Video Segmentation by MAP Labeling of Watershed Segments

Ioannis Patras, *Student Member*, *IEEE*,
E.A. Hendriks, and
R.L. Lagendijk, *Senior Member*, *IEEE*

Abstract—This paper addresses the problem of spatio-temporal segmentation of video sequences. An initial intensity segmentation method (watershed segmentation) provides a number of initial segments which are subsequently labeled, with a known number of labels, according to motion information. The label field is modeled as a Markov Random Field where the statistical spatial and temporal interactions are expressed on the basis of the initial watershed segments. The labeling criterion is the maximization of the conditional a posteriori probability of the label field given the motion hypotheses, the estimate of the label field of the previous frame, and the image intensities. For the optimization, an iterative motion estimation-labeling algorithm is proposed and experimental results are presented.

Index Terms—Markov Random Fields, motion-based segmentation, region labeling, watershed segmentation, motion estimation.

1 Introduction

One of the key issues in the design of many vision systems is their ability to decompose image sequences into the "objects" that are depicted in them. Motion information is one of the main elements that are used for segmenting video sequences. However, extracting and coupling motion information with the segmentation process is by no means a trivial task. For the estimation of motion, spatial constraints need to be imposed in a form of a support region where the motion is assumed either to be smooth or to follow a parametric model. In general, if the region of support is arbitrarily chosen then the motion estimate will be deteriorated either because the single motion assumption within the region is violated or because the texture pattern is too low to constrain enough the estimation. Furthermore, in the motion-based segmentation framework issues like the occlusions and the temporal coherency of the segmentation mask need to be addressed.

In this paper, we present an approach in which spatial and temporal constraints are incorporated into a single framework to allow the joint estimation of the segmentation field and of the motion information. The method operates in two levels (Fig. 1). At the lower level (LEVEL 1 in Fig. 1), a segmentation algorithm operating on the current frame's intensities provides a set of segments with relatively small intensity variation. At the next level, (LEVEL 2 in Fig.1) these segments are grouped into regions that move with the same motion parameters by assigning an "object" label to each segment. We use the well-known notion of Markov Random Fields (MRF) in order to express spatial and temporal constraints at the level of the "intensity" segments. The labeling criterion is the maximization of the conditional a posteriori probability (MAP) of the label field given the motion hypotheses, the label field of the previous frame, and the image intensities. For the optimization procedure, we propose a method which minimizes the corresponding objective function in an iterative way with respect to the motion parameters and the label field. A three frame approach is adopted in order to deal with occlusions.

• *The authors are with the Information and Communication Theory Group, Delft University of Technology (TU Delft), P.O. Box 5031, 2600 GA, Delft, The Netherlands.*
E-mail: {I.Patras, E.A.Hendricks, R.L.Langendijk}@its.tudelft.nl.

Manuscript received 24 Apr. 2000; revised 20 Sept. 2000; accepted 6 Nov. 2000.
Recommended for acceptance by M. Irani.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 111989.

In comparison to other works in the area of motion-based segmentation, our method is mostly related to two categories. To the first category belong methods which simultaneously estimate the motion information and its region of support. Depending on if the label field is explicitly defined, temporal and spatial constraints are imposed either on motion [1], [2], [3], [4] and/or on the label field itself [5], [6], [7]. Our work can be regarded as an extension of methods of this category that define cliques at pixel level in the Markovian framework and jointly estimate the motion and the label field. We exploit the ability of such approaches to incorporate the spatial and temporal constraints in the optimization procedure. However, by defining cliques on segment level, we provide tighter constraints for the labeling and reduce the dimensionality of the problem. The initial intensity segmentation groups together pixels in which the low degree of texture implies inadequate information about their temporal behavior. These segments are more reliable entities than pixels used as primary elements for the labeling problem. The relation of our approach with existing pixel-based methods will become more apparent once the modeling and the optimization procedure are described. It can be shown [8] that in the degenerate case where the segments that result from the initial segmentation contain a single pixel, our method reduces to an approach that falls in this category.

To the second category belong methods that combine an initial intensity segmentation with motion information. From the prism of our method, we distinguish between the following four general directions: To the first one belong the top-down approaches [9], [10]. To the second one, methods in which a region merging process is driven by motion-based distance measures [11], [12], [13], [14], [15]. To the third direction belong methods that utilize an initial intensity segmentation in order to incorporate spatial constraints in the Expectation Maximization framework [16], [17]. Finally, to the last one belong methods that combine the MRF modeling with an initial segmentation [18], [19], [20]. From the above mentioned approaches, the work of Fablet et al. [9], and Gelgon and Bouthemy [20], [19] is the most related to our proposal. However, their way of combining the motion information with the labeling is quite different. The dominant motion estimation/outlier detection paradigm which is adopted in [9] has the shortcomings of the hierarchical approaches. Such top-down approaches are faced with the problem of estimating the dominant motion in the presence of multiple independent motion patterns and, furthermore, they impose an artificial hierarchy in determining the motion characteristics of the objects which may lead in situations where outlier segments do not belong to any object [14]. In [20], [19], motion is estimated independently per segment. Estimating motion parameters per segment requires sufficient local intensity structure which often implies that the size of segment should be rather large. In search for sufficient texture, the initial intensity segmentation method might violate significant borders. In our approach, where a region-based motion estimation is employed, it is not crucial if some of the segments do not provide sufficient constraints. The ensemble of the constraints in the whole region is what determines the accuracy of the motion estimation. Furthermore, in both [20] and [19], the temporal constraints are introduced only in the initialization phase for the prediction of the initial label field. In comparison, in our approach, the temporal constraints are incorporated in the optimization procedure itself.

The remainder of the paper is organized as follows: Section 2 briefly discusses the initial intensity segmentation method. Sections 3 and 4 contain the formulation of the labeling problem in the MAP framework and the optimization procedure, respectively. In Section 5, experimental results are presented and, finally, conclusions are drawn in Section 6.

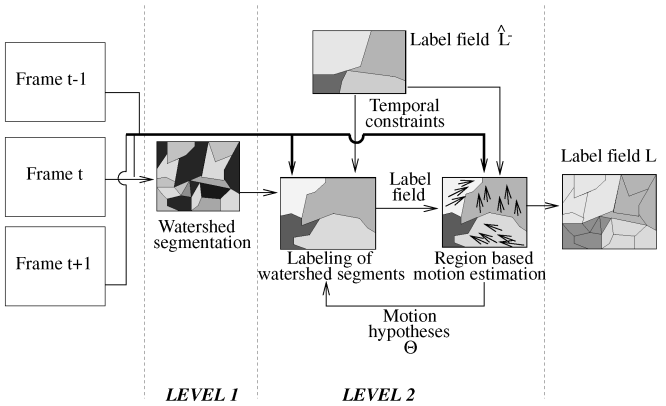


Fig. 1. Outline of the approach.

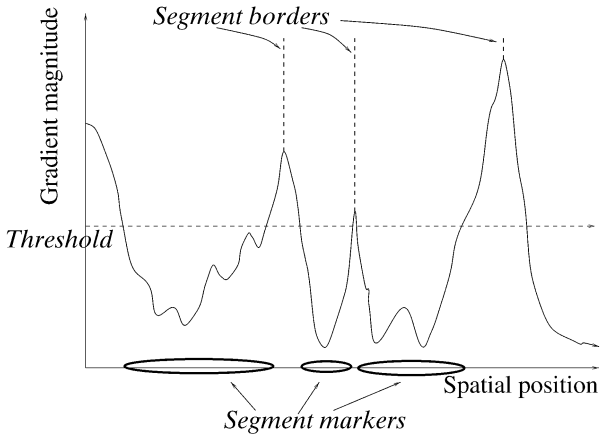


Fig. 2. Initial intensity segmentation in the 1D case.

2 Intensity Segmentation

At the lower level of the proposed method (Fig. 1), a segmentation algorithm is applied on the intensities in the current frame. We aim for a conservative partitioning of the current frame, such that significant boundaries, that is object boundaries, are not violated. That is, we favor an oversegmentation since the proposed method is not able to recover from an initial undersegmentation by splitting a segment that does not belong entirely to a single object. Although the choice of the segmentation method is not restrictive to the generality of our approach, we favor methods which consider the intensity gradient rather than clustering approaches. For its low-computational complexity and good edge localization accuracy, we use the watershed segmentation algorithm [21]. A filtering with morphological operators [22] with a small (3×3) structuring element is used for a nonlinear smoothing of the current frame. Once the noise level is reduced, the morphological gradient is estimated and segment *markers* are extracted as areas where the gradient is lower than a threshold. The flooding procedure described by Vincent and Soille [23] provides the final partition (Fig. 2).

The threshold for the marker extraction is a user-specified prediction of the smallest gradient magnitude of the significant edges. Edges with smaller gradient magnitude are not preserved. It should be noted that the threshold is not directly related with the amount of texture within a segment. During the flooding procedure, a segment will encapsulate some of the pixels which lie between its marker and the marker of the neighboring segment and have higher gradient magnitude than the threshold (Fig. 2).

Once the intensity segments are extracted, a Region Adjacency Graph (RAG) can be built to express neighborhood relations between them. We denote $\{s : s \in [1 \dots K]\}$ as the set of the watershed segments, G_s as the set of the pixels in the watershed segment s, and $N_s = \{s'\}$ as the set of neighbors of segment s as they are defined on the Region Adjacency Graph.

3 Problem Modeling

We consider the supervised framework where the number of independently moving objects in the scene, denoted by N, is considered as known. We assume that the 2D apparent motion field induced by them can be approximated by 6-parameter affine models. We seek for the unknown label field $L = \{l_s : l_s \in [1 \dots N], s \in [1 \dots K]\}$ and the motion hypotheses $\Theta = \{\theta_n : n \in [1 \dots N]\}$ at time instant t, considering as known the estimate $\hat{L}^-$ of the label field at the previous time instant. We consider the Bayesian framework and, more specifically, we adopt, as the labeling criterion, the maximization of the a posteriori probability (MAP) of the label field. The conditional probability distributions to which the MAP criterion decomposes are modeled as Gibbs distributions. This MAP-MRF framework has been used extensively for regularization and for expressing contextual constraints in numerous problems in computer vision [3], [24].

More specifically, we aim for the maximization of the a posteriori probability:

$$P(L|I,\Theta,\hat{L}^-,I^-,I^+) \propto P(I|L,\Theta,\hat{L}^-,I^-,I^+)P(\hat{L}^-|L,\Theta)P(L|\Theta) \quad (1)$$

with respect to L and Θ. With I^-, I, and I^+, we denote the image intensities in the previous, current, and next frame, respectively.

The first term on the right-hand side of (1) is the conditional probability distribution $P(I|L,\Theta,\hat{L}^-,I^-,I^+)$, which expresses how well the current motion and label field conform with the image intensities. We model it as a Gibbs distribution where the energy term, denoted with $E_d(I,L,\Theta,I^-,I^+)$, is defined as the sum of local Gibbs potentials $V_{ds}(I,s,\theta_{l_s},I^-,I^+)$. The local Gibbs potentials $V_{ds}(I,s,\theta_{l_s},I^-,I^+)$ are defined over single-site(segment) cliques as follows:

$$V_{ds}(I,s,\theta_{l_s},I^-,I^+) = \min\left(\sum_{\mathbf{i}\in G_s}\left(f_{\mathbf{i}}^-(\theta_{l_s})\right)^2, \sum_{\mathbf{i}\in G_s}\left(f_{\mathbf{i}}^+(\theta_{l_s})\right)^2\right), \quad (2)$$

where $f_{\mathbf{i}}^+(\theta_{l_s})$ and $f_{\mathbf{i}}^-(\theta_{l_s})$, respectively, are the forward and backward motion compensated intensity differences at pixel $\mathbf{i}$ $(\mathbf{i} \in G_s)$.

Note that we are using a three frame approach, where the motion compensated intensity differences are defined on the basis of segments either in the previous or in the next frame using the $\min$ operator. By doing so, we are dealing in a simple and efficient way with appearing and disappearing areas. The underlying assumption is that these areas are visible in at least two consecutive frames, that is each watershed segment has a correspondence either in the next or in the previous frame. A similar approach, where a visibility set for each pixel is defined in a forward/backward way, is presented by Dubois and Konrad [25]. In their work, the direction in which the motion estimation is constrained is determined by an occlusion field. In our segment based approach, the direct association of an occlusion field with the direction of the motion is not trivial since it may be the case that segments are only partially occluded. However, our assumption that each segment has a correspondence either in the next or in the previous frame, is violated only when a segment is large enough to partially appear and partially disappear from the scene. In this case, the relative size and texture structure of the visible

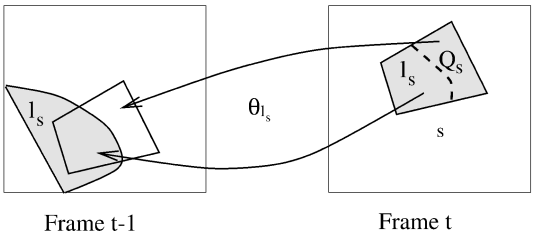


Fig. 3. The temporal projection of watershed segment s (gray polygon in Frame t). The shaded area in frame $t-1$ represents a *region* labeled with label l_s.

area will determine the accuracy of the temporal evidence for the segment in question. In such cases, the spatial constraints need to provide the additional cues for the correct labeling.

The second term on the right-hand side of (1) expresses the temporal constraints in terms of how well the estimate of the label field in the previous frame $\hat{L}^-$ conforms with the motion hypotheses Θ and the label field L in the current frame. The conditional probability $P(\hat{L}^-|L,\Theta)$ is modeled as a Gibbs distribution. The corresponding energy term $E_t(L,\Theta,\hat{L}^-)$ is defined as the sum of local energy terms $V_{ts}(\hat{L}^-,s,\theta_{l_s})$. The local energy terms are defined over single-site(segment) cliques as follows:

$$V_{ts}(\hat{L}^-,s,\theta_{l_s}) = \sum_{\mathbf{i}\in G_s}\left(O^-_{l_s}(\mathbf{i}-\mathbf{v_i}(\theta_{l_s}))\right)^2, \tag{3}$$

where $\mathbf{v_i}(\theta_{l_s})$ is the motion vector at pixel i that is generated under the motion hypothesis θ_{l_s}. An object field O^-_n is defined for each object n. O^-_n is equal to 0 at all the points that belonged to object n at the previous frame and equal to $\sqrt{z_t}$ at all other points. Each pixel of the watershed segment s is projected in the object field $O^-_{l_s}$ of the previous frame using the motion parameters θ_{l_s}. It is easy to verify that $V_{ts} = z_t Q_s$, where the number Q_s is the number of pixels of segment s whose motion-based projections in the previous frame have a label different than l_s (Fig. 3). The term z_t is a constant that controls the temporal consistency of the label field. From a global point of view, an optimization with respect to the spatial energy term results in a label field L and a set of motion hypotheses Θ such that each region n in the current frame is projected entirely within a region with the same label in the previous frame.

Finally, the third term on the right hand side of (1) models the probability of the label field. We model it as a Gibbs distribution whose energy term $E_c(L)$ is the sum of spatial clique potentials $V_c(s,s')$ which are defined over pair-site(segment) cliques as follows:

$$V_c(s,s') = \begin{cases} -z_c b(s,s') & \text{if } l_s = l_{s'} \\ z_c b(s,s') & \text{if } l_s \neq l_{s'}, \end{cases} \tag{4}$$

where the segments s and s' are neighbors in the neighborhood system N_s defined on the Region Adjacency Graph. The term z_c is a constant that controls the weight of the spatial constraints relative to the temporal constraints and to the constraints that the intensity preservation principle imposes. The term $b(s,s')$ denotes the length of the common border between s and s'. It is estimated as the number of pairs of pixels $(\mathbf{i},\mathbf{i}')$ which are neighbors in the image grid and belong to the borders of s and s', respectively. From a global point of view, an optimization with respect to the spatial energy term $E_c(L)$ tends to minimize the total border length between neighboring objects.

The parameters z_c and z_t control the relative weights that the spatial and the temporal constraints have in the estimation of the label field with respect to the data energy term E_d. So far as the parameter z_t is concerned, our formulation implies that a balance in the relative influence of the data energy in correspondence with the temporal energy is achieved when z_t is set according to the expected variance of the motion compensated intensity differences. However, the correctness of the influence of the temporal constraints depends on the degree of the accuracy of the estimated label field $\hat{L}^-$ in the previous frame. In order to insure that the algorithm is flexible enough to correct errors in $\hat{L}^-$, the value of z_t should be chosen rather conservatively. In all of our experiments, the value of z_t was chosen in the range between 1 and 2.

So far as the spatial energy term is concerned, we note that at segment level the spatial energy term is proportional to the perimeter of the segment s while the data energy term and the temporal energy term is proportional to the number of pixels of s. In general, this implies that the larger a segment is, the larger the ratio between the relative contribution of the local data and temporal energy terms with respect to the local spatial energy term. Thus, for larger segments, the emphasis is placed on the evidence that the segments themselves provide about their temporal behavior, while for smaller segments the emphasis is placed on the evidence provided by the label field in their neighborhood. For the manual setting of the value of z_c, a rule of thumb can be provided by an analysis of the ratio between the size and the perimeter of the segments. Taking into consideration that the data energy term has the characteristics of the variance of the motion compensated intensity differences and that the temporal energy term is scaled by the factor z_t, we chose values of z_c in the range between 3 and 10 for our experiments. In that range of values, the different energy terms are roughly normalized for segments with around 100 pixels size.

4 MAP Estimation

Once the energy functions are defined, the MAP estimation is equivalent to the minimization of the quantity

$$\begin{aligned} E(L,\Theta,I,\hat{L}^-,I^-,I^+) = \\ E_d(I,L,\Theta,I^-,I^+) + E_t(L,\Theta,\hat{L}^-) + E_c(L). \end{aligned} \tag{5}$$

In order to solve the nonlinear optimization problem of (5), we propose a method which iterates between a minimization with respect to the label field L (labeling phase) and a minimization with respect to the motion parameters Θ (motion estimation phase). In the labeling phase, a relaxation algorithm is employed to solve the combinatorial problem of assigning object labels to watershed segments. In the motion estimation phase, (5) is linearized with respect to Θ and a gradient-based approach is adopted. More specifically, the MAP estimation iterates between the following phases:

$$L_{m+1} = \arg\min_{L} E(L,\Theta_m,I,\hat{L}^-,I^-,I^+) \tag{6}$$

$$\Theta_{m+1} = \arg\min_{\Theta} E(L_{m+1},\Theta,I,\hat{L}^-,I^-,I^+), \tag{7}$$

where m denotes the iteration index. For the first iteration (i.e., for $m=0$), the motion hypotheses are initialized with the motion parameters estimated for the previous frame.

The optimization procedure described by (6) and (7) bears similarities with the *Expectation-Maximization* method where a hard classification is employed. Indeed, it can be shown [8] that our method can be formulated as an EM algorithm with hard decisions and that with a minor modification can incorporate soft decisions too.

4.1 Labeling Phase

In the labeling phase, (6), the minimization of (5) with respect to L takes place, keeping the motion parameters Θ "frozen" ($\Theta = \Theta_m$). We consider an iterative deterministic relaxation algorithm known as Iterative Conditional Modes (ICM). Proposed by Besag [26], ICM

TABLE 1
Modified ICM for Segment Labeling

1. $c_0 = \{s : s \in [1 \dots K]\}$, $c_1 = \emptyset$, $k = 0$
2. Choose randomly a segment s from c_k
3. Assign to s the label l that minimizes the local energy:

$$V_{ds}(I, s, \theta_l) + \sum_{s' \in N_s} V_c(s, s') + V_{ts}\left(\hat{L}^-, s, \theta_l\right)$$

4. $c_k = c_k - \{s\}$
5. If s has changed label update the candidate list for the next iteration

$$c_{k+1} = c_{k+1} \cup \{s' : s' \in N_s\}$$

6. If $c_k \neq \emptyset$ goto step 2
7. If $c_{k+1} \neq \emptyset$ then $k = k + 1$ and goto step 2
   else STOP

maximizes the conditional probability of a label at each site iteratively, given the labeling at all other sites. In the original algorithm at each iteration, each site is visited and is assigned the label that maximizes that conditional probability. This can be regarded as a scheme in which approximations of the conditional probabilities of the labels are estimated and hard decisions are employed.

As stated in [26], the order in which the sites are visited is important for the final configuration. Furthermore, a site can contribute to the reduction of the energy only if the local labeling configuration has changed, that is, at each iteration not all of the sites should be visited. In order to cope with the latter, we maintain a set of candidate segments c_k for each iteration k *within* the labeling phase. In order to avoid the influence of a predetermined ordering, we adopt a random visit schedule on the elements of the candidate set. The steps of the algorithm are summarized in Table 1.

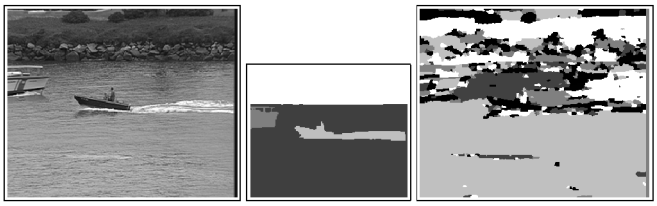
An initialization of the label L is estimated by an optimization of the data and temporal energy terms $(E_d + E_t)$ with respect to L. Since no spatial constraints are introduced at this point, the labeling is performed independently for each segment.

4.2 Motion Estimation Phase

In the motion estimation phase, the minimization of (5) with respect to Θ takes place, keeping the label field L "frozen" $(L = L_m)$. This minimization is a nonlinear optimization problem since neither E_d nor E_t are linear with respect to the motion parameters. For E_d this is because: first, the image intensities are nonlinear with respect to Θ and, second, because of the nonlinear minimum operator in (2). In order to overcome the latter, we turn the minimization of (5) into an equivalent optimization problem by introducing a binary **direction** field $\{d_s : d_s \in [0, 1], s \in [1 \dots K]\}$. This field determines the direction, backward or forward, in which the temporal intensity variation constrains the motion estimation. Let us define $C_e(d, \Theta)$ as:

$$C_e(d, \Theta) = E_t(L, \Theta, \hat{L}^-) + \sum_{s=1}^{K}\left(d_s \sum_{\mathbf{i} \in G_s}\left(f_{\mathbf{i}}^+(\theta_{l_s})\right)^2 + (1 - d_s)\sum_{\mathbf{i} \in G_s}\left(f_{\mathbf{i}}^-(\theta_{l_s})\right)^2\right), \quad (8)$$

where the functional dependence of $C_e(d, \Theta)$ on the label fields and on the image intensities is omitted for notational simplicity. The new energy term $C_e(d, \Theta)$ is derived from the terms of (5) that depend on Θ. It is almost straightforward to show [8] that if $(\hat{\Theta}, \hat{d})$ are the arguments which minimize (8), then $\hat{\Theta}$ is the argument which minimizes (5).



(a) (b) (c)

Fig. 4. The 14th frame of "Coastguard" sequence, the corresponding validation label field, and the watershed segmentation.

We minimize (8) with a method which iterates between a minimization with respect to Θ and with respect to d. More specifically,

$$d^{k+1} = \arg\min_d C_e\left(d, \Theta^k\right) \quad (9)$$

$$\Theta^{k+1} = \arg\min_\Theta C_e\left(d^{k+1}, \Theta\right), \quad (10)$$

where k is the iteration index *within* the motion estimation phase. Θ^0 are the motion hypotheses obtained at convergence in the previous motion estimation phase. For the first motion estimation phase in the current frame, Θ^0 are the motion hypotheses estimated for the previous frame.

Clearly, the minimization of $C_e\left(d, \Theta^k\right)$ with respect to d yields

$$d_s = \begin{cases} 1 & \text{if } \sum_{\mathbf{i} \in G_s}\left(f_{\mathbf{i}}^+\left(\theta_{l_s}^k\right)\right)^2 \leq \sum_{\mathbf{i} \in G_s}\left(f_{\mathbf{i}}^-\left(\theta_{l_s}^k\right)\right)^2 \quad \text{(forw.)} \\ 0 & \text{otherwise} \quad \text{(back.)} \end{cases} \quad (11)$$

For the minimization of $C_e\left(d^{k+1}, \Theta\right)$ with respect to Θ, we use first order Taylor approximations of $f_{\mathbf{i}}^-(\theta)$ (or $f_{\mathbf{i}}^+(\theta)$) and $O_n^-(\mathbf{i} - \mathbf{v}_{\mathbf{i}}(\theta_n^k))$ after smoothing with a Gaussian filter with a small variance. This results in a well-known form of the optical flow constraint. In order to solve for the motion parameters in an incremental way [27], we first express the motion parameters as $\Theta = \Theta^k + \Delta\Theta$. At iteration $k + 1$, we solve for the $\Delta\Theta$ for which the gradient of C_e with respect to the motion parameters is zero.

$$\nabla C_e\left(d^k, \Theta^k + \Delta\Theta\right) = 0. \quad (12)$$

Equation 12 results in N linear systems with six unknowns, one for each of the N sets of motion parameters θ_n.

5 Results

We have applied the proposed algorithm in a number of sequences in order to test the validity of our approach. We present results for three sequences in each of which different challenges arise. In the first one, the apparent motions of the objects are quite small in magnitude which makes the distinction between them rather difficult. In the second one, the motions are large, a fact which generates large occlusions and even blurs the edges of one of the objects. Finally, in the third sequence, difficulties arise on the one hand because of small deformations in the shape of the moving object and on the other because of the large rotational components which are present in the motion pattern. Extended results for these sequences can be found in [8].

5.1 "Coastguard" Sequence

For the MPEG validation sequence "Coastguard," Fig. 4a and Fig. 4b depict an original frame and the corresponding validation labeling mask which is used only for illustrative purposes. Given

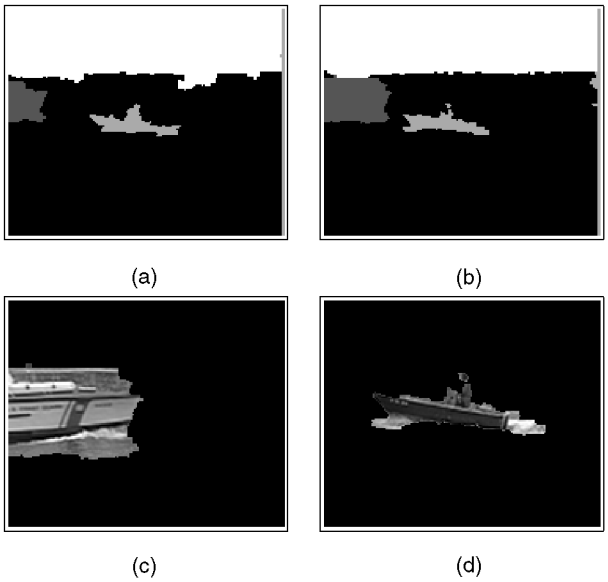


Fig. 5. The "Coastguard" sequence: (a) The label field at the 10th frame. (b) The label field at the 20th frame. (c) Object mask for the left ship at the 20th frame. (d) Object mask for the middle boat at the 20th frame ($z_c = 3, z_t = 1$).

this labeling mask, four different objects are present. The camera follows the ship in the middle, while another ship is entering the scene. The water of the river globally appears to move to the right, but deviations from the dominant motion pattern occur locally. The motion behavior of the different objects is quite similar; a distinction between the "Shore" and the "Water" is possible only at subpixel level.

The result of the watershed segmentation is depicted in Fig. 4c, where an area with constant intensity represents a watershed segment. Our primary goal of obtaining a well-localized, edge preserving segmentation is achieved. Each watershed segment belongs entirely to a single object, a result which validates our choice of a small structuring element.

In Fig. 5, we present results obtained at convergence for $N = 4$. The algorithm is capable of distinguishing the different objects in the scene by successfully grouping the watershed segments into regions that move in the same way and produces temporally coherent label fields. Both of the ships are well-localized and the "Water" is well-separated from the "Shore." The main difference with the "ground truth" segmentation of Fig. 4b remains the trail of the ship in the water. However, it is questionable if it is possible, without any semantic reasoning, to classify with the same label the ship and a trail whose apparent movement is quite arbitrary.

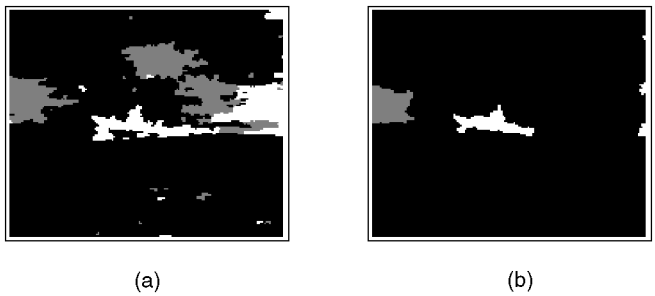


Fig. 6. Label fields under the assumption that three objects are present in the scene. (a) Label field for the 10th frame without spatial and temporal constraints. (b) Label field for the 10th frame with spatial and temporal constraints ($z_c = 3, z_t = 1$).

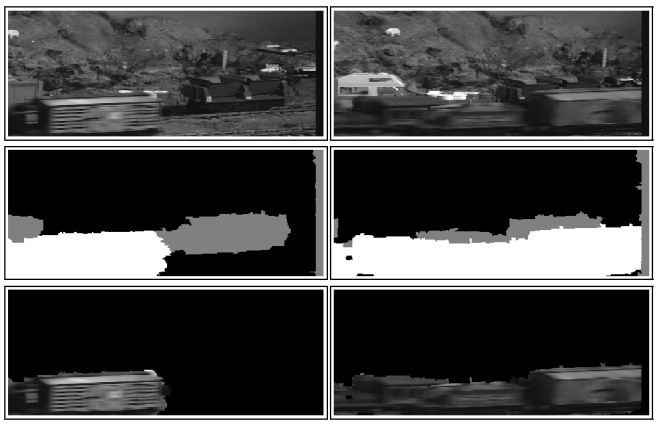


Fig. 7. Original frames 10 and 26 of the "train" sequence, the corresponding label masks, and the mask for the train in the foreground. ($z_c = 4.5, z_t = 2$).

Finally, we have applied our method assuming that only three objects are present in the scene. In Fig. 6a, we present the label field obtained at convergence for the 10th frame of the sequence when both temporal and spatial constraints are disabled. In the absence of temporal and spatial constraints, we could not obtain a good localization of the two ships. In comparison, in Fig. 6b, we present the label field obtained at convergence at frame 11, with both the temporal and spatial constraints enabled.

5.2 "Train" Sequence

The algorithm has been also tested on the even field of the interlaced "train" sequence. The original fields for frames 10 and 26 are presented in the left column of Fig. 7. The movement of the camera is generating an apparent motion of the background of about 4 to 8 pixels per frame (depending on the relative depth), one train is moving with 6 pixels per frame and the other train with about 45 pixels per frame. Due to the large apparent motion, there are large areas that appear and disappear from the scene, namely, the areas in front and in the back of the second train and the areas that border the image. For the same reason, there are even areas that appear only for one frame; for example, the area between the wagons of the train in the foreground ("train two").

In Fig. 7, we present the original frames 10 and 26, the corresponding label fields, and the mask for the train in the foreground. The algorithm exhibits good temporal stability, good localization properties, and the areas that appear and disappear are also classified successfully due to the bidirectional way in which we validate the motion hypotheses. However, problems occur for the areas between the wagons of the train in the foreground, that appear only for one frame. Since there is no correspondence neither in the previous or in the next frame and the temporal constraint is also invalid, they are likely to be misclassified. Rough manual initializations were used for the horizontal motion components of the three different objects. For the 10th frame, the temporal constraints were disabled.

In order to illustrate the internals of the iterative procedures, we have applied the algorithm at the 10th frame of the sequence with a bad initialization for the motion parameters and the temporal constraints disabled. In Fig. 8, we present the label fields obtained

Fig. 8. Label masks at external iterations 0, 3, and 12 with a bad initialization of the motion parameters and disabled temporal constraints.

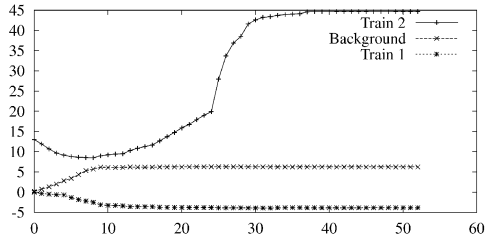


Fig. 9. Evolution of the horizontal translational components of the affine motion model for the 10th frame of the "train" sequence.

at the end of the different labeling phases and, in Fig. 9, we present the horizontal motion components obtained at the subsequent iterations within the motion estimation phases.

Finally, in order to demonstrate the influence of the bidirectional way in which the motion hypotheses are estimated and validated, we present experimental results which were obtained by setting $\{d_s = 1, s \in [1 \dots K]\}$. This way, only the forward direction is considered. In Fig. 10, we present the label field and the corresponding object masks for the 10th frame of the sequence which are to be compared with the results presented in the first column of Fig. 7. Misclassifications occur in the areas that are covered in the next frame (frame 11), namely, the areas in front of the two trains as well as at the right edge of the field of view.

5.3 "Pig" Sequence

In the "pig" sequence, which was obtained for the needs of a project for monitoring animal behavior, a pig is moving against a static background, with slowly changing illumination conditions. There is strong rotation in some of the frames of the sequence and deformations of the body of the pig. Moreover, the assumption of rigid motion is violated in areas like the pig's ears and legs. In Fig. 11 the label fields at frames 411, 416, 421, and 426 are presented. The localization accuracy and the temporal stability are preserved, even though the motion of the pig changes quite fast and in a strong rotational sense. However, the motion of the ears and the leg, in some cases, deviate significantly from the estimated parametric model and merge with the background.

6 Conclusions

In this paper, we have proposed a method for segmentation of video sequences in which spatial and temporal consistency is expressed in terms of interactions between segments that result from an initial intensity segmentation. We express the solution in terms of the MAP criterion and propose an optimization strategy which iteratively maximizes the conditional a posteriori probability of the label field with respect to the motion and the label field.

We have presented results for various image sequences, that show that with the proposed modeling it is possible to group segments that result from a "fine" initial segmentation, based on motion information. The proposed method exhibits good localization properties, temporal stability and deals successfully with motion occlusions.

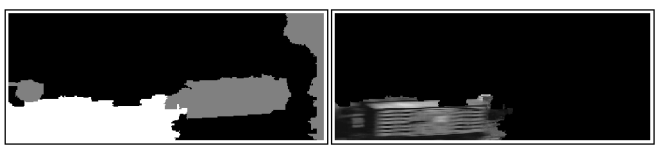


Fig. 10. Label field and object mask for the foreground train for the 10th frame of the "train" sequence with forward motion estimation and labeling ($\{d_s = 1, s \in [1K]\}$). The absence of a correct match in the next frames causes misclassifications in occluded areas.

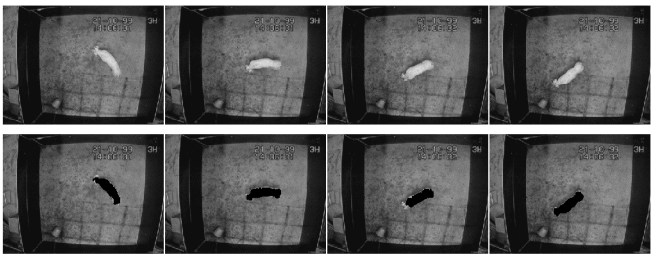


Fig. 11. Original frames and masks at frames 411 to 426 of the "pig" (Courtesy of R. Griekspoor).

For future work, an explicit treatment of the occlusions and, more specifically, of occlusions in the previous frame could be beneficial. This implies the identification of segments that have just appeared in the scene and the relaxation of the assumption of the temporal continuity of the label map in such cases. Finally, the automatic determination of the number of the objects using the Minimum Description Length principle [28], [29] might be an interesting extension.

Acknowledgments

The video material for the "pig" sequence was provided by the Deptartment of Animal Health and Welfare, Danish Institute of Agricultural Sciences. The results for the same sequence were obtained at Noldus Information Technology b.v., Wageningen, The Netherlands.

References

[1] M.J. Black and P. Anandan, "The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow-Fields," *Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 75-104, Jan. 1996.
[2] H.H. Nagel, "On the Estimation of Optical Flow: Relations between Different Approaches and Some New Results," *Artificial Intelligence*, vol. 33, no. 3, pp. 299-324, Nov. 1987.
[3] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no. 11, pp. 721-741, Nov. 1984.
[4] J. Konrad and E. Dubois, "Bayesian Estimation of Motion Vector Fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 9, pp. 910-927, Sept. 1992.
[5] C. Stiller, "Object-Based Estimation of Dense Motion Fields," *IEEE Trans. Image Processing*, vol. 6, no. 2, pp. 234-250, Feb. 1997.
[6] M.M. Chang, A.M. Tekalp, and M.I. Sezan, "An Algorithm for Simultaneous Motion Estimation and Scene Segmentation," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, Apr. 1994.
[7] P. Bouthemy and E. Francois, "Motion Segmentation and Qualitative Dynamic Scene Analysis from an Image Sequence," *Int'l J. Computer Vision*, vol. 10, no. 2, pp. 157-182, May 1993.
[8] I. Patras, E.A. Hendriks, and R.L. Lagendijk, "Video Segmentation by Map Labeling of Watershed Segments," Techical Report ICT-00-01, Delft Univ. of Technology, 2000.
[9] R. Fablet, P. Bouthemy, and M. Gelgon, "Moving Object Detection in Color Image Sequences Using Region-Level Graph Labeling," *Proc. IEEE Int'l Conf. Image Processing*, Oct. 1999.
[10] N. Diehl, "Object-Oriented Motion Estimation and Segmentation in Image Sequences," *Signal Processing: Image Comm.*, vol. 3, no. 1, pp. 23-56, Jan. 1991.
[11] F. Dufaux, F. Moscheni, and A. Lippman, "Spatiotemporal Segmentation Based on Motion and Static Ssegmentation," *Proc. IEEE Int'l Conf. Image Processing*, vol. 1, pp. 306-309, Oct. 1995.
[12] F. Moscheni, F. DuFaux, and M. Kunt, "A New Two-Stage Global/Local Motion Estimation Based on a Background/Foreground Segmentation," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, May 1995.
[13] J.Y.A. Wang and E.H. Adelson, "Representing Moving Images with Layers," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 625-638, Sept. 1994.
[14] F. Moscheni, S. Bhattacharjee, and M. Kunt, "Spatiotemporal Segmentation Based on Region Merging," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 9, pp. 897-915, Sept. 1998.

[15] D. Wang, "Unsupervised Video Segmentation Based on Watersheds and Temporal Tracking," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 539-546, Sept. 1995.
[16] N. Brady and N.E. O'Connor, "Object Detection and Tracking Using an Em-Based Motion Estimation and Segmentation Framework," *Proc. IEEE Int'l Conf. Image Processing*, p. 17A2, 1996.
[17] Y. Weiss and E.H. Adelson, "A Unified Mixture Framework for Motion Segmentation: Incorporating Spatial Coherence and Estimating the Number of Models," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 321-326, 1996.
[18] J. Konrad and V.N. Dang, "Coding-Oriented Video Segmentation Inspired by MRF Models," *Proc. Int'l Conf. Image Processing*, vol. 1, pp. 909-912, 1996.
[19] M. Gelgon and P. Bouthemy, "A Region-Level Graph Labeling Approach to Motion-Based Segmentation," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 514-519, 1997.
[20] M. Gelgon and P. Bouthemy, "A Region-Level Motion-Based Graph Representation and Labeling for Tracking a Spatial Image Partition," *Pattern Recognition*, vol. 33, pp. 725-740, Apr. 2000.
[21] S. Beucher, "Watersheds of Functions and Picture Segmentation," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 1928-1931, May 1982.
[22] P. Shalembier and J. Serra, "Morphological Multiscale Image Segmentation," *Proc. SPIE Visual Comm. and Image Processing*, vol. 1818, pp. 620-631, Nov. 1992.
[23] L. Vincent and P. Soille, "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583-589, June 1991.
[24] S.Z. Li, *Markov Random Field Modeling in Computer Vision*. Springer-Verlag, 1995.
[25] E. Dubois and J. Konrad, "Estimation of 2D Motion Fields from Image Sequences with Application to Video Coding," *Motion Analysis and Image Sequence Processing*, M.I. Sezan and R.L. Lagendijk, eds., Kluwer Academic Publishers, pp. 53-87, 1993.
[26] J. Besag, "On the Statistical Analysis of Dirty Pictures," *J. Royal Statistical Soc.*, vol. 48, no. 3, pp. 259-302, 1986.
[27] J.M. Odobez and P. Bouthemy, "Robust Multiresolution Estimation of Parametric Motion Models," *J. Visual Comm. and Image Representation*, vol. 6, no. 4, pp. 348-365, Dec. 1995.
[28] J Rissanen, "A Universal Prior for Integers and Estimation by Minimum Description Length," *The Annals of Statistics*, vol. 11, no. 2, pp. 416-431, 1983.
[29] C.S. Wallace and D.M. Boulton, "An Information Measure for Classification," *Computing J.*, vol. 11, no. 2, pp. 185-195, 1968.