ELSEVIER

Signal Processing 66 (1998) 143–155

SIGNAL PROCESSING

Direct incremental model-based image motion segmentation for video analysis

Jean-Marc Odobez, Patrick Bouthemy*

IRISA/INRIA, Campus universitaire de Beaulieu, 35042 Rennes Cedex, France

Received 22 April 1997; revised 2 October 1997

Abstract

Dynamic analysis of image sequences is an important task in object-oriented video applications. It often relies on the segmentation of each image of the sequence into region entities of apparent homogeneous motion. In this paper, we present an original motion segmentation algorithm based on 2D polynomial motion models, a multiresolution robust estimator to compute these motion models, and appropriate local observations supplying both motion relevant information and their reliability. Motion segmentation is formulated as a contextual statistical labeling problem exploiting multiscale Markov random field (MRF) models. One of its main features is that it avoids time consuming alternate iterations between motion model estimation and spatial support identification. An original detection step allows us to estimate and to update the number of required motion models, and thus to handle the appearance of new objects. Numerous experiments performed with real indoor and outdoor image sequences demonstrate the efficiency of the method. © 1998 Published by Elsevier Science B.V. All rights reserved.

Zusammenfassung

Dynamische Analyse von Bildsequenzen ist eine wichtige Aufgabe in objektorientierten Videoanwendungen. Diese hängt oft von der Segmentierung jedes Bildes einer Sequenz in Bereiche offenbar gleicher Bewegung ab. In dieser Arbeit stellen wir einen Algorithmus zur Bewegungssegmentierung, der auf 2D polynomialen Bewegungsmodelle basiert, einen robusten Multiresolutions-Schätzer, um diese Bewegungsmodelle zu berechnen, und angemessene lokale Beobachtungen vor, die sowohl bewegungsrelevante Informationen, als auch deren Zuverlässigkeit unterstützen. Die Bewegungssegmentierung wird als ein kontextuelles, statistisches Zuordnungsproblem formuliert, daß multiskalen Markov Random Field (MRF) Modelle ausnutzt. Ein wesentliches Merkmal dieses Algorithmus ist, daß er zeitintensive wechselnde Iterationen zwischen Schätzung des Bewegungsmodells und räumlicher Support-Identifikation vermeidet. Ein Original-Detektionsschritt erlaubt es, die Anzahl der benötigten Bewegungsmodelle zu schätzen und zu aktualisieren und so das Auftreten neuer Objekte zu handhaben. Zahlreiche Experimente mit echten In- und Outdoor Bildsequenzen demonstrieren die Effizienz dieses Verfahrens. © 1998 Published by Elsevier Science B.V. All rights reserved.

Résumé

L'analyse dynamique de séquences d'images est une tâche importante dans les applications vidéo orientées objet. Elle s'appuie souvent sur la segmentation de chaque image de la séquence en régions de mouvement apparent

* Corresponding author. E-mail: bouthemy@irisa.fr.

0165-1684/98/$19.00 © 1998 Published by Elsevier Science B.V. All rights reserved.
PII S0165-1684(98)00003-6

homogène. Dans cet article, nous présentons un algorithme de segmentation original basé sur des modèles polynomiaux du mouvement 2D, un estimateur robuste multi-résolution permettant de calculer ces modèles, ainsi que des observations locales appropriées fournissant aussi bien l'information pertinente sur le mouvement que sur leur fiabilité. La segmentation du mouvement prend la forme d'un problème de classification statistique, en exploitant des modèles de champs de Markov aléatoires multi-échelles. Une de ses principales caractéristiques est d'éviter les itérations alternées entre l'estimation des modèles de mouvement et l'identification des supports spatiaux, pénalisantes en temps de calcul. Une approche originale de détection nous permet d'estimer et d'actualiser le nombre de modèles de mouvement requis, et de traiter ainsi l'apparition de nouveaux objets. De nombreuses expériences avec des séquences réelles d'intérieur et d'extérieur ont demontré l'efficacité de cette méthode.

Keywords: Motion segmentation; MRF models; Robust estimation; 2D parametric motion models

1. Introduction and related work

The apparent motion existing in an image sequence contains a rich source of information about camera motion and the composition of the viewed scene. Therefore, substantial research efforts have been devoted very early in the computer vision domain to motion analysis. Motion-based segmentation [4,13], i.e., the partitioning of the image in relevant regions that are homogeneous according to a given motion criterion, emerged as an essential tool in dynamic scene analysis applications. More recently, the need for motion segmentation algorithms arose in the image communication domain too [19]. Promising approaches for visual communication are now oriented towards content-based functionalities. This is of course the case for the MPEG-4 compression standard, but newly emerging objectives as in particular content based video indexing are also involving such requirements [9,18,20]. They therefore greatly depend on the ability of the analysis step to provide an object-based representation of the scene.

Motion-based segmentation plays two essential roles. First, motion segmentation naturally leads to motion compensation. The block effect observed with the more traditional schemes is much reduced with such a region-based approach. Secondly, it is useful to analyse and interpret the scene, as well as to extract the different moving parts visually important for the task at hand, sometimes referred to as layers [1,2]. The choice of motion as a cue for image segmentation allows us to deal with a small number of meaningful entities, as opposed for instance to a gray-level based segmentation approach, and to build a compact and structured representation of the spatio-temporal content of a video sequence.

Different techniques have been investigated for motion segmentation. A first category is composed of top-down hierarchical schemes [10,22], which consist in the computation of successive dominant motions. Significant areas consistent with the current dominant motion are associated with a single label, while the process is iterated on the remaining data. A drawback of these methods is that they generally break down in the absence of a well defined dominant motion. Moreover, the determination of the support layers greatly suffers from the lack of competition between different motion models to explain the motion measurements at pixel locations. Clustering methods [7,1,11], in a wide sense, fall into the second category. For instance, in [7], an unsupervised clustering technique mixes information on the position, color and motion-based residual at every pixel in a competitive learning scheme, whereas in [1], a k-mean technique is employed to group regions based on their pre-computed affine motion. One important shortcoming of these methods is that clustering in the parameter space is usually sensitive to the number of specified clusters. Besides, these methods generally start the segmentation process using elementary and somewhat arbitrary spatial areas. The resulting motion estimates, on which the clustering relies, are usually noisy. Moreover, these bottom-up algorithms [7,11,24], suited for the two-frame case, cannot easily incorporate the temporal aspect and the benefit from a predicted segmentation map.

The third category is composed of the techniques that address the segmentation issue in a Markovian framework, as a contextual labeling problem [4, 13,23].

Finally, a fourth related category is composed of other methods, that are based on mixture models [2,21], MDL criterion principle [2,24], or the robust regression framework [3]. In [2], motion models and spatial supports of the mixture are simultaneously estimated as the two steps of an EM algorithm. One drawback of mixture-based methods, also true for clustering methods, is that they do not inherently incorporate spatial coherence in the estimation of the spatial supports. Spatial smoothness is usually introduced in an ad-hoc way between iterations [2], or postponed to post-processing stages, as done for example in [11]. Recently, [21] proposed a formulation to overcome this problem. The main advantage of the MDL approach is that no threshold on a discrepancy measure needs to be set, while allowing the joint estimation of the number of motion models [2,24]. However, the coding optimality principle may not coincide with motion analysis purposes. In practice, algorithms using the MDL principle or mixture models need to start with a number of motion models greater than the actual one. As in clustering techniques, they cannot easily exploit a temporally predicted segmentation map, especially when the number of regions is growing. As pointed out in [2], no mechanism to allow for the variation of the number of regions in time has been proposed.

In this paper, we present an original and efficient motion-based segmentation method. In Section 2, the general features of our approach, its main contribution, and an overview of the algorithm are given. In Section 3, which forms the main part of this paper, we describe our motion segmentation algorithm. Results that validate our approach are reported in Section 4, while Section 5 contains concluding remarks.

2. General features and overview of the algorithm

Our motion-based segmentation algorithm relies on the use of 2D parametric motion models, the robust estimation of these models, and the introduction of multiscale Markov Random Field (MRF) models. The goal of the segmentation is to jointly estimate the motion models $\{(\Theta_k)_t^{t+1}\}$ between time t and time $t+1$ in each delimited region $R_k(t)$, $k \in \{1,\ldots,N_r(t)\}$, and the associated partition into regions represented by a label field $e(t)$ at time t, whose labels are in the set $\{1,\ldots,N_r(t)\}$. $N_r(t)$ stands for the number of regions in the image, and has to be estimated on-line also. It has also to be updated in time in accordance with the scene content. The time indexes will be dropped when there is no ambiguity.

To satisfy these requirements, we have designed a *direct* model-based segmentation algorithm which presents the following attributes:

1. A segmentation technique relies as much on the theoretical modeling as on the considered measurements. We thus pay attention to the definition of the latter, and look for local motion-related measurements that quantitatively indicate whether the modeled displacement $\boldsymbol{d}$ at pixel $s = (x,y)$ in region R_k, $\boldsymbol{d}_{\Theta_k}(s)$, constitutes a good approximation of the underlying true flow at this pixel, $\boldsymbol{d}_{\text{true}}(s)$. More precisely, we would like to have in each region R_k:

$$\forall s \in R_k, \quad \|\Delta\boldsymbol{d}(s,k)\| \leqslant \eta,$$

$$\text{with } \Delta\boldsymbol{d}(s,k) = \boldsymbol{d}_{\text{true}}(s) - \boldsymbol{d}_{\hat{\Theta}_k}(s), \tag{1}$$

where $\hat{\Theta}_k$ represents an estimate of Θ_k and η a constant desired accuracy. Since the computation of a reliable flow field is a difficult problem per se, we only resort to partial local motion measurements adequately defined as explained later, to locally assess the validity of the estimated model according to Eq. (1). This is the first meaning of the term ‘direct’ used to qualify our method.

2. Most approaches to region segmentation generally proceed in two steps that are iterated until convergence. The first one consists in estimating the motion models given the current image partition into regions; the second one in determining the optimal partition, the motion models being kept unchanged [1,2,4,21]. This optimization procedure happens to be computationally expensive, especially when stochastic minimization is utilized [13]. In our case, these

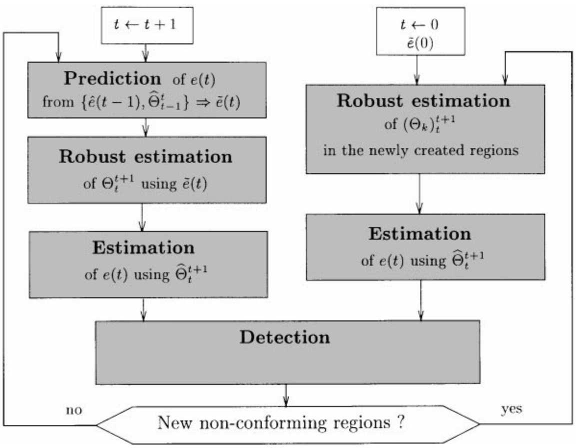
$t \leftarrow t+1$
Prediction of $e(t)$ from $\{\hat{e}(t-1), \widehat{\Theta}_{t-1}^{t}\} \Rightarrow \tilde{e}(t)$
Robust estimation of Θ_t^{t+1} using $\tilde{e}(t)$
Estimation of $e(t)$ using $\widehat{\Theta}_t^{t+1}$
$t \leftarrow 0$
$\tilde{e}(0)$
Robust estimation of $(\Theta_k)_t^{t+1}$ in the newly created regions
Estimation of $e(t)$ using $\widehat{\Theta}_t^{t+1}$
Detection
no
New non-conforming regions ?
yes

Fig. 1. Flow chart of the algorithm. $e(t)$ denotes the label map at time t, and $\Theta_t^{t+1} = \{(\Theta_k)_t^{t+1}\}$ is the set of motion models associated with $e(t)$ and accounting for the description of the motion field between time t and $t+1$.

successive alternate iterations are avoided thanks to the use of a robust multiresolution motion estimator, as explained below. This is the second motivation to present our motion segmentation method as a ‘direct’ one.

Fig. 1 shows a flow chart of the algorithm. The determination of the segmentation map at time t, given the segmentation map at time $t-1$, involves four main steps: the prediction of an initial segmentation map, the estimation of the motion models, the updating of the predicted map given the computed motion models, and finally the detection of new regions. This four step approach is now rather usual [4]; the way they are defined is original and efficient. The motion-based segmentation corresponding to the first image of the sequence is conducted in the same manner, starting with an arbitrary segmentation map (in practice, an initial map composed of one single region). Before describing the different processing steps in the next section, we introduce the motion model and briefly describe the robust motion estimation algorithm.

2.1. Motion model and estimation

We consider 2D parametric motion models to represent the projection of the 3D motion field of the different parts of the scene. Though less general than the full 3D rigid case, the choice of 2D models leads to an efficient motion computation scheme. In all the experiments we have carried out so far, the 2D affine motion model proved to be a good compromise between its relevance as a motion descriptor and the efficiency of its estimation. With this model, the displacement $\boldsymbol{d}_{\Theta_k}$ at point $s = (x,y)$ is described by

$$\boldsymbol{d}_{\Theta_k}(s) = \begin{bmatrix} a_1^k + a_2^k x + a_3^k y \\ a_4^k + a_5^k x + a_6^k y \end{bmatrix}, \tag{2}$$

where $\Theta_k = (a_i^k)$, $i = 1,\dots,6$, is the parameter vector to be estimated characterizing the motion in the region R_k. Using this model, each region of the segmentation can be interpreted as being the projection of a planar surface of the scene, provided that the slant of this plane is not too important.

To estimate this motion model between two successive frames I_t and I_{t+1}, we have developed a gradient-based multiresolution incremental robust estimation method described in [15]. To ensure the goal of robustness, we minimize an M-estimator criterion with a hard-redescending function [8]. The constraint is given by the usual assumption of brightness constancy of a projected surface element over its 2D trajectory. In order to handle large displacements, we adopted an incremental approach. Given a current estimate $\hat{\Theta}_k$ of the motion parameters, an increment $\Delta\Theta_k$ is computed according to

$$\widehat{\boldsymbol{\Delta\Theta_k}} = \underset{\Delta\Theta_k}{\operatorname{argmin}} \sum_{s \in R_k} \rho(r_{\Delta\Theta_k}(s,\hat{\Theta}_k)) \tag{3}$$

with

$$\begin{aligned} r_{\Delta\Theta_k}(s) = {} & \nabla I_{t+1}(s + \boldsymbol{d}_{\hat{\Theta}_k}(s)).\boldsymbol{d}_{\Delta\Theta_k}(s) \\ & + I_{t+1}(s + \boldsymbol{d}_{\hat{\Theta}_k}(s)) - I_t(s), \end{aligned} \tag{4}$$

where $\rho(x)$ is a function which is bounded for high values of x, and $\nabla I_{t+1}(s)$ denotes the spatial gradient of the intensity function at location s and at time $t+1$. The incremental minimization is conducted within a multiresolution framework. For more details about the method and its performances, the reader is referred to [15].

3. Segmentation algorithm

The main goal of our segmentation algorithm, stated in Eq. (1), is to recover coherent regions where the estimated motion models account for the true motion field up to a given precision η. Another way to formulate this goal is to state that we aim at discriminating regions whose 'motion difference' is greater than η. The principal steps of the algorithm ensuring the constraint (Eq. (1)) are the *updating step*, and the *detection step* (cf. Fig. 1), which localizes subregions where this constraint is violated.

To achieve this goal, we utilize only locally measured quantities that can be *straightforwardly* computed from the images. They supply valuable information accounting for the motion compensation accuracy obtained using the estimated motion models. Since such local measurements can be too noisy or insufficient to reach a correct decision, we state both the updating and detection issues as statistical contextual labeling problems. Thanks to such regularization schemes, information on motion compensation errors at reliable points (e.g., corners) will be propagated to points where ambiguities might exist (e.g., straight edge line) or points with no information (uniform areas), as explained below.

3.1. Definition of local motion-related measurements

Experience shows that the displaced frame difference expression, given by

$$\mathrm{DFD}(s,\Theta_k) = I_{t+1}(s + \boldsymbol{d}_{\Theta_k}(s)) - I_t(s), \tag{5}$$

is not a reliable measure to assess whether a modeled flow $\boldsymbol{d}_{\hat{\Theta}_k}(s)$ is a 'good' or a 'bad' approximation of the true flow. Indeed, in uniform intensity areas, the response of this measure is always very low, whatever the accuracy of the estimated motion model is; along highly contrasted edges, the response is large whenever there exists even a small residual motion. As a more appropriate measure, we have adopted a weighted average of the normal residual flow $\Delta d_n(s,k)$, given by $|\mathrm{DFD}(s,\hat{\Theta}_k)|/\|\nabla I(s)\|$, over a 3×3 neighborhood $\mathcal{N}(s)$ of site s. As weights, we take $\|\nabla I\|^2$, since the spatial image gradient is often considered as a good indicator of the relevance of the normal flow measurements. Thus, we consider the following measurement:

$$\varepsilon_s(k) = \frac{\sum_{p\in\mathcal{N}(s)}|\mathrm{DFD}(p,k)| \times \|\nabla I(p)\|}{\max(9\times G_m^2, \sum_{p\in\mathcal{N}(s)}\|\nabla I(p)\|^2)}, \tag{6}$$

where G_m is a predetermined constant related to the noise level in uniform areas.

An interesting property of this local measure is that at each site s, we can derive two bounds l_s and L_s, such that

$$\begin{aligned} \varepsilon_s(k) < l_s &\Rightarrow \|\Delta\boldsymbol{d}(s,k)\| < \eta, \\ \varepsilon_s(k) > L_s &\Rightarrow \|\Delta\boldsymbol{d}(s,k)\| > \eta. \end{aligned} \tag{7}$$

Let us point out that l_s and L_s do not depend on the motion models, but only on the preset parameter η and the local distribution of the directions of the local spatial intensity gradients. Thus, they allow us to adapt the meaning of a measure to the local intensity structure. For example, at a linear isointensity contour s, the bound l_s is zero, indicating that even very low measurement $\varepsilon_s(k)$ cannot be trusted as expressing the motion adequacy. Indeed, in this example, due to the aperture problem, a motion model k may be consistent with the motion of the contour (i.e. $\varepsilon_s(k) \simeq 0$), although there is no coincidence between the true velocity and the velocity obtained with this model.

Bounds l_s and L_s, in the general case, are given by

$$(l_s,L_s) = (\tau\times\eta\times\lambda'_{\min},\eta) \tag{8}$$

with

$$\tau = \frac{\sum_{p\in\mathcal{N}(s)}(\|\nabla I(p)\|^2)}{\max(9\times G_m^2,\sum_{p\in\mathcal{N}(s)}(\|\nabla I(p)\|^2))} \quad \text{and}$$

$$\lambda'_{\min} = \frac{\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}, \tag{9}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are, respectively, the smallest and highest eigenvalues of the following matrix (with $\nabla I(p) = (I_x(p),I_y(p))$):

$$M = \begin{bmatrix} \sum_{p\in\mathcal{N}(s)} I_x(p)^2 & \sum_{p\in\mathcal{N}(s)} I_x(p)I_y(p) \\ \sum_{p\in\mathcal{N}(s)} I_x(p)I_y(p) & \sum_{p\in\mathcal{N}(s)} I_y(p)^2 \end{bmatrix}. \tag{10}$$

By exploiting a modelisation of the local spatial intensity gradient distribution, tighter bounds can be obtained, as reported in [16].

3.2. Algorithm steps

We now describe in more details each step of the motion segmentation algorithm.

3.2.1. Step 1: prediction map determination

The prediction map of the partition at time t, denoted $\tilde{e}(t)$, is determined using the segmentation map along with the estimated motion models obtained at time $t-1$. This step, along with step 3, allows us to supply a coherent labeling of the same moving scene element in the successive partitions over time. The label k present at site s in the computed label map at time $t-1$, $\hat{e}(t-1)$, is assigned to each point on the grid around $s + \boldsymbol{d}_{(\hat{\Theta}_k)^t_{t-1}}(s)$ in $\tilde{e}(t)$. Pixels that receive no label are given a special label, as well as pixels that get several labels. They respectively correspond to uncovered regions between $t-1$ and t, and to occlusion areas, which are thus effectively and straightforwardly handled in our method.

3.2.2. Step 2: robust motion estimation

The motion models $\Theta_t^{t+1} = \{(\Theta_k)_t^{t+1}\}$ between image I_t and image I_{t+1} are estimated using the initial partition $\tilde{e}(t)$. Since we use a robust estimator, an imprecise predicted map or the appearance of new objects do not perturb this estimation process. Therefore, parameters $(\Theta_k)_t^{t+1}$ are computed only once, according to the incremental scheme outlined previously, where the estimation support is now $\tilde{R}_k(t)$, the kth predicted region support at time t.

3.2.3. Step 3: updating of the predicted partition

The estimation of the optimal partition $\hat{e}(t)$ given the predicted map $\tilde{e}(t)$ and the estimated models $(\hat{\Theta}_k)_t^{t+1}$, is achieved through a statistical regularization approach based on *multiscale* Markov Random Field (MRF). We adopt a Bayesian MAP criterion, which, due to the use of MRF models, leads to the minimization of an energy function $U(e,o,\tilde{e})$. o is the field of observations and is composed of the images I_t and I_{t+1}, from which we can compute the bounds (l_s,L_s) and the local measurements $\varepsilon_s(k)$. We have established an energy function comprising three terms:

$$U(e,o,\tilde{e}) = U_1(e,o) + U_2(e) + U_3(e,\tilde{e}), \tag{11}$$

where each term, that can be written as the sum of local potential functions, are described hereafter.

- We have paid particular attention to the data-driven term U_1 expressing the adequacy between the labels and the observations. Since the role of this term is to indicate, *given the local intensity structure*, which motion model is 'adequate', and which is not, the potentials involved in this energy term U_1 has to convert somehow the inequalities (Eq. (7)) into an energy-based formulation. This is done as follows:

$$U_1(e,o) = \sum_{s\in S} \alpha . F_s . V_1(e_s,o) \tag{12}$$

with

$$V_1(e_s,o) = A_{l_s}(\varepsilon_s(e_s)) - (1 - A_{L_s}(\varepsilon_s(e_s))) \tag{13}$$

where $A_{tr}(x)$ is a smooth version of a step edge. We use the normalized arctangent: $A_{tr}(x) = (1/\pi)\arctan(2\pi(x - tr)) + 0.5$ instead of a sigmoide, because it reaches saturation levels 0 or 1 less rapidly. $F_s = \max(A_G(\|\nabla I(s)\|),\mathrm{At}_{\max})$ is a damping factor. A site with low image gradient usually carries poor and unreliable information about the adequacy of a given motion model. The role of the damping factor F_s is to reduce the amplitude of the energy term V_1 provided by the observations at such a site, which *conversely* increases the relative contributions of the regularization terms. The parameter $\mathrm{At}_{\max}$ fixes the minimal value allowed to avoid over-damping.

Let us comment the choice of V_1. First, since the potential energy function V_1 is bounded, it behaves similarly to a 'robust estimator'. It avoids isolated strong spurious mesurements to have a sufficient influence to locally impose the wrong label even if all the neighbours 'disagree'. Secondly, the potential function V_1 reflects the inequalities on the measurements. It is illustrated in Fig. 2 where we have plotted this potential function for two representative cases. A very low potential value should indicate that the discrepancy between the true and the modelled motion at site s is below a given predefined value η. In a straight edge configuration, due to the aperture problem, a measurement,

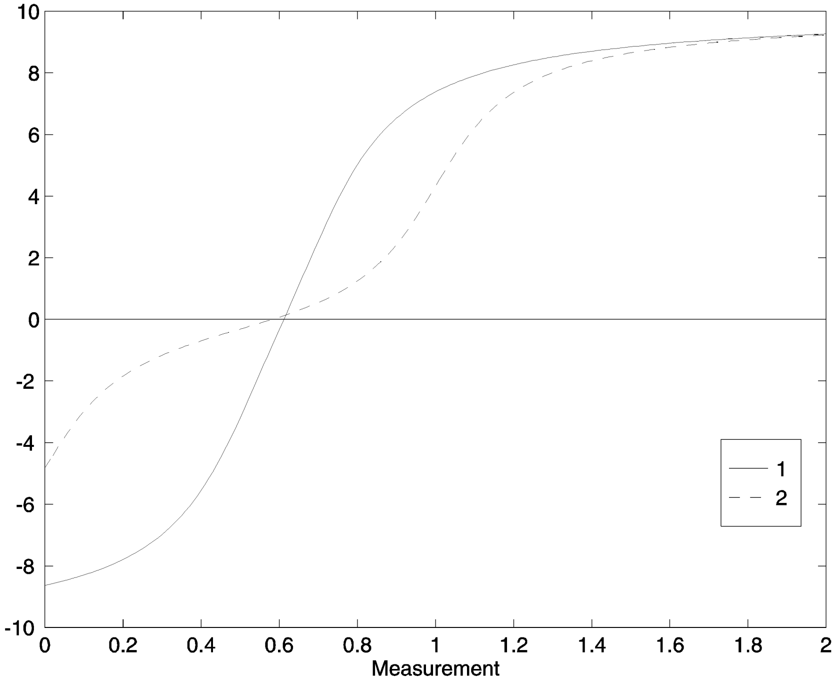

Fig. 2. Potential function V_1 for two specific configurations: case of a site located on a corner (curve 1), and on straight edge (curve 2).

whatever low it is, cannot indicate with certainty whether the discrepancy magnitude is below η. Therefore, the potential value does not fall as low as in other cases, like in the corner configuration, where the aperture problem does not arise.

When locally there exists an ambiguity, energy terms U_2 and U_3 described below introduce the contextual information necessary to remove it and to perform a correct labeling.

- Energy term U_2 accounts for the expected spatial properties (homogeneity) of the label field. It has the following expression, which is a usual one:

$$U_2(e) = \sum_{(s,t)\in\mathscr{C}} \beta_d(1 - \delta_{e_s = e_t}) \tag{14}$$

  where $\mathscr{C}$ represents the set of cliques of two elements associated to a second order neighbourhood system, and δ is the Kronecker function.

- U_3 favours the conservation of labels over time, except in occluded and uncovered regions between I_{t-1} and I_t where no label is favoured. It is given by

$$U_3(e,\tilde{e}) = \sum_{s\in S} F_s \times \beta_{dt}(1 - \delta_{e_s = \tilde{e}_s}) \tag{15}$$

This step of the algorithm performs the updating of the boundaries between two regions R_k and $R_{k'}$, taking into account the new motion model estimates. When completed, the image segmentation into regions of homogeneous motion is achieved using the number of motion models intervening in the segmentation at the previous instant. The purpose of the next step is to test: (a) if new moving objects have appeared in the scene; (b) if the current number of motion models is still adequate to provide a good description of the apparent motion in the image.

3.2.4. Step 4: detection of new regions

Within each region R_k, sub-areas whose motion does not conform to the estimated motion model

$\hat{\Theta}_k$ are detected. This is achieved using a scheme similar to the one described in [14], which was concerned with the detection of moving regions not conforming with the global dominant motion model estimated in the whole image. Briefly, this detection phase uses a statistical regularization leading to the minimization of an energy function U_d similar to Eq. (11). In particular, the data driven term in U_d can be directly computed from Eq. (13) at no cost, leading to an extremely fast detection procedure.

The significant connected components are extracted from the set of all the detected sub-areas, and the number of regions $N_r(t)$ is updated accordingly. If there is no new region, the final partition at time t is that obtained at the end of the last relaxation iteration performed at step 3. Otherwise, the motion models, *in the newly created regions only*, are estimated, still using the multiresolution robust estimator, and the partition is updated again according to the relaxation scheme described in step 3. The relaxation then involves only few computations, since it is only concerned with the adjustment of the boundaries of the newly created regions.

3.3. Dealing with occlusions

In the motion-based segmentation of a scene using two images only, a problem arises with areas at time t that are occluded at time $t+1$. In these areas, no correct local measurements can be extracted. The energy term U_1 has been extended to deal with occluded area. The most intuitive way to correctly label these areas consists in looking towards the past. In addition to motion estimates from time t to time $t+1$, $(\hat{\Theta}_k)_t^{t+1}$, we have also considered motion models from t to $t-1$, $(\Theta_k)_t^{t-1}$. Then, we can define local motion measurements $\varepsilon_s'(k)$ involving these models and the images I_t and I_{t-1}, in a way similar to the definition of $\varepsilon_s(k)$. When the right label k is assigned to s, at least one of the measurement $\varepsilon_s(k)$ or $\varepsilon_s'(k)$ should be small, whereas with a wrong label, both should be high. Consequently, the energy term U_1 is simply modified by replacing $\varepsilon_s(e_s)$ in Eq. (13) by: $\min(\varepsilon_s(e_s),\varepsilon_s'(e_s))$.

3.4. Computational issues

The minimization of the energy functions (steps 3 and 4) is performed using the multiscale approach described in [17]. At a given scale j, the solution, constrained to be constant within blocks of size $2^j \times 2^j$, is computed using the Highest Confidence First [6] minimization procedure. Experiments have shown that both the use of the multiscale approach and the HCF minimization algorithm (instead of the standard ICM) improve the results. Especially, the multiscale scheme reinforces the homogeneity constraint without having to overweigh the corresponding a priori energy term.

Parameter setting is an important issue in any image processing algorithm. Results should not exhibit a high sensitivity to the choice of parameter values. In our algorithm, no parameter setting appeared to be crucial. Indeed, results shown in the next section are all obtained with the same set of parameter values in the different steps of the segmentation algorithm, except two parameters. The values of the unchanged parameters are the following: $\alpha = 200$, $\beta_d = 20$, $\beta_{dt} = 40$, $G = 1$, $\mathrm{At}_{\max} = 0.5$, and we use four levels in the multiscale minimization algorithm.

The first variable parameter, G_m, is related to the image quality, and is not sensitive. Its value is 6 for the 'van' sequence, and 3 for the four other sequences. The second one is the parameter η involved in the constraint (1), and which represents the precision with which a motion model in a given region represents the underlying true motion field. From its definition, it is clear that it influences the number of motion models needed to describe the true flow field, and therefore the number of regions in the segmentation. It offers a real flexibility to the method, and is a typical user-defined parameter of obvious physical interpretation, that can easily be set according to the needs and requirements.

Finally, let us note that, prior to the updating step (step 3), a procedure merging regions undergoing similar motions is performed. However, only very few region merging events happen indeed (approx. one every 10 images). This procedure is required to overcome specific problems that occur in the processing of a whole sequence [16]. The main one stems from the variability of the motion complexity along a video sequence. When this complexity decreases, less motion models are required to describe the flow field, and thus regions must be merged. The merging procedure relies on the estimated motion models in a simple way and is very fast. It is developed in [16].

4. Experimental results

The proposed algorithm has been validated on many real sequences comprising indoor and outdoor scenes. Here, we give results obtained on five different sequences. Boundaries of the segmentation are overprinted in white on the original images. The value of the precision parameter η is given in the caption. We also indicate, for the 'Van', 'Renata' and 'Mobile' sequences the range of values for parameter η leading to almost identical results. With the current non-optimized implementation of the algorithm, the whole algorithm has a computational cost of 15–20 s per image on a Sun Sparc 20, for images of size 256 × 256. However, this cost could be greatly reduced since most of the calculus are local and regular.

In the 'Van' sequence (Fig. 3(a–d)), the scene takes place at a crossroad. A white van, followed by a black car driving at almost the same speed, is coming from the left of the scene and going to the right. A white car is coming from the opposite

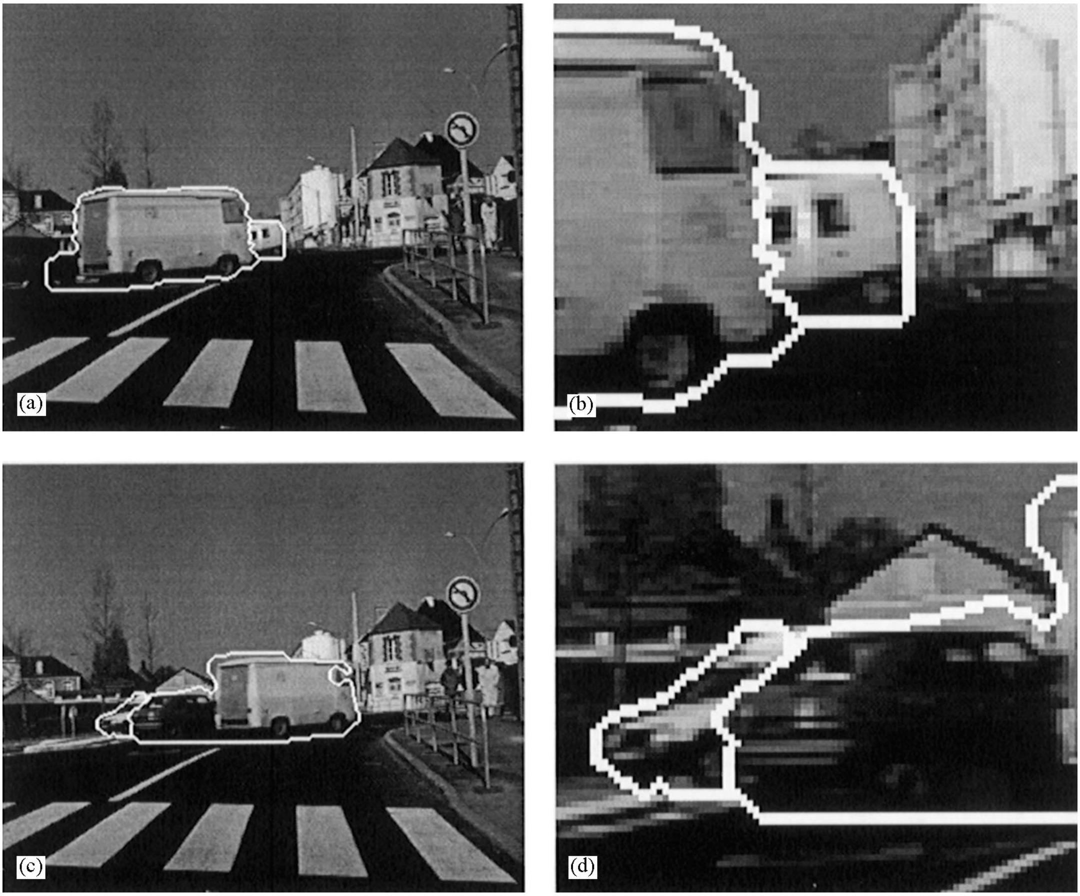

Fig. 3. 'Van' sequence: (a) Segmentation map obtained at time t_{34}. (b) Enlarged detail of (a). (c) Segmentation map obtained at time t_{54}. (d) Enlarged detail of (c). $\eta = 0.8$. We obtain similar results for values of η ranging from 0.5 to 1.5.

direction. The results demonstrate the accuracy of our algorithm in determining the motion discontinuities, even when the white car is getting occluded (Fig. 3(b)). The creation of a new region is perfectly realized as soon as the front of the white car reappears, as shown by Fig. 3(c,d).

In the 'Garden' sequence, the camera is translating to the right, while the scene is static. Several motion models are needed to structure this scene into its different layers. An affine model is a reasonable approximation of the motion of rather fronto-parallel planar surfaces or shallow objects. The garden sequence illustrates nicely this property. The ground as well as the houses in the background can be considered as planar surfaces with different orientations, while the two regions in the tree can be considered as shallow objects. Boundaries between regions correspond not only to motion or depth discontinuities but also to orientation discontinuities, as shown by the estimated modeled flow field displayed in Fig. 4(b).

Fig. 5(a,b) displays the results obtained for the 'Renata' sequence. The small swing of the woman arm, as well as the slight nod of the head at the beginning of the sequence are taken into account. Let us note that segmentation boundaries usually extend outside the moving object border, especially when the background is uniform. This effect is due to the facts that we only rely on motion information, and that of course all motions are valid in

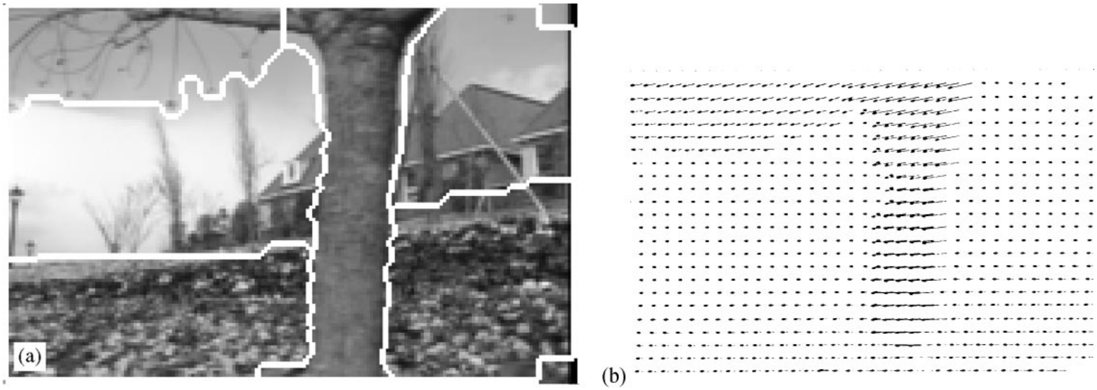

Fig. 4. 'Garden' sequence: segmentation map at time t_2 (a) and associated estimated model flow fields (b).

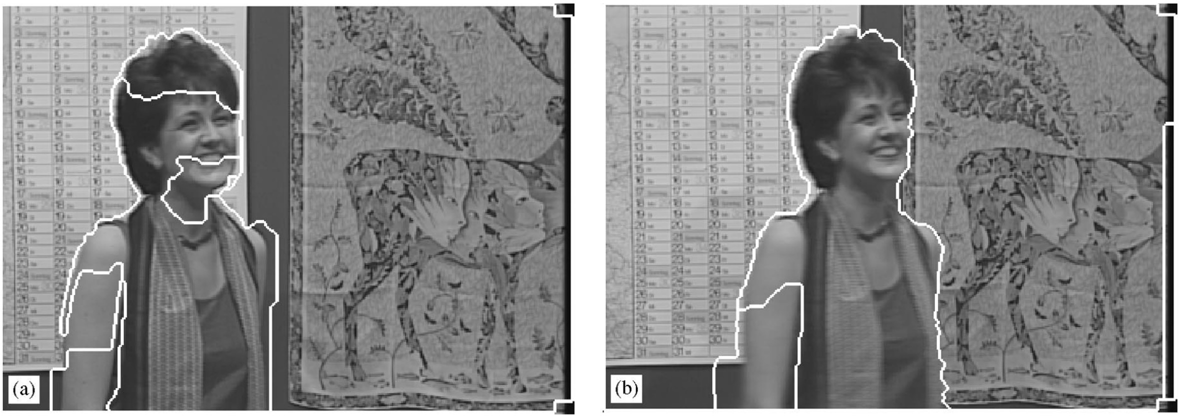

Fig. 5. 'Renata' sequence: segmentation map obtained at time t_1 (a) and t_{15} (b). $\eta = 0.5$. We obtain similar results for values of η ranging from 0.4 to 1.0.

uniform areas. However, as soon as there exists some texture on both sides of a motion discontinuity, we recover this discontinuity with very high precision. This is illustrated for instance by comparing the boundary location near the woman left arm in Fig. 5(a,b), or by the boundary around the train in the 'Mobile' sequence, Fig. 6(a,b). In this example, the camera is panning the scene while the calendar is sliding vertically, the ball is rolling and the train is moving forward.

In the 'Interview' sequence, the camera is tracking the woman on the right, who is standing up while moving her arms. Besides, casted shadows of these arms are sliding over her pants. Fig. 4(a,b) emphasis the role of the precision parameter η, and displays results obtained with two different values of η. As expected, a larger value of η is more suitable for motion analysis, as in that case the algorithm captures only the few principal motion components of the scene. A smaller value (Fig. 7(b)) can be used if a better description of the motion field is preferred, for motion-compensation purposes in image coding for instance. It usually leads to the recovery of more regions, as in this example, where the

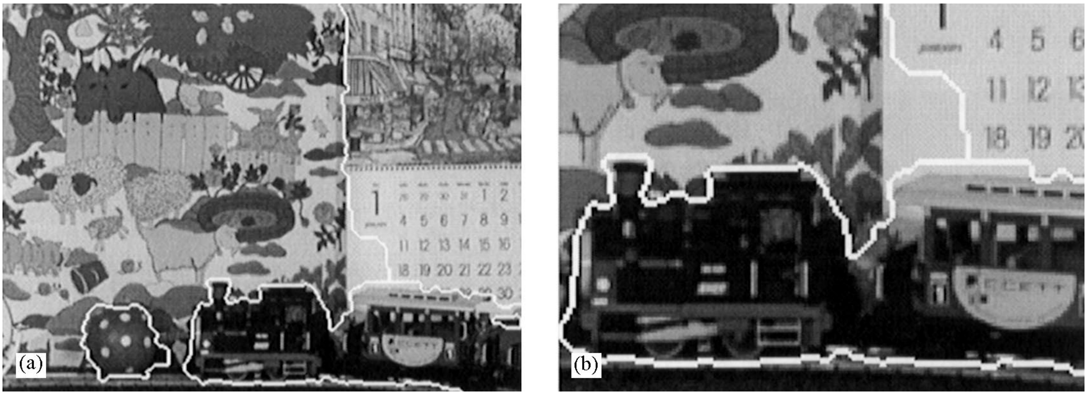

Fig. 6. 'Mobile' sequence. (a) Segmentation map obtained at time t_1. (b) Enlargement of one part of the segmentation map obtained at time t_4. $\eta = 0.6$. We obtain similar results for values of η ranging from 0.4 to 1.5.

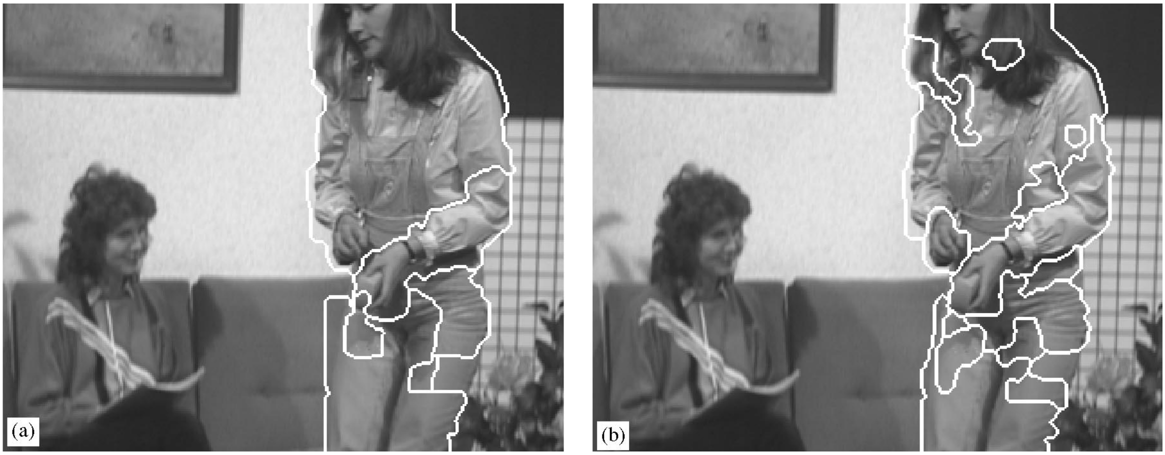

Fig. 7. 'Interview' sequence. Motion-based segmentation map obtained at time t_{37} with two different 'precision' levels: (a) $\eta = 1.25$ and (b) $\eta = 0.75$.

algorithm creates regions to account for the motion of the right hand or the hairs.

5. Conclusion

We have described a motion-based segmentation method based on 2D motion models. The method aims at recovering a relevant partition of each image of the sequence, where in each region the estimated motion model is able to describe the true motion field with a predefined precision level.

The algorithm relies, apart from the prediction step, on three essential steps. The first one is the motion estimation step, based on a multiresolution robust estimator which enables for the computation of accurate motion model estimates using a predicted segmentation map, even in the presence of prediction errors and changes of scene content like the appearance of new objects. As a consequence, this step is performed only once. The second step uses local appropriate *motion-related* measurements, and is embedded in a multiscale MRF framework to favour spatio-temporal consistency of the segmentation maps. A key feature is that the aperture problem is explicitly and directly acknowledged in this framework, allowing us to differentiate between informative sites and non-informative ones. Moreover, the scheme has been extended to handle occlusions. Finally, the third step achieves an efficient detection of new regions. Experiments carried out on many different sequences have demonstrated the robustness and the validity of our approach.

As illustrated by the results, our algorithm constitutes a powerful tool for the extraction of relevant information from video sequences. The short term temporal link provided by our algorithm can be efficiently exploited to build long term object trajectories, even in the presence of long occultation [12]. Background motion estimates can be exploited to compute mosaic or key frames [9] as well as to identify and to characterize different shots of a video sequence [5]. The use of motion and trajectory of objects along with the detection of dynamic events (e.g. object emergence or disappearance, object stopping, occultation) to build storyboard-like representation is currently studied.

References

[1] E.H. Adelson, J.Y.A Wang, Representing moving images with layers, IEEE Trans. Image Process. 3 (5) (September 1994) 625–638.

[2] S. Ayer, H.S. Sawhney, Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding, in: Proc. IEEE Internat. Conf. Computer Vision, Boston, June 1995, pp. 777–784.

[3] M. Black, A. Jepson, Estimating optical flow in segmented images using variable-order parametric models with local deformations, IEEE Pattern Anal. Machine Intell. PAMI 18 (10) (October 1996) 972–996.

[4] P. Bouthemy, E. François, Motion segmentation and qualitative dynamic scene analysis from an image sequence, Internat. J. Comput. Vision 10 (2) (April 1993) 157–182.

[5] P. Bouthemy, F. Ganansia, Video partitioning and camera motion characterization for content-based video indexing, in: Proc. 3rd ICIP, September 1996, pp. 905–909.

[6] P.B. Chou, C.M. Brown, The theory and practice of Bayesian image modeling, Internat. J. Computer Vision 4 (1990) 185–210.

[7] M. Etoh, Y. Shirai, Segmentation and 2D motion estimation by region fragments, in: Proc. 4th IEEE Internat. Conf. on Computer Vision, Berlin, May 1993, pp. 192–199.

[8] P.J. Hubert, Robust Statistics, Wiley, New York, 1981.

[9] M. Irani, P. Anandan, J. Bergen, R. Kumar, S. Hsu, Efficient representation of video sequences and their applications, Signal Processing: Image Communication 8 (1996) 327–351.

[10] M. Irani, B. Rousso, S. Peleg, Detecting and tracking multiple moving objects using temporal integration, in: Proc. 2nd ECCV, May 1992, pp. 282–287.

[11] S.-M. Kruse, Scene segmentation from dense displacement vector fields using randomized Hough transform, Signal Processing: Image Communication 9 (1996) 29–41.

[12] F.G. Meyer, P. Bouthemy, Region-based tracking using affine motion models in long image sequences, CVGIP: Image Understanding 60 (2) (September 1994) 119–140.

[13] D.W. Murray, H. Buxton, Scene segmentation from visual motion using global optimization, IEEE Trans. Pattern Anal. Machine Intell., PAMI 9 (2) (March 1987) 220–228.

[14] J.-M. Odobez, P. Bouthemy, Detection of multiple moving objects using multiscale MRF with camera motion compensation, in: Proc. 1st ICIP, November 1994, pp. II:257–261.

[15] J.-M. Odobez, P. Bouthemy, Robust multiresolution estimation of parametric motion models, J. Vis. Comm. Image Representation 6 (4) (December 1995) 348–365.

[16] J-M. Odobez, P. Bouthemy, Direct incremental model-based image motion segmentation for video analysis, Technical Report 1129, IRISA/INRIA-Rennes, October 1997.

[17] P. Pérez, F. Heitz, P. Bouthemy, Multiscale minimization of global energy functions in some visual recovery problems, CVGIP: Image Understanding 59 (1) (January 1994) 125–134.

[18] H. Sawhney, S. Ayer, M. Gorkani, Model-based 2D and 3D dominant motion estimation for mosaicing and video representation, in: Proc. 5th ICCV, 1995, pp. 583–590.

[19] G. Tziritas, C. Labit, Motion analysis for image sequence coding, in: Advances in Image Communication, Vol. 4, Elsevier, Amsterdam, 1994.

[20] H. Ueda, T. Miyatake, S. Yoshizawa, Impact: an interactive natural-motion picture dedicated multimedia authoring system, in: Proc. ACM Conf. INTERCHI'91, 1991, pp. 343–350.

[21] Y. Weiss, E. Adelson, A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models, in: Proc. IEEE Conf. Computer Vision Pattern Recognition, 1996, pp. 321–326.

[22] S.F. Wu, J. Kittler, A gradient-based method for general motion estimation and segmentation, J. Visual Comm. Image Representation 4 (1) (March 1993) 25–38.

[23] W. Xiong, C. Graffigne, A hierarchical method for detection of moving objects, in: Proc. 1st IEEE Internat. Conf. Image Processing, Austin, TX, November 1994, pp. 795–799.

[24] H. Zheng, D. Blostein, Motion-based object segmentation and estimation using the MDL principle, IEEE Trans. Image Process. 4 (9) (September 1995) 1223–1235.