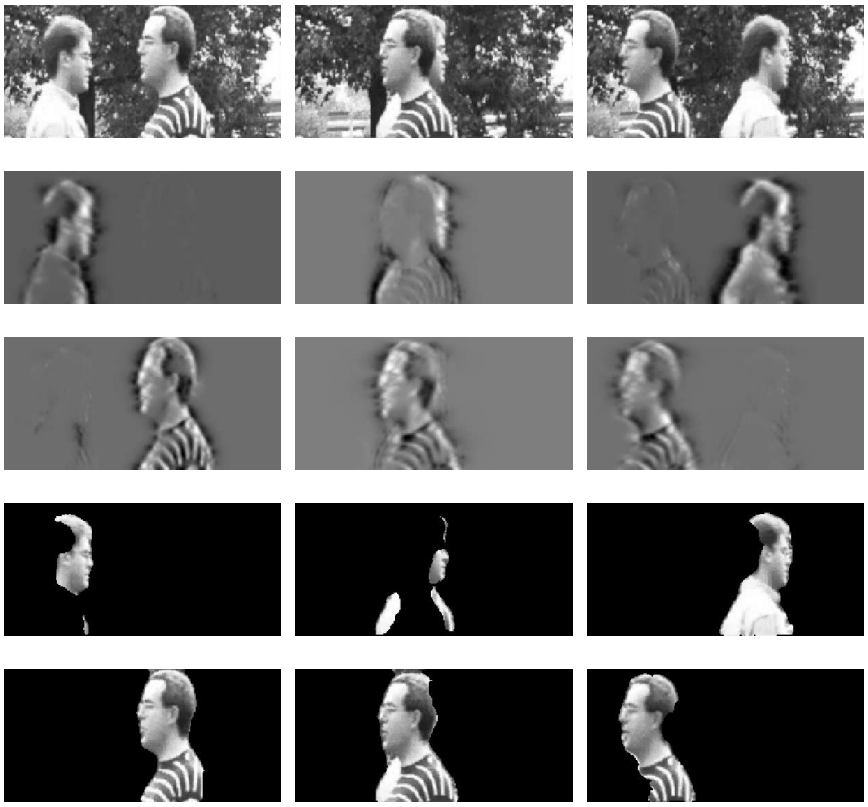


Fig. 12. Three frames of a sequence showing two occluding motion patterns. *1st row*: Input data. *2nd and 3rd rows*: $\Psi_{even}$ of two of the obtained composite-features, corresponding to the two motion patters. *4th and 5th rows*: Segmentations produced using composite-features from rows 2nd and 3rd respectively.

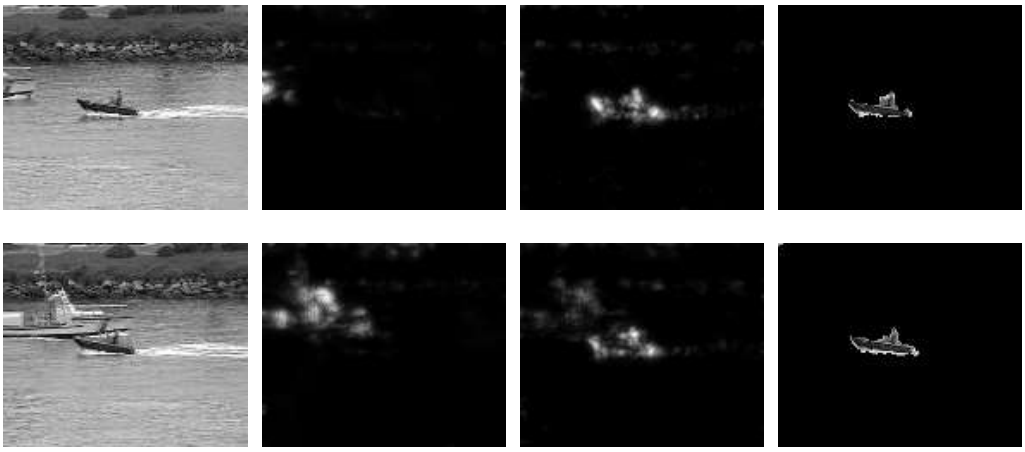


Fig. 13. Top and bottom row correspond to two frames of the "cost guard" video sequence: *1st column*: Input data. *2nd and 3rd columns*: $\Psi_{abs}$ of the two motion patterns isolated by the composite-feature representation mode and *4th column*: segmentations produced using the composite feature in 3rd column.

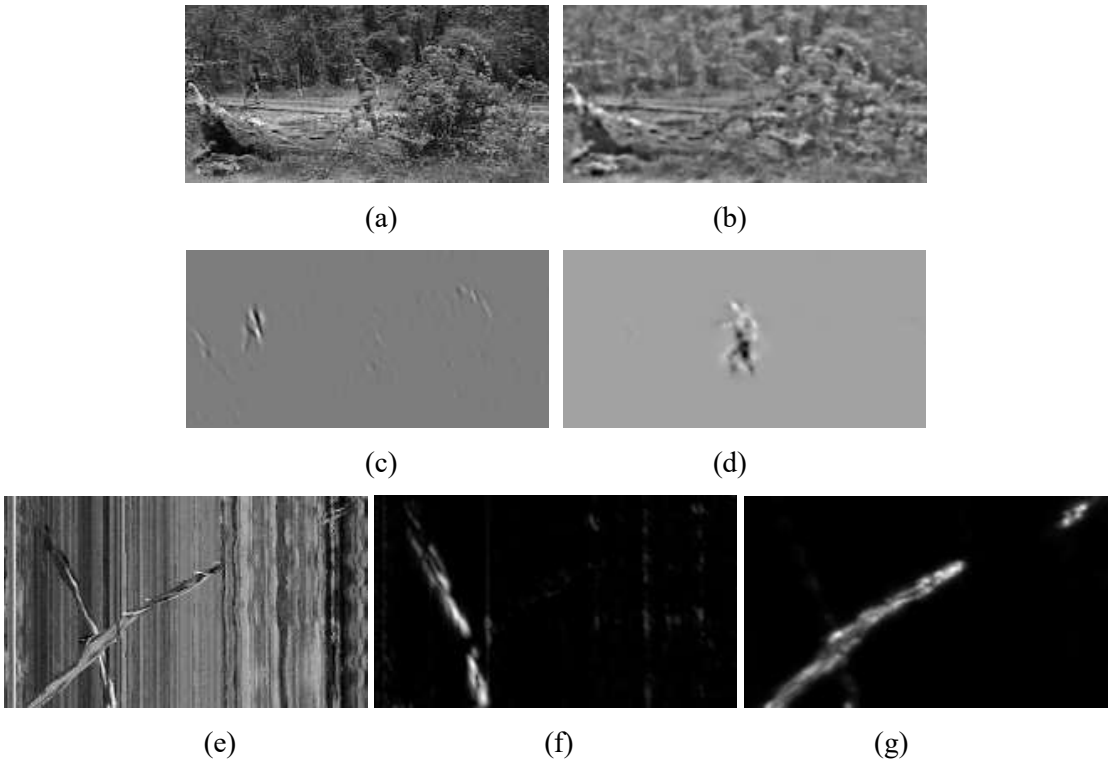



Fig. 14. (a) A frame of the input sequence. (b) $\Psi_{real}$ from the group of non active filters. (c) and (d). $\Psi_{real}$ from the composite features detected by the model. (e) Traversal cut of the input sequence –frame index increases from top to bottom. (f) and (g) Traversal cut of $\Psi_{abs}$ from the two composite features detected. Occlusions are better appreciated in this view.

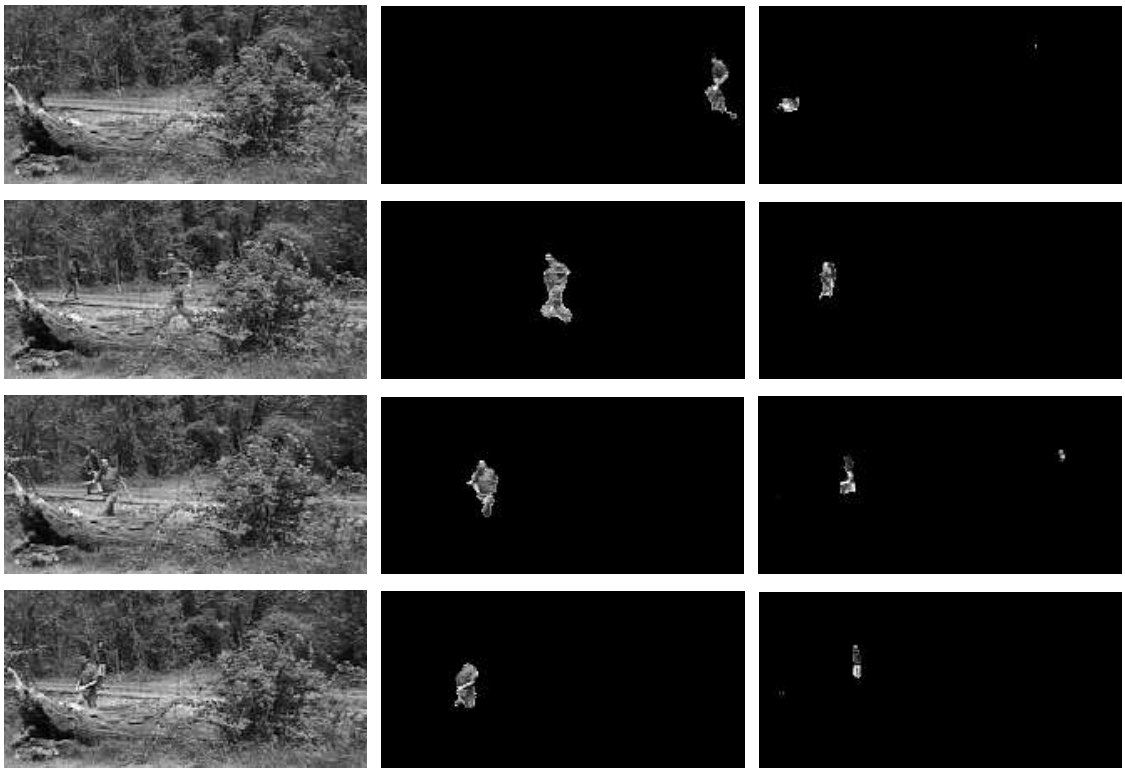


Fig. 15. *Left column*: Several frames of the input sequence. *Center and Left columns*:

Corresponding segmentations produced using each of the two detected composite features.