# Motion and appearance nonparametric joint entropy for video segmentation

Sylvain Boltz, Ariane Herbulot, Eric Debreuve, Michel Barlaud, Gilles Aubert

# Motion and appearance nonparametric joint entropy for video segmentation

S. Boltz* A. Herbulot* É. Debreuve* M. Barlaud* G. Aubert†

PREPRINT – Published in *International Journal on Computer Vision*
ISSN: 0920-5691 (print version) and 1573-1405 (electronic version)
DOI: 10.1007/s11263-007-0124-2

## Abstract

This paper deals with video segmentation based on motion and spatial information. Classically, the motion term is based on a motion compensation error (MCE) between two consecutive frames. Defining a motion-based energy as the integral of a function of the MCE over the object domain implicitly results in making an assumption on the MCE distribution: Gaussian for the square function and, more generally, parametric distributions for functions used in robust estimation. However, these assumptions are not necessarily appropriate. Instead, we propose to define the energy as a function of (an estimation of) the MCE distribution. This function was chosen to be a continuous version of the Ahmad-Lin entropy approximation, the purpose being to be more robust to outliers inherently present in the MCE. Since a motion-only constraint can fail with homogeneous objects, the motion-based energy is enriched with spatial information using a joint entropy formulation. The resulting energy is minimized iteratively using active contours. This approach provides a general framework which consists in defining a statistical energy as a function of a multivariate distribution, independently of the features associated with the object of interest. The link between the energy and the features observed or computed on the video sequence is then made through a nonparametric, kernel-based distribution estimation. It allows for example to keep the same energy definition while using different features or different assumptions on the features.

**Keywords:** Spatio-temporal segmentation, nonparametric distribution, joint entropy, active contour

*Laboratoire I3S, UMR CNRS 6070, Sophia Antipolis, France (boltz,herbulot,debreuve,barlaud@i3s.unice.fr).

†Laboratoire Dieudonné, UMR CNRS 6621, Nice, France (gaubert@math.unice.fr).

## 1 Introduction

Video segmentation aims at partitioning some video frames into objects and background. (For simplicity, it will be supposed that there is a single object.) This task can be performed without motion computation. If reference values of some descriptors are available (*e.g.*, mean color, color variance, color distribution... of the object of interest), an object can be segmented by minimizing a distance between the actual values of the descriptors computed on a candidate object domain and the reference values [21]. However, the lack of sensitivity of some descriptors near the object boundary (*e.g.*, the color distribution might not vary significantly if the candidate domain is slightly deformed) and the degree of freedom of the object motion (*a priori* infinite) may increase the number of potential solutions. Therefore, the segmentation framework involving motion computation will be considered.

Let us first consider the motion estimation task. Dense flow field estimation (*i.e.*, one motion vector per pixel) is an underdetermined problem. Moreover, when using the first order approximation of the brightness/color constancy constraint, only the motion component in the direction of the image gradient can be estimated. This limitation is known as the aperture problem. Motion estimation is therefore an ill-posed problem. It needs to be regularized, *i.e.*, constrained. On the one hand, the so-called global methods estimate a dense flow field while imposing the solution to be smooth [2, 8, 41]. On the other hand, local methods constrain the motion to follow a parametric model (*e.g.*, translation, affine motion, homography) with constant parameters, either in the whole image or within blocks or regions [30, 33, 43]. Both approaches have also been combined [10]. Given the link between motion estimation and object segmentation in a video, it can be noted that global methods require anisotropic smoothing to preserve object boundaries [41] whereas local methods are characterized by a chicken-and-egg dilemma: *(i)* estimating motion knowing the object

boundary while *(ii)* the boundary is defined as an optimal partition knowing the motion of the object and its neighborhood. This suggests to perform motion estimation and segmentation jointly [17], which will be the approach followed here.

Focusing on motion estimation again, imposing the brightness/color constancy constraint is equivalent, in variational terms, to minimizing a function of the motion compensation error (MCE), or of its first order approximation, as already mentioned. There is a correspondence between the choice of one such function and an assumption on the distribution of the MCE, *e.g.*, the square function and the assumption of a Gaussian distribution or the absolute value [43] and the assumption of a Laplacian distribution. This point of view will be referred to as parametric since the underlying distribution is characterized by a small set of parameters. In contrast, it is proposed to get rid of the parametric assumption on the data by trying to estimate the actual distribution as proposed for various related problems [13, 4, 27, 31] or in the context of shape prior [14, 28]. This approach will be referred to as nonparametric. Distribution approximation methods have been developed in statistics, most notably Parzen windowing [35, 40, 39] and the $k^{\text{th}}$-nearest neighbor (kNN) framework [22, 23].

In this nonparametric framework, we propose to use a unique statistical measure to both estimate the motion and segment the object. (A review of statistical methods in image segmentation was done recently [15].) Among the popular measures such as entropy [32], mutual information [42], or the Kullback-Leibler divergence [36], the entropy [7, 24] was chosen for its interesting properties (it is a measure of dispersion and it is robust to outliers - see Section 3.1) and because manipulating a single distribution (the distribution of the MCE) was preferred over taking the reference/target distribution comparison approach.

Motion-based segmentation can fail in areas insufficiently textured. In particular, the MCE is equal to zero in any homogeneous region. Therefore, adding such a region to, or subtracting it from, a given segmentation still produces potential solutions. This can be solved with the help of shape regularization [17], by adding spatial terms to the motion-based energy [9, 34], or by processing color and motion sequentially [19]. This last alternative is interesting but asks the difficult question of ordering the features, say, by importance (especially if involving even more features). The first two ones often requires a nontrivial adjustment of the weighting of the different terms. It will be shown that, using joint distributions, an objective choice of the weighting of the motion term and the spatial term can be made (namely, equal weighting or, equivalently, weight-free).

In brief, we propose to define a single spatio-temporal energy[1] to perform joint motion estimation and segmentation. To account for noise and model mismatch, the energy will be based on a statistical measure, namely entropy. In order to adapt to the data, no assumption will be made on the MCE or color distributions; they will be estimated using a nonparametric method. Finally, it will be shown that, with the proposed approach, the motion term need not be weighted relative to the spatial term. In a way, this offers a solution to the implementation of the operator AND between several properties (related here to motion "and" color) jointly describing the object of interest.

Also note that, although some equations below have some similarities with existing, likelihood-based or Bayesian methods, the philosophy here is different and somewhat more general. Bayesian methods are directly tied to the definition of the probability of the (observed) image or sequence given a segmentation. Assuming independence between the pixels, an energy is derived, which usually writes as a sum or integral of log probabilities. In the proposed approach, each region of the segmentation is regarded as a set of samples or realizations. The energy is defined as a function of a multivariate distribution in order to best fit the needs of the specific application. The link between the energy and the samples is then made through a nonparametric, Parzen-like or variable-size kernel-based distribution estimation. This allows for example to keep the same energy definition while using different object features or different assumptions on the features. In particular, one could think of discarding the assumption of independence between the pixels and use a patch-based (or neighborhood-based) approach [3] to change the spatial information from color to texture.

The paper is organized as follows: Section 2 details the problem statement. In Section 3, the classical parametric assumption on the MCE distribution is discarded and the proposed nonparametric framework for video segmentation, involving the actual residual and color distributions, is described. A single spatio-temporal energy is proposed to perform motion estimation and motion-based segmentation simultaneously. A piecewise motion model is introduced to allow enough flexibility for segmenting articulated objects. An active contour procedure is proposed in Section 4 to minimize

[1] Note that, here, the usage of the terms "temporal" and "spatio-temporal" should be understood as "motion-based" and "based on motion and color", respectively. "Spatio-temporal" more typically refers to a process performed in the $xyt$-space where $x$ and $y$ are video frame coordinates and $t$ is the time coordinate. As far as active contours are concerned, such a process would manipulate a tube oriented along the time dimension as opposed to a planar curve here.

the energy. Finally, Section 5 presents some results on synthetic and natural video sequences.

## 2 Problem Statement

The motion of an object domain $\Omega$ can be computed by choosing a motion model and finding the motion parameters that minimize a function of the MCE over $\Omega$. At a pixel level, making the assumption of brightness/color constancy, the MCE is classically equal to the following residual

$$e_n(v(x), x) = I_n(x) - I_{n+1}(x + v(x)) \tag{1}$$

where $x$ is a pixel of $\Omega$, $I_n$ is the $n^{\text{th}}$ grayscale or color frame of the sequence, and $v(x)$ is the apparent motion between $I_n$ and $I_{n+1}$ at $x$ (known as the optical flow). Ideally, $e_n(v(x), x)$ is equal to zero up to some noise. In grayscale, this condition provides a single equation for two unknowns (the components of $v(x)$) and, both in grayscale and color, it is likely that several pixels $y$ have the same value $I_{n+1}(y)$. As a consequence, the motion estimation problem cannot be solved without additional constraints. A possible way to constrain the problem is to assume that the motion is coherent with a chosen model inside $\Omega$ [43]. Then, the motion estimate $v$ can be computed as

$$v = \arg\min_{w} \int_{\Omega} \varphi(e_n(w, x))\ \mathrm{d}x \tag{2}$$

where $\varphi$ can be, for example, the square function, the absolute value, or a function typical of the robust estimation framework [5, 12].

The motion-based segmentation of frame $I_n$ can be formulated as the largest domain $\Omega$ inside which the motion is coherent with model (2), formally,

$$\begin{cases} \hat{\Omega} = \arg\min_{\Omega} \int_{\Omega} \varphi(e_n(v(\Gamma), x))\ \mathrm{d}x \\ v(\Gamma) = \arg\min_{w} \int_{\Omega} \varphi(e_n(w, x))\ \mathrm{d}x \end{cases} \tag{3}$$

where $\Gamma$ is the boundary $\partial\Omega$ of $\Omega$. Note that writing $v(\Gamma)$ or $v(\Omega)$ is only a matter of notation since $\Omega$ is completely determined by $\Gamma$ and conversely. Let us denote by $E_t$ the following domain energy[2]

$$E_t(\Gamma) = \int_{\Omega} \varphi(e_n(v(\Gamma), x))\ \mathrm{d}x\ . \tag{4}$$

Choosing $\varphi$ results in making an assumption on the distribution of the residual $e_n$ in $\Omega$. For example, if $\varphi$ is equal to the square function, the motion estimation is performed based on the assumption that the distribution is Gaussian; if $\varphi$ is equal to the absolute value, the distribution is assumed to be Laplacian[3]. However, these assumptions may not be appropriate. In particular, the presence of outliers in the residual (*e.g.*, due to occlusions, mismatch between the chosen motion model and the actual motion, variation of luminance...) may result in a complex, multimode distribution. As a consequence, the motion estimator in (3) may be biased, leading to a loss of accuracy of the motion-based segmentation.

## 3 Proposed segmentation energy

Three steps will be taken to derive the proposed energy: the definition of an ideal energy, its simplification, and its "symmetrization" (Sections 3.1, 3.2 and 3.3, respectively).

### 3.1 Nonparametric, entropy-based energy

To account for the true distribution of the residual $e_n$, and in general any feature that will be used for segmenting, it is proposed to make the energy depend on an estimation of the feature distributions rather than on the features themselves as it was the case in (4) concerning the residual. For the present segmentation task, the residual $e_n$ will be combined with the spatial feature $I_n$ (similar combinations of geometry and radiometry have been proposed [20, 29]). The proposed energy has the following form

$$\begin{cases} E(\Gamma) = -\dfrac{1}{|\Omega|} \displaystyle\int_{\Omega} \log f(e_n(v(\Gamma), x), I_n(x))\ \mathrm{d}x \\ v(\Gamma) = \arg\min_{w} E_{\Gamma}(w) \end{cases} \tag{5}$$

where $f$ is the joint distribution of the residual $e_n(v(\Gamma))$ and the image color $I_n$ inside the object domain $\Omega$, and

$$E_{\Gamma}(w) = -\frac{1}{|\Omega|} \int_{\Omega} \log f(e_n(w, x), I_n(x))\ \mathrm{d}x\ . \tag{6}$$

Energy (5) is the continuous version of the Ahmad-Lin approximation of differential entropy [1]. In both (5) and (6), $f$ is the joint distribution of the residual and the color. The residual being a function of the motion, $f$ is itself a function of $v(\Gamma)$ in the former and $w$ in the latter.

Let us see why this choice of energy is interesting. First, entropy is a measure of dispersion. If the segmentation is optimal, the residual should be distributed around zero with a minimal dispersion. Similarly, if the object is assumed to be piecewise homogeneous, the

[2] Subscript $t$ stands for *temporal*.

[3] Approach (3) is referred to as parametric since the underlying distributions are defined by a small number of parameters.

color distribution has a small dispersion. Moreover, entropy coincides locally asymptotically with likelihood at the optimum[4]. Thus, a minimum entropy criterion should have near optimal performances in case of a parametric distribution while being able to adapt to nonparametric cases. In particular, entropy appears to be less sensitive to outliers in practice.

Note that this approach defines a general framework for multimodal segmentation: the joint entropy allows to combine an arbitrary number of features/modalities. In practice, though, the number of modalities that can be combined together is limited by the number of samples available, *i.e.*, the number of pixels of the image or sequence frame. Indeed, if the samples fill the distribution space too sparsely, then the entropy (or any other statistical measure) cannot be approximated accurately. This problem, known as the curse of dimensionality, can be solved to a certain extent by the use of estimators based on the $k^{\text{th}}$-nearest neighbor (kNN) framework [6, 22, 23].

### 3.2 Simplification using marginal distributions

A fixed-size kernel-based procedure will be employed to estimate the distributions (see Section 3.5). To avoid that the entropy estimation be biased as an effect of the curse of dimensionality, energy (5) will be "simplified". Thus, the residual and the color will be assumed to be independent (See Appendix B). As a consequence, energy (5) can be rewritten as the following sum involving the marginal distributions

$$\begin{aligned} E(\Gamma) &= -\frac{1}{|\Omega|}\int_{\Omega} \log f_t(e_n(v(\Gamma),x))\ \mathrm{d}x \\ &\quad -\frac{1}{|\Omega|}\int_{\Omega} \log f_s(I_n(x))\ \mathrm{d}x \qquad (7) \\ &= E_t(\Gamma) + E_s(\Gamma) \qquad (8) \end{aligned}$$

where subscript $t$, respectively $s$, in $f_t$ and $E_t$, respectively $f_s$ and $E_s$, stands for temporal, respectively spatial. Note that the second integral in (7) was proposed for image segmentation [27].

The temporal energy in (7) is of the form

$$E_t(\Gamma) = \int_{\Omega} \psi(f_t(e_n(v(\Gamma),x)))\ \mathrm{d}x\ . \qquad (9)$$

One can say that the parametric approach (4) is extended to nonparametric distributions by substituting for a function of the residual $\varphi(e_n)$ a function of its distribution $\psi(f_t(e_n))$.

By making the assumption of independence, one obtains a sum of two energies, meeting the philosophy usually adopted when one want to simultaneously minimize several energies. However, in general, weighting parameters are introduced to tune the influence of the respective energies whereas, here, there are no such weights (it seems indeed natural not to favor any of the two terms since they have the same *unit*).

As suggested at the end of Section 3.1, this energy simplification might not be necessary if the distributions were estimated using the kNN framework [22, 23] as opposed to using the Parzen windowing approach (see Section 3.5). There is one concern about using the kNN framework in the present case though: as will be shown in Section 4.1, estimation of the distributions will be necessary for the proposed active contour process while kNN-based estimations are known to be noisy.

### 3.3 Region competition

In practice, due to approximations and roundoff errors, energy (7) might have the empty set as a unique global minimizer. A common solution is known as region competition: the energy of the background is added to the energy (7) of the object. It is not mandatory to use the same energy for the object and the background. However, it can be appropriate to do so. As a result, the segmentation will represent a tradeoff between the minimization of the object energy and the minimization of the background energy. It can also be interpreted as the maximal separation between object and background descriptors [44], here, the respective joint distributions.

To account for the relative areas of the object and the background, or, in other words, to account for the probability of a pixel to belong to either of them, the following weighted sum will be used

$$E_{\text{rc}}(\Gamma) = \frac{|\Omega|}{|D|}\ E(\Gamma) + \frac{|\Omega^c|}{|D|}\ E(\Gamma^c) \qquad (10)$$

where $\Omega^c$ is the complement of $\Omega$ in $D$, the image domain, and $\Gamma^c$ is its boundary $\partial\Omega^c$ (the division by $|D|$ of $|\Omega|$ and $|\Omega^c|$ can be omitted since it has no influence on the minimization).

Energy (10) can be rewritten as

$$E_{\text{rc}}(\Gamma) = p(C=1)\ E(\Gamma) + p(C=0)\ E(\Gamma^c) \qquad (11)$$

where $C$ is the characteristic function of the object and $p(C=i)$ denotes the probability of the event $C=i$. As defined in (5), energy $E(\Gamma)$ is (an approximation of) the joint entropy of the residual and the color conditional on $C=1$. Let us denote it by $H(e_n, I_n|C=1)$. Equivalently, $E(\Gamma^c)$ is equal to $H(e_n, I_n|C=0)$. Then, Eq. (11) is equal to

$$E_{\text{rc}}(\Gamma) = \sum_{i\in\{0,1\}} p(C=i)\ H(e_n, I_n|C=i) \qquad (12)$$

[4] This is interesting since the maximum likelihood estimator is optimal when the distribution of data is parametric.

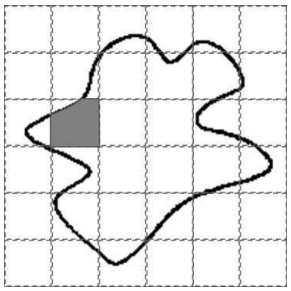


Figure 1: Solid line: contour $\Gamma = \partial\Omega$; Dashed blocks: $B_i$; Gray-filled block: an example of a domain $\Omega_i$ with boundary $\Gamma_i$.

$$= H(e_n, I_n|C) \ . \tag{13}$$

Therefore, energy (10) is equal[5] to the conditional joint entropy of the residual and the color $H(e_n, I_n|C)$.

### 3.4 Motion estimation

As mentioned in Section 2, the motion $v$ is assumed to follow a given model inside $\Omega$. For example, it can be defined by a set of parameters $p$ [33]. Then, estimating $v$ in (9) is only a matter of estimating $p$. This task is certainly made easier if the relation between $v$ and $p$ is linear

$$v(\Gamma) = M \ p(\Gamma) \tag{14}$$

where $M$ is a $2 \times l$ matrix if $p$ is an $l$-vector. Even if the motion model is complex, it will hardly account for general motions such as motions of articulated objects, and if it does, solving for the model parameters is likely to be an ill-posed inverse problem. Instead, we propose to keep the model simple while solving for its parameters locally. Frame $I_n$ is divided into $k$ blocks $B_i$ of identical size, where $k$ depends on the frame size. Let $\Omega_i$ be the intersection of $\Omega$ with $B_i$ and let $\Gamma_i$ be the boundary $\partial\Omega_i$ of $\Omega_i$ (see Fig. 1). The temporal energy (9) is replaced with

$$E_t^{\text{local}}(\Gamma) = -\frac{1}{|\Omega|} \int_\Omega \log f_t(e_n(v_1, \ldots, v_k, x)) \ \mathrm{d}x \tag{15}$$

where $v_i$ is a short notation for $v(\Gamma_i)$, the motion of $\Omega_i$. (The consequence of using this local approach is discussed in Appendix D.) In this context, the motion model can simply be translation. Therefore, Eq. (1) is replaced with

$$e_n(v_i, x) = I_n(x) - I_{n+1}(x + v_i), \ x \in \Omega_i \ . \tag{16}$$

This local approach will be used when the object of interest is articulated (see Section 5.4). In the other experiments, a global translation will be used. It corresponds to decomposing $I_n$ into a single block $B_1$ covering the whole frame. Note that in the following, for clarity, the notations $E_t$ and $e_n(v(\Gamma), x)$ will be preferred over $E_t^{\text{local}}$ and $e_n(v_1, \ldots, v_k, x)$, respectively.

Finally, to minimize the influence of occlusions, expression (16) is regarded as the forward residual and compared with the backward version as follows

$$e_n(v, x) = \min_{\text{abs}}\{I_n(x) - I_{n+1}(x+v), \ I_n(x) - I_{n-1}(x-v)\} \tag{17}$$

where $\min_{\text{abs}}$ is equal to

$$\min_{\text{abs}}\{a, b\} = \begin{cases} a & \text{if} \quad \min\{|a|, |b|\} = |a| \\ b & \text{if} \quad \min\{|a|, |b|\} = |b| \end{cases} \ . \tag{18}$$

Function (18) is not differentiable. However, in the present work, it does not need to be differentiated (see Appendix C).

### 3.5 Distribution estimation

Parzen windowing is a classical distribution estimation procedure [35]. The following continuous version was used

$$f(r) = \frac{1}{|\Omega|} \int_\Omega K_\sigma(r - g(x)) \ \mathrm{d}x \tag{19}$$

where $|\Omega|$ is the measure of $\Omega$, $K_\sigma$ is a Gaussian kernel with zero mean and a variance equal to $\sigma^2$, and $g$ is a random variable whose distribution is to be estimated (*i.e.*, $e_n(v(\Gamma))$ or $I_n$). It is usual to adapt $\sigma^2$ to the data [40, 39].

## 4 Segmentation using active contours

### 4.1 Shape gradient of the energy

Minimization of energy (7) requires the computation of its derivative with respect to $\Gamma$. There exists an infinite number of ways of deforming $\Gamma$. The shape derivative [18, 25, 26, 4] of (7) can be interpreted as the derivative in a direction $F$, a vector field defined on $\Gamma$. It can be shown that the shape derivative of (9) is equal to (see Appendix C)

$$\begin{aligned} \mathrm{d}E_t(\Gamma, F) = \frac{1}{|\Omega|} \int_\Gamma \Bigg[ & \log f_t(e_n(v(\Gamma), s)) - 1 + E_t(\Gamma) \\ & + \frac{1}{|\Omega|} \int_\Omega \frac{K_\sigma(e_n(v(\Gamma), s) - e_n(v(\Gamma), x))}{f_t(e_n(v(\Gamma), x))} \ \mathrm{d}x \Bigg] \\ & N(s) \cdot F(s) \ \mathrm{d}s \end{aligned} \tag{20}$$

where $N$ is the inward unit normal of $\Gamma$.

Note that the distribution $f_t$ appears explicitly in (20), hence the necessity to estimate it even though

[5] In fact, would be equal if the assumption of independence between $e_n$ and $I_n$ had not been made (see Section 3.2).

the kNN framework allows to compute (9) without computation of the underlying distribution [22, 23].

The expression of $\mathrm{d}E_s$ is similar to (20) (see Appendix C). Finally, the shape derivative of (7) is equal to

$$\mathrm{d}E(\Gamma, F) = \mathrm{d}E_t(\Gamma, F) + \mathrm{d}E_s(\Gamma, F) \ . \tag{21}$$

The shape derivative (21) has the following form

$$\begin{aligned} \mathrm{d}E(\Gamma, F) &= \int_\Gamma ((\alpha_t(s) + \alpha_s(s))\ N(s)) \cdot F(s)\ \mathrm{d}s & (22) \\ &= \langle \alpha\ N, F \rangle & (23) \end{aligned}$$

where $\langle , \rangle$ is the $L^2$-inner product on $\Gamma$. Therefore, $\alpha\ N$ is, by definition, the gradient of (7) at $\Gamma$ associated with this inner product.

### 4.2 Region competition

The shape derivative of (10) can be obtained by applying the traditional differentiation rule $(u\ v)' = u'\ v + u\ v'$ and determining the shape derivative of $|\Omega|$ (see Appendix C.3). The terms related to the object and the terms related to the background can be gathered together by noting that $\Gamma$ and $\Gamma^c$ are identical up to a change of orientation. In particular, the inward unit normal $N^c$ of $\Gamma^c$ is equal to $-N$.

### 4.3 Evolution equation

Based on the notion of gradient defined in Section 4.1, energy (10) can be minimized using a steepest descent procedure in the space of contours. The following contour evolution process is known as the active contour technique [11, 25]: an initial contour[6] is iteratively deformed in the opposite direction of the gradient until a convergence condition is met. The evolution equation of the active contour is written as follows

$$\left\{ \begin{array}{l} \Gamma(\tau = 0) = \Gamma_0 \\ \dfrac{\partial \Gamma}{\partial \tau} = (\alpha^c - \alpha)\ N \end{array} \right. \tag{24}$$

where $\tau$ is the evolution parameter and $\alpha^c$ has the same expression as $\alpha$ but is evaluated on $\Omega^c$. The convergence condition is $\alpha^c - \alpha = 0$.

## 5 Experimental results

### 5.1 Test settings

As a reminder, the proposed segmentation energy has the following form

$$E_{\mathrm{rc}}(\Gamma) = |\Omega|\ E(\Gamma) + |\Omega^c|\ E(\Gamma^c) \tag{25}$$

where

$$E(\Gamma) = E_t(\Gamma) + E_s(\Gamma) \ . \tag{26}$$

For comparison purposes, energy (25) will also be used in two incomplete forms: when $E_s$ is removed from the definition of $E$ in (26), the energy will be called temporal energy; when $E_t$ is removed from the definition of $E$, the energy will be called spatial energy. In its complete form, it was already defined as the spatio-temporal energy.

The tests were performed on synthetic and natural sequences composed of $300 \times 300$-pixel frames and `cif`[7] frames, respectively, all defined in the $YUV$-color space. The $V$ channel was discarded. Therefore, the distributions of $I_n$ and $e_n$ are functions from $\mathbb{R}^2$ to $\mathbb{R}$ with support $[0, 255]^2$ and $[-255, 255]^2$, respectively. In computing $e_n$, $I_{n+1}(x + v)$ was bilinearly interpolated. Independence between spatial and temporal information was assumed in Section 3.2 in order to write $E$ as the sum of $E_t$ and $E_s$. The computation of these two components was also simplified by assuming independence between the channels $Y$ and $U$. As a consequence, $E_t$ and $E_s$ were themselves estimated as the sum of a $Y$-based entropy and a $U$-based entropy.

The standard deviation $\sigma$ of the Parzen kernel (see Eq. (19)) was adapted to the data[8] by using the empirical standard deviation $\hat{\sigma}$ of the residual or the color in $\Omega$

$$\sigma = 0.9 \min(\hat{\sigma}, \hat{p}/1.34)\ |\Omega|^{-1/5} \tag{27}$$

where $\hat{p}$ is the interquartile range of the data in $\Omega$. Therefore, $\sigma$ should be regarded as a function of $\Omega$. This would add some terms to the shape derivative $\mathrm{d}E$ since expression (52) would not be valid anymore. However, these terms can be neglected because $\sigma$ does not change significantly between two iterations of the active contour process.

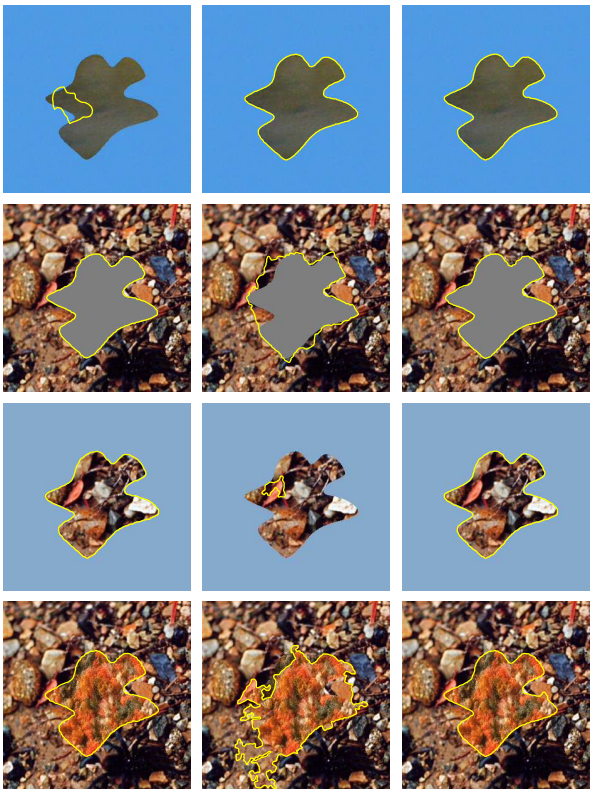
As mentioned in Section 3.4, translation was chosen as the motion model. The motion estimation in (9) was performed by fast, suboptimal (as opposed to exhaustive) search [45] within a search window of -12/+12 pixels in both directions and a quarter of a pixel precision. This procedure was used whether the motion was estimated globally in $\Omega$ or locally in each $\Omega_i$.

In the following, "segmentation" refers to object detection with an initialization far from the solution (typically a circle) while "tracking" refers to object detection with an initialization obtained by translating by $v_{\mathrm{global}}$ the object contour as detected in the previous frame, where $v_{\mathrm{global}}$ is the motion of $\Omega$ computed by the suboptimal procedure described above.

[6] For example, a user-defined contour.

[7] The frame size `cif` corresponds to $352 \times 288$ pixels.

[8] Adapting the kernel bandwidth to the data is known as a plug-in procedure [40].



Temporal only Spatial only Spatio-temporal

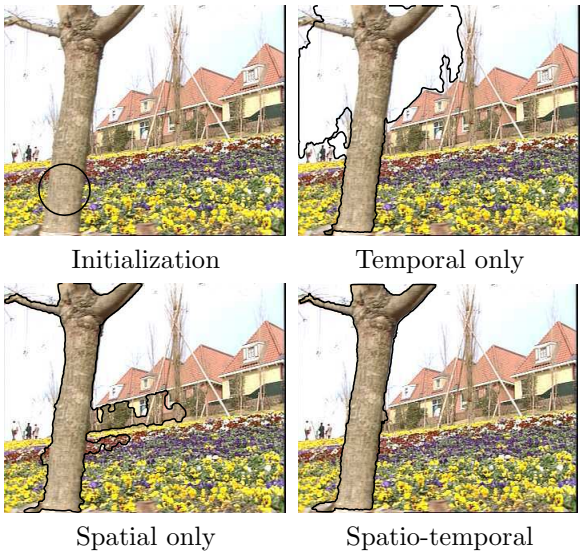
Figure 2: Segmentation of synthetic sequences accounting for motion, color, or both. First row: homogeneous object over homogeneous background; second row: homogeneous object over textured background; third row: textured object over homogeneous background; last row: textured object over textured background.

## 5.2 Comparing spatial, temporal, and spatio-temporal energies

In this section, motion is estimated globally on $\Omega$ (see Section 3.4).

### 5.2.1 Synthetic sequences

Several synthetic sequences were designed by combining different textures and homogeneous areas with a given *motion scenario*: an object is translating horizontally by -3 pixels over a background translating horizontally by 1 pixel. Segmentation was performed with the spatial energy, the temporal energy, and the spatio-temporal energy (see Fig. 2). These results suggest that the temporal energy is adapted whenever there is texture. On the contrary, the spatial energy seems more reliable in homogeneous areas. Finally, the combination of temporal and spatial information appears appropriate for segmenting sequences that contain homogeneous areas, textured areas, or both.



Initialization Temporal only

Spatial only Spatio-temporal

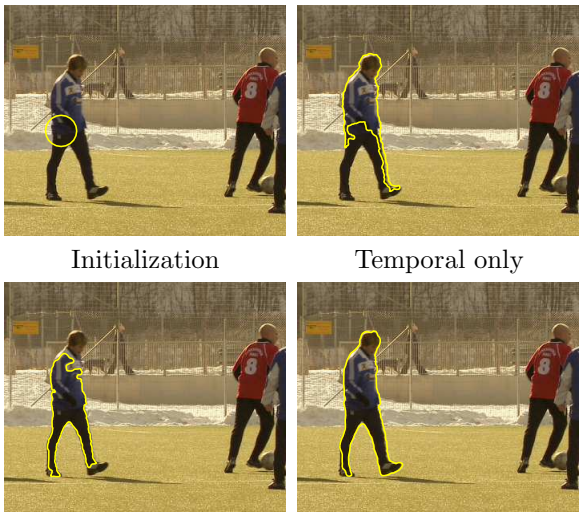
Figure 3: Segmentation of frame 237 of sequence `'Flowers and garden'` accounting for motion, color, or both.

### 5.2.2 Standard test sequences

The same comparison as in Section 5.2.1 was performed with standard test sequences `'Flowers and garden'` and `'Soccer'`.

In sequence `'Flowers and garden'`, the sky bordering the tree is rather homogeneous (see Fig. 3). Therefore, oversegmentation occurs with the temporal energy, as noted in Appendix A. With the spatial energy, the segmentation process also fails because part of the houses in the background have colors similar to the tree. Finally, the spatio-temporal segmentation mostly excludes the sky since it has a different color (spatial information) and also excludes the houses since they have a different motion (temporal information).

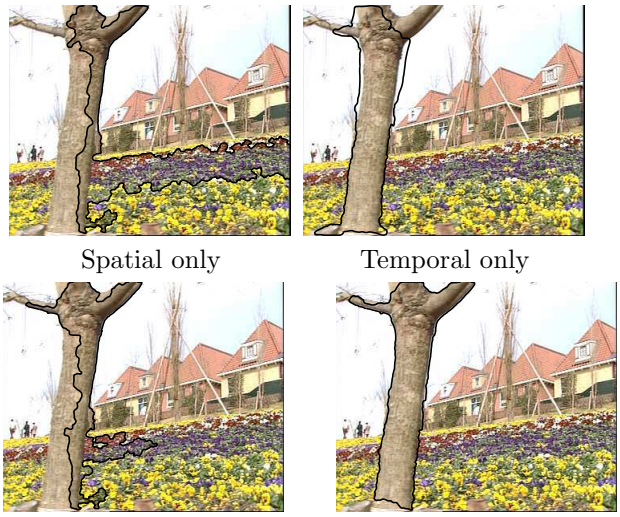
In sequence `'Soccer'`, the soccer player has a complex, articulated motion (see Fig. 4). The temporal energy only captures the rigid part of the body while the spatial energy does not capture the head as it has colors similar to the background. The spatio-temporal energy provides a good tradeoff, although it sometimes *miss* a foot of the player (see Fig. 9) for which both the temporal information and the spatial information (the color of the shoe is similar to background colors in the $YU$-color space) are unreliable. This satisfying result can be explained by the fact that the spatial energy helps the temporal term when the motion model mismatches the actual motion, and the temporal energy helps the spatial term when the color is not discriminating.



Initialization Temporal only

Spatial only Spatio-temporal

Figure 4: Segmentation of frame 162 of sequence `'Soccer'` accounting for motion, color, or both.



Spatial only Temporal only

Spatio-temporal with $\alpha = 1$ Spatio-temporal with $\alpha \simeq 9$

Figure 5: Segmentation of frame 237 of sequence `'Flowers and garden'` assuming parametric distributions and using the same initialization as in Fig. 3. *(Lower left)* The spatio-temporal energy relies equally on space and time and *(Lower right)* the spatio-temporal energy favors the temporal term.

### 5.3 Parametric vs. nonparametric

In this section, motion is estimated globally on $\Omega$ (see Section 3.4).

One can wonder the practical benefits of relying on nonparametric estimations of the residual distribution and the color distribution as opposed to using classical error terms corresponding to parametric assumptions. For a fair comparison, the parametric assumptions for the residual and the color distributions have to be chosen appropriately. The residual is corrupted by outliers mainly due to noise, illumination variations, motion model mismatch, and occlusion. The Sum of Absolute Differences (SAD) [43] was chosen since it is robust to outliers. Note that it follows from a Laplacian assumption (see Appendix E.1).

It is clear that there is no ideal parametric assumption concerning the spatial term. Nevertheless, noting that the spatial entropy in (7) can be interpreted as a piecewise color homogeneity criterion, it seems reasonable to make the assumption of a Gaussian distribution[9] (see Appendix E.2).

The continuous form of criteria (80) and (82) can be linearly combined to define a parametric, space-time segmentation energy

$$E_p(\Gamma) = \int_\Omega (I_n(x) - \mu_I(\Gamma))^2 \, \mathrm{d}x + \alpha \int_\Omega |I_n(x) - I_{n+1}(x + v(\Gamma))| \, \mathrm{d}x \quad (28)$$

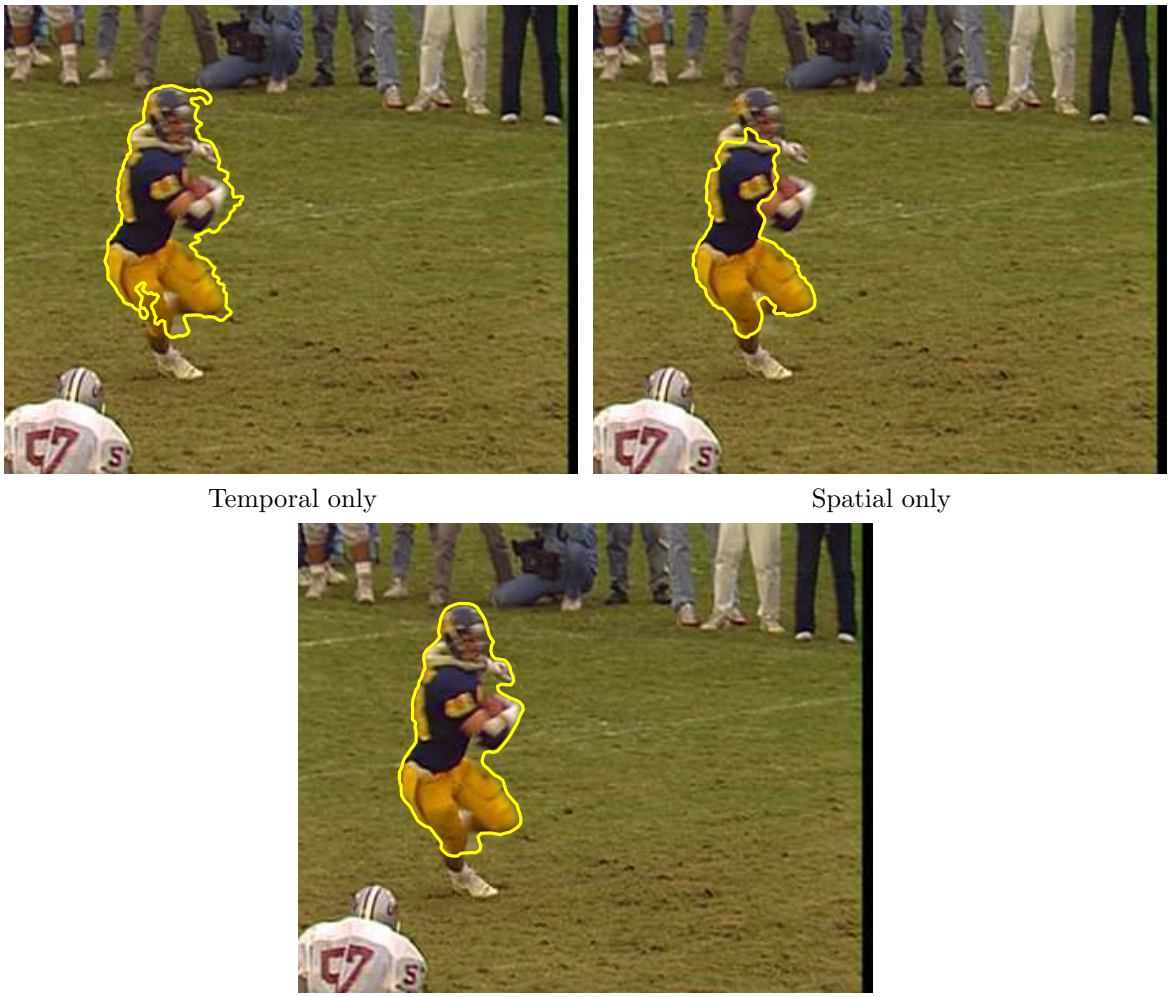
[9]To be coherent with the *piecewiseness* property of entropy, a mixture of Gaussians would be more appropriate. However, the purpose of this section is to compare the proposed approach with classical error terms such as the Sum of Squared Differences (SSD).

where

$$\begin{cases} \mu_I(\Gamma) = \int_\Omega I_n(x) \, \mathrm{d}x / \int_\Omega \, \mathrm{d}x \\ v(\Gamma) = \arg\min_w \int_\Omega |I_n(x) - I_{n+1}(x + w)| \, \mathrm{d}x \end{cases} \quad (29)$$

and $\alpha$ is a positive constant. The nonparametric energy (7) does not weight the spatial term relatively to the temporal term. Therefore, to be coherent, $\alpha$ should be equal to one. However, the results on sequence `'Flowers and garden'` suggest to choose $\alpha$ greater than one (see Fig. 5). In each experiment, the optimal value was determined empirically. Moreover, to give an idea of the behavior of each term of the parametric energy (28), segmentation was also performed using each term separately (same procedure as in Section 5.2). The parametric approach was also tested on the other, more challenging sequence `'Football'`. Even when assigning a higher weight to the temporal term, the segmentation is not satisfying (see Figs. 6, 7, and 8).

In light of these results, three intuitive conclusions can be made. *(i)* As expected, when the parametric assumptions are roughly in accordance with the actual distributions (sequence `'Flowers and garden'`), the parametric approach can perform well. *(ii)* The Laplacian assumption for the residual distribution is more reliable than the Gaussian assumption for the color distribution. Indeed, with sequence `'Flowers and garden'`, the correct segmentation is obtained only when the temporal term is weighted significantly more



Temporal only Spatial only

Spatio-temporal

Figure 8: Segmentation of frame 72 of sequence `'Footbal l'` with the nonparametric approach using the same initialization as in Fig. 6.

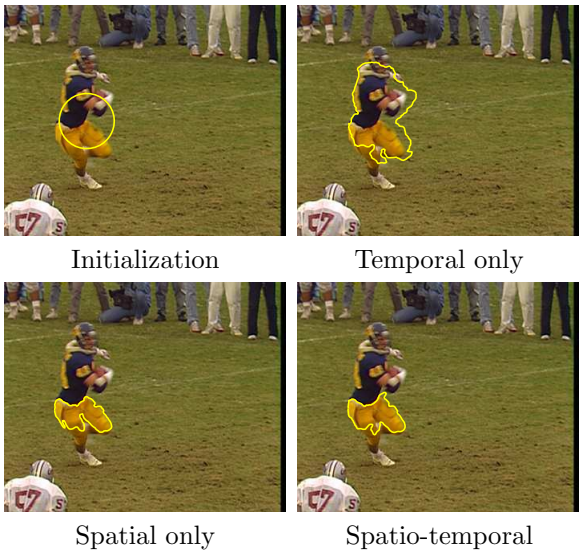



Figure 6: Segmentation of frame 72 of sequence `'Football'` assuming parametric distributions. The spatio-temporal energy relies equally on space and time ($\alpha = 1$).

Figure 7: Segmentation of frame 72 of sequence `'Football'` assuming parametric distributions. The spatio-temporal energy favors the temporal term ($\alpha \simeq 9$).

than the spatial term. *(iii)* Again, as expected, when the parametric assumptions clearly mismatch the actual distributions for the motion being complex or the object and background being composed of several colors (sequence `'Football'`), the parametric approach fails, as opposed to the proposed nonparametric approach (see Figs. 3 and 4).

### 5.4 Tracking and piecewise motion estimation

In this section, an object of interest is tracked in two standard test sequences using the proposed method. In both sequences, the object of interest is composed of several colors and has a complex, articulated motion. Therefore, they are appropriate for comparing the global (on $\Omega$) motion approach and the local (on $\Omega_i$) motion approach (see Section 3.4). Sequence `'Soccer'` (already seen in Section 5.2.2) is less complex than sequence `'Football'` (already seen in Section 5.3) since the latter suffers from motion blur. For the piecewise motion estimation, each frame was divided into $16 \times 16$ blocks $B_i$ of size $22 \times 18$ pixels. As a reminder, domain $\Omega_i$ is defined as $\Omega \cap B_i$. The comparison between the two approaches is presented in Figs. 9 and 10. Although the local approach clearly improves the segmentation, it is not perfect in sequence `'Football'`. This can be explained by the combined effects of motion blur and a domain $\Omega_i$ too small (which may happen for blocks $B_i$ that intersect $\Gamma$), resulting in a less reliable local motion estimate.

## 6 Conclusion

The addressed problem was the segmentation of a video sequence. A spatio-temporal approach was chosen in order to make use of both spatial and temporal coherence. As opposed to the classical approach consisting in dealing with time by involving the MCE directly, the proposed method is based on the use of the distribution of the MCE. This allowed to combine temporal and spatial information coherently using joint distributions. The distributions were estimated nonparametrically to fit the data. Entropy was chosen as the energy to minimize, in particular because, in practice, it is robust to outliers. In order to make the motion model complex enough to describe articulated objects, it was proposed to keep it simple (namely, translation) while estimating its parameters locally.

The proposed method was qualitatively compared with a classical, parametric approach followed by some existing methods. Thorough comparison with specific methods is out of scope of this paper, though. Nevertheless, on sequence `'Flowers and garden'`, our results (see Fig. 3) are comparable to those of recent

Figure 9: Tracking on sequence `'Soccer'`: comparison between the global motion approach (first column) and the local motion approach (second column) on frames 162, 172, 182, 192, and 202. The segmentation is more accurate with the local motion approach (the main differences are highlighted by circles).

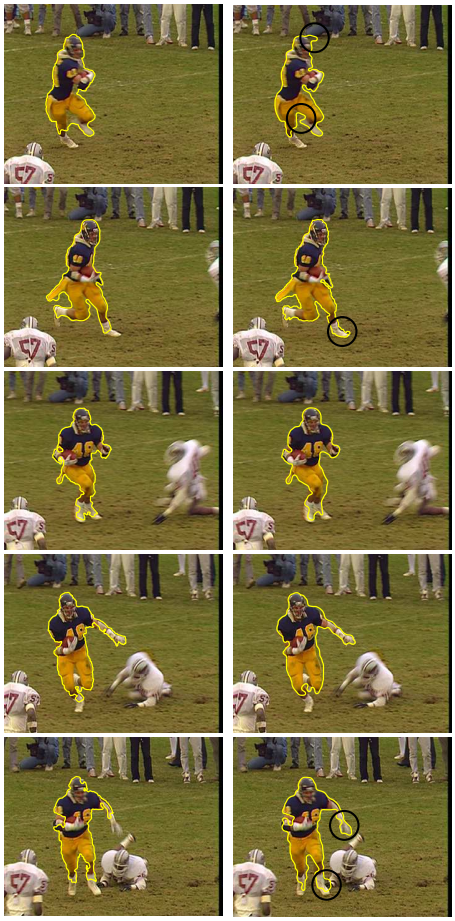


Figure 10: Tracking on sequence `'Football'`: comparison between global motion approach (first column) and local motion approach (second column) on frames 73, 77, 81, 85, and 89. The segmentation is more accurate with the local motion approach (the main differences are highlighted by circles).

segmentation methods [16, 38], Fig. 4 in both articles.

# A Disambiguation using spatial information

Energy (9) is well suited for segmenting objects over a textured background. However, it might cause segmentation to include homogeneous or quasi-homogeneous areas of the background. Indeed, this type of areas has a low residual even if compensated with the motion estimated for the object, at least as long as the motion-compensated object domain remains in the homogeneous area. Therefore, the energy might increase only negligibly when expanding in such areas. Since the notions of object and background are arbitrary and can be swapped for one another, one can note that an equivalent undersegmentation phenomenon can occur if the object contains homogeneous areas near its boundary.

On the other hand, the entropy of the object color increases if the object domain includes some background since it adds new colors to the object[10] and, therefore, increases the dispersion of its color distribution. Consequently, the joint entropy of the residual and the color also increases.

# B "Independence" between residual and color

Let us consider the following sequence model

$$I_{n+1}(x) = I_n(T(x)) + n(x) \tag{30}$$

where $T$ is a transformation and $n$ is a Gaussian white noise. The residual is equal to

$$e_n(v(x), x) = I_n(x) - I_{n+1}(v(x)) \ . \tag{31}$$

If the transformation $T$ exists and the motion is perfectly estimated, then $v$ is equal to $T^{-1}$ and $e_n(v(x), x) = -n(T^{-1}(x))$, which is independent of $I_n$. However, model (30) is an approximation: in general, there is no such transformation $T$, frame $I_{n+1}$ being a projection on a two-dimensional plane of a three-dimensional scene. Often, some parts of objects in $I_n$ become invisible in $I_{n+1}$ while others become visible. Therefore, frame $I_{n+1}$ cannot be deduced entirely from $I_n$. In the unpredictable areas, the residual is by definition independent of $I_n$. Overall, whether a transformation $T$ exists or not, if the motion $v(\Gamma)$ is fairly well estimated, then the assumption of independence should be acceptable.

[10] If the background has the same color as the object near the boundary, there is no objective information to find the object boundary.

# C Energy derivative

## C.1 Temporal energy

The temporal energy is equal to

$$E_t(\Gamma) = -\frac{1}{|\Omega|} \int_\Omega \log f_t(e_n(v(\Gamma), x)) \ \mathrm{d}x \tag{32}$$

where

$$\begin{cases} f_t(r) &= \dfrac{1}{|\Omega|} \displaystyle\int_\Omega K_\sigma(e_n(v(\Gamma), x) - r) \ \mathrm{d}x \\ e_n(v(\Gamma), x) &= \min_{\text{abs}}(I_n(x) - I_{n+1}(x + v(\Gamma)), \\ & \qquad I_n(x) - I_{n-1}(x - v(\Gamma))) \\ v(\Gamma) &= \arg\min_w -\dfrac{1}{|\Omega|} \displaystyle\int_\Omega \log f_t(e_n(w, x)) \ \mathrm{d}x \end{cases} \tag{33}$$

Note that, for simplicity, residual $e_n$ has been defined for a translation motion model. However, the following development is valid for any motion model.

The definition of the shape derivative of (32) is based on a domain transformation $T$ whose amplitude continuously depends on a parameter $\tau$ such that $T(\Omega, \tau = 0)$ is equal to $\Omega$ and $T(\Omega, \tau)$ is equal to $\Omega(\tau)$ [18, 25, 26, 4]. Functions of $\Omega$, or $\Gamma$, can then be rewritten as functions of $\tau$. In this context, the shape derivative of

$$E(\Gamma) = \int_\Omega G(\Gamma, x) \ \mathrm{d}x \tag{34}$$

is equal to

$$\mathrm{d}E(\Gamma, F) = \frac{\mathrm{d}E}{\mathrm{d}\tau}(\tau = 0) \tag{35}$$

$$= \int_\Omega \frac{\partial G}{\partial \tau}(\tau = 0, x) \ \mathrm{d}x - \int_\Gamma G(\Gamma, s) \ N(s) \cdot F(s) \ \mathrm{d}s \tag{36}$$

where $F$ is a vector field defined on $\Gamma$ and linked to $T$, $s$ is the arclength parameter of $\Gamma$, $G(\Gamma, s)$ is a short notation for $G(\Gamma, \Gamma(s))$, and $N$ is the inward unit normal of $\Gamma$.

Let us define $\mathcal{E}_t$ as follows

$$\mathcal{E}_t(\Gamma, w) = -\frac{1}{|\Omega|} \int_\Omega \log f_t(e_n(w, x)) \ \mathrm{d}x \ . \tag{37}$$

Hence,

$$E_t(\Gamma) = \mathcal{E}_t(\Gamma, v(\Gamma)) \tag{38}$$

and

$$v(\Gamma) = \arg\min_w \mathcal{E}_t(\Gamma, w) \ . \tag{39}$$

Then, the shape derivative of (32) is equal to

$$\mathrm{d}E_t(\Gamma, F) = \frac{\mathrm{d}\mathcal{E}_t}{\mathrm{d}\tau} \ (\tau, v(\tau))|_{\tau=0} \tag{40}$$

$$= \frac{\partial \mathcal{E}_t}{\partial \tau} \ (\tau, v(\tau))|_{\tau=0} + \frac{\partial \mathcal{E}_t}{\partial w} \ (\tau, v(\tau))|_{\tau=0} \ \frac{\mathrm{d}v}{\mathrm{d}\tau}(\tau = 0) \ . \tag{41}$$

Recalling that $\tau = 0$ corresponds to $\Gamma$ and according to (39), the second term in (41) is equal to zero. Therefore, expression (41) is equal to

$$\mathrm{d}E_t(\Gamma, F) = \frac{\partial \mathcal{E}_t}{\partial \tau}\ (\tau, v(\tau))|_{\tau=0}\ . \tag{42}$$

Note that the derivative of $\mathcal{E}_t$ is taken with respect to the first variable, $v(\tau)$ being considered as a constant (including in all the following calculations). The classical rule for differentiating a product leads to (see Section C.3 for the shape derivative of $\frac{1}{|\Omega|}$)

$$\begin{aligned}\mathrm{d}E_t(\Gamma, F) = \frac{E_t(\Gamma)}{|\Omega|} &\int_\Gamma N(s)\cdot F(s)\ \mathrm{d}s \\ &-\frac{1}{|\Omega|}\ \frac{\mathrm{d}}{\mathrm{d}\tau}\int_{\Omega(\tau)} \log f_t(e_n(v(\Gamma), x))\ \mathrm{d}x\Big|_{\tau=0} \quad (43)\\ = \frac{1}{|\Omega|}&\left[\int_\Gamma E_t(\Gamma)\ N(s)\cdot F(s)\ \mathrm{d}s - \mathcal{A}\right] \quad (44)\end{aligned}$$

where $f_t$ is also seen as a function of $\tau$

$$\begin{aligned}f_t(e_n(v(\Gamma), x)) = \frac{1}{|\Omega(\tau)|}\int_{\Omega(\tau)} &K_\sigma(e_n(v(\Gamma), y) \\ &- e_n(v(\Gamma), x))\ \mathrm{d}y\ . \quad (45)\end{aligned}$$

Remember that $v(\Gamma)$ is considered as a constant and not as a function of $\tau$ as a result of the decoupling (38). Therefore, for clarity, $e_n(v(\Gamma), \cdot)$ will be denoted by $e_n(\cdot)$.

Term $\mathcal{A}$ can be computed by applying the general rule (36) successively

$$\begin{aligned}\mathcal{A} &= \int_\Omega \frac{\partial \log f_t}{\partial \tau}(\tau = 0, x)\ \mathrm{d}x \\ &\quad - \int_\Gamma \log f_t(e_n(s))\ N(s)\cdot F(s)\ \mathrm{d}s \quad (46)\\ &= \int_\Omega \mathcal{B}\ \mathrm{d}x - \int_\Gamma \log f_t(e_n(s))\ N(s)\cdot F(s)\ \mathrm{d}s \quad (47)\end{aligned}$$

Then,

$$\begin{aligned}\mathcal{B} &= \frac{\frac{\partial f_t}{\partial \tau}(\tau = 0, x)}{f_t(e_n(x))} \quad (48)\\ &= \frac{1}{f_t(e_n(x))}\ \left[\frac{f_t(e_n(x))}{|\Omega|}\ \int_\Gamma N(s)\cdot F(s)\ \mathrm{d}s\right. \\ &\quad \left. + \frac{1}{|\Omega|}\ \frac{\mathrm{d}}{\mathrm{d}\tau}\int_{\Omega(\tau)} K_\sigma(e_n(y) - e_n(x))\ \mathrm{d}y\Big|_{\tau=0}\right] \quad (49)\\ &= \frac{1}{|\Omega|}\ \left[\int_\Gamma N(s)\cdot F(s)\ \mathrm{d}s + \frac{\mathcal{C}}{f_t(e_n(x))}\right]\ . \quad (50)\end{aligned}$$

Finally,

$$\begin{aligned}\mathcal{C} = \int_\Omega \frac{\mathrm{d}K_\sigma(e_n(y) - e_n(v(x))}{\mathrm{d}\tau}(\tau = 0)\ \mathrm{d}y \\ - \int_\Gamma K_\sigma(e_n(s) - e_n(v(x)))\ N(s)\cdot F(s)\ \mathrm{d}s \quad (51)\\ = -\int_\Gamma K_\sigma(e_n(s) - e_n(x))\ N(s)\cdot F(s)\ \mathrm{d}s \quad (52)\end{aligned}$$

since $K_\sigma(\ldots)$ does not depend on $\tau$. Gathering all the intermediate results together, the shape derivative of (32) is equal to

$$\begin{aligned}\mathrm{d}E_t(\Gamma, F) = \frac{1}{|\Omega|}\ \int_\Gamma \Big(&E_t(\Gamma) - 1 + \log f_t(e_n(s)) \\ &+ \frac{1}{|\Omega|}\ \int_\Omega \frac{K_\sigma(e_n(s) - e_n(x))}{f_t(e_n(x))}\ \mathrm{d}x\Big) \\ &N(s)\cdot F(s)\ \mathrm{d}s\ . \quad (53)\end{aligned}$$

## C.2 Spatial energy

The spatial energy is equal to

$$E_s(\Gamma) = -\frac{1}{|\Omega|}\ \int_\Omega \log f_s(I_n(x))\ \mathrm{d}x \tag{54}$$

where

$$f_s(r) = \frac{1}{|\Omega|}\ \int_\Omega K_\sigma(I_n(x) - r)\ \mathrm{d}x\ . \tag{55}$$

Following the same approach as in Section C.1, it can be shown that the shape derivative of (54) is equal to

$$\begin{aligned}\mathrm{d}E_s(\Gamma, F) = \frac{1}{|\Omega|}\ \int_\Gamma \Big(&E_s(\Gamma) - 1 + \log f_s(I_n(s)) \\ &+ \frac{1}{|\Omega|}\ \int_\Omega \frac{K_\sigma(I_n(s) - I_n(x))}{f_s(I_n(x))}\ \mathrm{d}x\Big) \\ &N(s)\cdot F(s)\ \mathrm{d}s\ . \quad (56)\end{aligned}$$

## C.3 Shape derivative of $|\Omega|$ and $\frac{1}{|\Omega|}$

The shape derivative of $|\Omega|$ is equal to

$$\begin{aligned}\mathrm{d}(|\Omega|)(\Gamma, F) &= \mathrm{d}\left(\int_\Omega \mathrm{d}x\right)(\Gamma, F) \quad (57)\\ &= \frac{\mathrm{d}}{\mathrm{d}\tau}\int_{\Omega(\tau)} \mathrm{d}x\Big|_{\tau=0} \quad (58)\\ &= \left[\int_\Omega \frac{\partial 1}{\partial \tau}(\tau = 0, x)\ \mathrm{d}x\right. \\ &\quad \left. - \int_\Gamma N(s)\cdot F(s)\ \mathrm{d}s\right] \quad (59)\\ &= -\int_\Gamma N(s)\cdot F(s)\ \mathrm{d}s\ . \quad (60)\end{aligned}$$

The shape derivative of $\frac{1}{|\Omega|}$ is equal to

$$\mathrm{d}(1/|\Omega|)(\Gamma, F) = \frac{\mathrm{d}}{\mathrm{d}\tau}\ \frac{1}{|\Omega(\tau)|}\Big|_{\tau=0} \tag{61}$$

$$= -\frac{1}{|\Omega|^2} \frac{\mathrm{d}}{\mathrm{d}\tau} \int_{\Omega(\tau)} \mathrm{d}x \Big|_{\tau=0} \quad (62)$$

$$= \frac{1}{|\Omega|^2} \int_{\Gamma} N(s) \cdot F(s) \, \mathrm{d}s \; . \quad (63)$$

# D Piecewise motion decomposition

The following development should give some intuitions to study the validity of the piecewise motion decomposition. As will be clear from the concluding remarks, it does not provide a full and rigorous analysis.

The frame $I_n$ is divided into blocks $B_i$ of identical size. Let $\Omega_i$ be the intersection of $\Omega$ with $B_i$ and let $\Gamma_i$ be the boundary $\partial\Omega_i$ of $\Omega_i$ (see Fig. 1). For clarity, $v(\Gamma_i)$ will be denoted by $v_i$. Energy (32) is replaced with

$$E_t^{\text{local}}(\Gamma) = -\frac{1}{|\Omega|} \int_{\Omega} \log f_t(e_n(v_1, \ldots, v_k, x)) \, \mathrm{d}x \quad (64)$$

where

$$\begin{cases} f_t(r) = \dfrac{1}{|\Omega|} \displaystyle\int_{\Omega} K_\sigma(e_n(v_1, \ldots, v_k, x) - r) \, \mathrm{d}x \\ e_n(v_1, \ldots, v_k, x) = \min_{\text{abs}}(I_n(x) - I_{n+1}(x + v_i), \\ \qquad\qquad I_n(x) - I_{n-1}(x - v_i)) \\ \text{if } x \in \Omega_i \\ v_i = \arg\min\limits_{w} -\dfrac{1}{|\Omega|} \displaystyle\int_{\Omega_i} \log f_t(e_n(v_1, \ldots, v_{i-1}, w, \\ \qquad\qquad v_{i+1}, \ldots, v_k, x)) \, \mathrm{d}x \end{cases} \quad (65)$$

Note that the motions $v_j, j \neq i$, in the energy minimized to solve for $v_i$ are irrelevant constants since they are not used in the computation of the residual $e_n$ on $\Omega_i$.

Let us define $\mathcal{E}_t^i$ as follows

$$\mathcal{E}_t^i(\Gamma, w_1, \ldots, w_k) = -\frac{1}{|\Omega|} \int_{\Omega_i} \log f_t(e_n(w_1, \ldots, w_k, x)) \, \mathrm{d}x \; . \quad (66)$$

According to the remark on the residual above, it can be concluded that $\mathcal{E}_t^i$ is independent of $w_j, j \neq i$.

Energy (64) is equal to,

$$E_t^{\text{local}}(\Gamma) = \sum_i \mathcal{E}_t^i(\Gamma, v_1, \ldots, v_k)) \quad (67)$$

and

$$v_i = \arg\min_{w} \mathcal{E}_t^i(\Gamma, v_1, \ldots, v_{i-1}, w, v_{i+1}, \ldots, v_k) \; . \quad (68)$$

Then, the shape derivative of (64) is equal to

$$\mathrm{d}E_t^{\text{local}}(\Gamma, F) = \sum_i \frac{\mathrm{d}\mathcal{E}_t^i}{\mathrm{d}\tau} (\tau, v_1(\tau), \ldots, v_k(\tau))|_{\tau=0} \quad (69)$$

$$= \sum_i \frac{\partial \mathcal{E}_t^i}{\partial \tau} (\tau, v_1(\tau), \ldots, v_k(\tau))|_{\tau=0} + \sum_i \sum_j \underbrace{\frac{\partial \mathcal{E}_t^i}{\partial w_j} (\tau, v_1(\tau), \ldots, v_k(\tau))|_{\tau=0}}_{\mathcal{A}_j^i} \times \frac{\mathrm{d}v_j}{\mathrm{d}\tau}(\tau = 0) \; . \quad (70)$$

Recalling that $\tau = 0$ corresponds to $\Gamma$ (and, therefore, $v_i(\tau = 0) = v_i$), $\mathcal{A}_j^i$ is equal to zero if $j$ is equal to $i$ because of (68). Moreover, according to the independence of $\mathcal{E}_t^i$ with respect to $w_j, j \neq i$, $\mathcal{A}_j^i$ is also equal to zero if $j$ is not equal to $i$. Therefore, expression (70) is equal to

$$\mathrm{d}E_t^{\text{local}}(\Gamma, F) = \sum_i \frac{\partial \mathcal{E}_t^i}{\partial \tau} (\tau, v_1(\tau), \ldots, v_k(\tau))|_{\tau=0} \; . \quad (71)$$

By definition, the shape derivative is based on a domain transformation $T$ operating on $\Omega$ (see Appendix C.1). Energy $\mathcal{E}_t^i$ is an integral over $\Omega_i$. Its shape derivative is naturally related with the restriction of $T$ to $\Omega_i$. However, $f_t$ is still an integral over $\Omega$. Keeping that in mind, the approach of Section C.1 can be followed to determine the shape derivative of (64)

$$\mathrm{d}E_t^{\text{local}}(\Gamma, F) = \frac{E_t^{\text{local}}(\Gamma)}{|\Omega|} \int_{\Gamma} N(s) \cdot F(s) \, \mathrm{d}s - \frac{1}{|\Omega|} \times \sum_i \frac{\mathrm{d}}{\mathrm{d}\tau} \int_{\Omega_i(\tau)} \log f_t(e_n(v(\Gamma), x)) \, \mathrm{d}x \Big|_{\tau=0} \quad (72)$$

$$= \frac{1}{|\Omega|} \left[ \int_{\Gamma} E_t^{\text{local}}(\Gamma) \; N(s) \cdot F(s) \, \mathrm{d}s - \sum_i \mathcal{A}_i \right] \; . \quad (73)$$

For clarity, $e_n(v_1, \ldots, v_k, \cdot)$ will be denoted by $e_n(\cdot)$.

Term $\mathcal{A}_i$ is equal to

$$\mathcal{A}_i = \int_{\Omega_i} \frac{\partial \log f_t}{\partial \tau}(\tau = 0, x) \, \mathrm{d}x - \int_{\Gamma_i} \log f_t(e_n(s)) \; N_i(s) \cdot F(s) \, \mathrm{d}s \quad (74)$$

$$= \int_{\Omega_i} \mathcal{B} \, \mathrm{d}x - \int_{\Gamma_i} \log f_t(e_n(s)) \; N_i(s) \cdot F(s) \, \mathrm{d}s \quad (75)$$

where $N_i$ is the inward unit normal of $\Gamma_i$. Term $\mathcal{B}$ is identical to the corresponding term (50) in Appendix C.1, *i.e.*,

$$\mathcal{B} = \frac{1}{|\Omega|} \int_\Gamma \left(1 - \frac{K_\sigma(e_n(s) - e_n(x))}{f_t(e_n(x))}\right) N(s)\cdot F(s)\ \mathrm{d}s\ . \tag{76}$$

Gathering all the intermediate results together, the shape derivative of (64) is equal to

$$\begin{aligned}\mathrm{d}E_t(\Gamma, F) = \frac{1}{|\Omega|} \Bigg[ & \int_\Gamma \left( E_t^{\text{local}}(\Gamma) - 1 \right. \\ & \left. + \frac{1}{|\Omega|} \int_\Omega \frac{K_\sigma(e_n(s) - e_n(x))}{f_t(e_n(x))}\ \mathrm{d}x \right) \\ & \times N(s)\cdot F(s)\ \mathrm{d}s \\ & \underbrace{- \sum_i \int_{\Gamma_i} \log f_t(e_n(s))\ N_i(s)\cdot F(s)\ \mathrm{d}s}_{\mathcal{S}} \Bigg]\ .\end{aligned} \tag{77}$$

Let $B_i$ and $B_j$ be 2 adjacent blocks with boundaries $\Gamma_i$ and $\Gamma_j$, respectively. On their common boundary, $\log f_t(e_n)$ and $F$ are uniquely defined. However, $N_i$ and $N_j$ have opposite directions, each pointing inward relatively to its (oriented) boundary. Therefore, the sum of the integrals over $\Gamma_i$ and $\Gamma_j$ in $\mathcal{S}$ on this common boundary is equal to zero. When considering all the blocks, the only portions of integral that remain of $\mathcal{S}$ are the ones which are not in common with any other block boundary. These portions sum to $\Gamma$. The normals $N_i$ on these portions are equal to $N$. In conclusion, shape derivative (77) is identical to (53): it seems that this hierarchical motion decomposition approach can be safely used with minimal changes to the implementation (only the residual computation changes). However, one condition has not been mentioned so far. The shape derivative framework is valid for smooth contours. In particular, the presence of the contour normal in the expressions implicitly requires that the contour be at least continuously differentiable. Unfortunately, the contours $\Omega_i$ of the proposed partition of $\Omega$ are not smooth, independently of the smoothness of $\Omega$. Actually, any paving of $\Omega$ using patches contains multiple junctions. As a consequence, the previous development is theoretically invalid. Nevertheless, the set of singularities is finite and it might be possible to rigorously confirm the result by studying the limit of a related, smooth setting similar to some works on classification [37]. Moreover, in practice, the (wrongly) obtained result can be easily implemented since it does not involve these singularities.

# E Parametric assumptions

## E.1 Residual

If the residual $e_n$ is a spatially uncorrelated random field with a Laplacian distribution with mean $\mu_e$ and scale $\sigma$, the probability of having a given field, conditional to a motion $v$, is equal to

$$p(e_n|v) = \frac{1}{(2\sigma)^{|\Omega|}} \prod_{x\in\Omega} \exp -\frac{|e_n(v,x) - \mu_e|}{\sigma}\ . \tag{78}$$

The maximum log-likelihood estimation of $v$ is given by

$$\arg\min_v \sum_{x\in\Omega} |e_n(v,x) - \mu_e|\ . \tag{79}$$

In practice, choosing $\mu_e$ different from zero can only be motivated by a global change of illumination occurring between frames $I_n$ and $I_{n+1}$. Making the assumption that the global illumination remains constant, $\mu_e$ will be set to zero. Therefore, estimation (79) is equivalent to

$$\arg\min_v \sum_{x\in\Omega} |I_n(x) - I_{n+1}(x+v)| \tag{80}$$

which is the SAD criterion.

## E.2 Color

If the color $I_n$ is a spatially uncorrelated random field with a Gaussian distribution with mean $\mu_I$ and standard deviation $\sigma$, the probability of having a given field, conditional to a motion $v$, is equal to

$$p(I_n|v) = \frac{1}{\sqrt{2\pi}\ \sigma} \prod_{x\in\Omega} \exp -\frac{(I_n(x) - \mu_I)^2}{2\sigma^2}\ . \tag{81}$$

The maximum likelihood estimation of $v$ is then equivalent to minimizing the Sum of Squared Differences (SSD)

$$\arg\min_v \sum_{x\in\Omega} (I_n(x) - \mu_I)^2\ . \tag{82}$$

In practice, $\mu_I$ can be approximated by the mean of $I_n$ in $\Omega$ [26].

# References

[1] I. A. Ahmad and P. E. Lin. A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Trans. Inform. Theory*, 22(3):372–375, 1976.

[2] L. Alvarez, J. Weickert, and J. Sánchez. A scale-space approach to nonlocal optical flow calculations. In *International Conference on Scale-Space Theories in Computer Vision*, Corfu, Greece, 1999.

[3] S. P. Awate and R. T. Whitaker. Unsupervised, Information-Theoretic, Adaptive Image Filtering for Image Restoration. *IEEE Trans. Pattern Analysis*, 28(3):364–376, 2006.

[4] G. Aubert, M. Barlaud, O. Faugeras, and S. Jehan-Besson. Image segmentation using active contours: Calculus of variations or shape gradients? *SIAM J. Appl. Math.*, 1(2):2128–2145, 2003.

[5] M. J. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. *Comput. Vis. Image Und.*, 63(1):75–104, 1996.

[6] S. Boltz, E. Debreuve, and M. Barlaud. High-dimensional statistical distance for region-of-interest tracking: Application to combining a soft geometric constraint with radiometry. In *International Conference on Computer Vision and Pattern Recognition*, Minneapolis (MN), USA, 2007.

[7] S. Boltz, A. Herbulot, E. Debreuve, and M. Barlaud. Entropy-based space-time segmentation in video sequences. In *ECCV Workshop on Statistical Methods in Multi-Image and Video Processing*, Graz, Austria, 2006.

[8] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision*, Prague, Czech Republic, 2004.

[9] T. Brox, M. Rousson, R. Deriche, and J. Weickert. Unsupervised segmentation incorporating colour, texture, and motion. In *International Conference on Computer Analysis of Images and Patterns*, Groningen, The Netherlands, 2003.

[10] A. Bruhn, J. Weickert, and C. Schnörr. Combining the advantages of local and global optic flow methods. In *DAGM Symposium on Pattern Recognition*, Zurich, Switzerland, 2002.

[11] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *Int. J. Comput. Vision*, 22(1):61–79, 1997.

[12] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Process.*, 6(2):298–311, 1997.

[13] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *International Conference on Computer Vision and Pattern Recognition*, Hilton Head Island (SC), USA, 2000.

[14] D. Cremers, S. J. Osher, and S. Soatto. Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. *Int. J. Comput. Vision*, 69(3):335–351, 2006.

[15] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. *Int. J. Comput. Vision*, 72(2):195–215, 2007.

[16] D. Cremers and S. Soatto. Variational space-time motion segmentation. In *International Conference on Computer Vision*,Nice, France, 2003.

[17] D. Cremers and S. Soatto. Motion competition: A variational framework for piecewise parametric motion segmentation. *Int. J. Comput. Vision*, 62(3):249–265, 2005.

[18] M. C. Delfour and J.-P. Zolésio. *Shapes and geometries: Analysis, differential calculus and optimization.* Advances in Design and Control. Society for Industrial and Applied Mathematics, Philadelphia, 2001.

[19] C. C. Dórea, M. Pardàs, and F. Marqués. Generation of long-term color and motion coherent partitions. In *International Conference on Image Processing*, Atlanta (GA), USA, 2006.

[20] A. Elgammal, R. Duraiswami, and L. S. Davis. Probabilistic tracking in joint feature-spatial spaces. In *International Conference on Computer Vision and Pattern Recognition*, Madison (WI), USA, 2003.

[21] D. Freedman and T. Zhang. Active contours for tracking distributions. *IEEE Trans. Image Process.*, 13(4):518–526, 2004.

[22] K. Fukunaga. *Introduction to statistical pattern recognition (2nd Ed.).* Academic Press Professional, Inc., 1990.

[23] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametr. Stat.*, 17(3):277–297, 2005.

[24] A. Herbulot, S. Boltz, E. Debreuve, and M. Barlaud. Robust motion-based segmentation in video sequences using entropy estimator. In *International Conference on Image Processing*, Atlanta (GA), USA, 2006.

[25] M. Hintermuller and W. Ring. A second order shape optimization approach for image segmentation. *SIAM J. Appl. Math.*, 64(2):442–467, 2004.

[26] S. Jehan-Besson, M. Barlaud, and G. Aubert. DREAM$^2$S: deformable regions driven by an eulerian accurate minimization method for image and video segmentation. *Int. J. Comput. Vision*, 53(1):45–70, 2003.

[27] J. Kim, J. W. F. Fisher, A. Yezzi, M. Çetin, and A. S. Willsky. A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Trans. Image Process.*, 14(10):1486–1502, 2005.

[28] M. Leventon, E. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contour. In *International Conference on Computer Vision and Pattern Recognition*, Hilton Head Island (SC), USA, 2000.

[29] M. E. Leventon, O. Faugeras, W. E. L. Grimson, and W. M. Wells III. Level set based segmentation with intensity and curvature priors. In *Workshop on Mathematical Methods in Biomedical Image Analysis*, Hilton Head Island (SC), USA, 2000.

[30] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, Vancouver, Canada, 1981.

[31] A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. In *International Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2004.

[32] H. Neemwuchwala and A. O. Hero. *Entropic Graphs for Registration*, Chapter 6, pages 185–235. Eds. R. S. Blum and Z. Liu, Marcel Dekker, Inc., 2005.

[33] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *J. Vis. Commun. Image R.*, 6(4):348–365, 1995.

[34] N. Paragios and R. Deriche. Geodesic active regions: A new framework to deal with frame partition problems in computer vision. *J. Vis. Commun. Image R.*, 13(1-2):249-268, 2002.

[35] E. Parzen. On the estimation of a probability density function and mode. *Ann Math Stat*, 33(3):1065–1076, 1962.

[36] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121, 2000.

[37] C. Samson, L. Blanc-Féraud, G. Aubert, and J. Zerubia. A level set model for image classification. *Int. J. Comput. Vision*, 40(3):187–197, 2000.

[38] T. Schoenemann and D. Cremers. Near real-time motion segmentation using graph cuts. In *DAGM Symposium on Pattern Recognition*, Berlin, Germany, 2006.

[39] D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

[40] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[41] J. Weickert and C. Schnörr. Variational optic flow computation with a spatio-temporal smoothness constraint. *J. Math. Imaging Vis.*, 14(3):245–255, 2001.

[42] W. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Med. Image Analysis*, 1(1):35–51, 1996.

[43] S. F. Wu and J. Kittler. A gradient-based method for general motion estimation and segmentation. *J. Vis. Commun. Image R.*, 4(1):25–38, 1993.

[44] A. Yezzi, A. Tsai, and A. Willsky. A fully global approach to image segmentation via coupled curve evolution equations. *J. Vis. Commun. Image R.*, 13(1):195216, 2002.

[45] S. Zhu and K.-K. Ma. A new diamond search algorithm for fast block-matching motion estimation. *IEEE Trans. Image Process.*, 9(2):287–290, 2000.