UNSUPERVISED MOTION DETECTION USING A MARKOVIAN TEMPORAL MODEL WITH GLOBAL SPATIAL CONSTRAINTS

P-M. Jodoin [‡] *M. Mignotte*[‡]

[‡] Département d'Informatique et de Recherche Opérationnelle (DIRO), Université de Montréal,
P.O. Box 6128, Stn. Centre-Ville, Montréal, Québec, H3C 3J7.
E-MAIL : JODOINP@IRO.UMONTREAL.CA

ABSTRACT

In this work, we propose an unsupervised Bayesian model for the detection of moving objects from dynamic scenes. This unsupervised solution is a three-step approach that uses a statistical model of an inter-frame gradient norm field (as likelihood model) with a local regularization term (as prior model) combined with strong intra-frame spatial constraints. In the first step, the spatial constraints are estimated by making an unsupervised Markovian spatial over-segmentation of two input frames. In the second step, the inter-frame gradient (derived from the input frames) is restored to minimize undesired noise. In the last step, an unsupervised Markovian temporal segmentation (with global spatial constraints) is performed to generate the desired motion label field. The Maximum A Posteriori (MAP) estimation of the label field associated with the spatial segmentations (in the first step) and the motion label field (in the third step) is performed by a classical Iterative Conditional Mode (ICM) algorithm. An Iterative Conditional Estimation (ICE) procedure is exploited for estimating the parameters of the spatial model and the region-constrained temporal model. This new statistical method of motion detection has been successfully applied to real dynamic scenes and seems to be well suited for the temporal detection of noisy image sequences.

1. INTRODUCTION

Motion segmentation plays an important role in image sequence analysis and thus has received a lot of attention in the past twenty years [1]. Motion segmentation refers to the general task of labeling image regions that contain uniform displacement vectors. Motion detection is a special case of motion segmentation since it aims at partitioning the image into spatially homogeneous regions that are either moving or stationary. In the literature [1, 2], the label field associated with this motion detection map is often called the *Change Detection Mask*(CDM).

Motion detection methods can be divided in two groups, namely the *motion-based* approaches versus the *spatio-temporal* techniques. Motion-based approaches segment image sequences based on temporal information only such as *optical flow* [3, 4, 5, 6] or *inter-frame difference* [7, 8, 2, 9, 10]. Some well known limitations of motion-based detection techniques come from their sensitivity to temporal noise and their difficulty to accurately preserve discontinuities at object boundaries. To overcome these limitations, spatio-temporal segmentation techniques make use of intra-frame spatial information to rectify and improve the temporal segmentation results. Consequently, such techniques are generally more robust but slower due to the extra computational effort required. In this context, two types of spatial constraints are often proposed, namely *contour-based* [2, 11] or *region-based* [12].

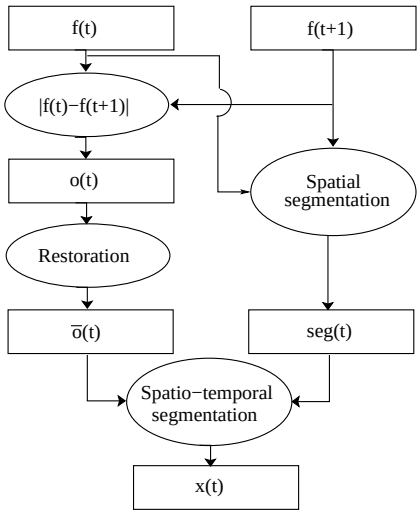

FIG. 1 – Pipeline used by our motion detection method

In this paper, we propose a new statistical spatio-temporal approach built over an inter-frame difference feature space with region-based spatial constraints. Our method differs from the others in that the proposed Bayesian strategy makes use of a restored version of the inter-frame gradient norm field and has strong intra-frame global spatial constraints.

Our approach is a three-step procedure. In the first step, the two input time frames f_f and f_{t+1} are partitioned (with an unsupervised spatial Markovian segmentation) into a set of uniform regions $\hat{R}$. The purpose of this step is to decompose every moving object into small uniform region that are likely to be either moving of stationary. The second step is a restoration procedure that minimizes undesired noise within the inter-frame gradient image o_t. This step adds robustness to the temporal segmentation procedure. The third and final step is a spatio-temporal Markovian segmentation that partitions the denoised inter-frame gradient $\overline{o}_t$ into three classes. This segmentation uses global spatial constraints (taken from $\hat{R}$) to overcome the problem of noise sensitivity and boundary inaccuracy. In other words, this step is a region-based segmentation that assigns motion labels to regions ($\hat{R}$) instead of pixels(see Fig.1 for a schematic view of our three-step motion detection method).

The search for the label field associated with the spatial segmentation (first step) and with the motion label field (third step) is performed by a classical markovian procedure.

The remainder of this paper is organized as follows. In Section 2, we provide a brief overview of the strategy used for both Markovian segmentations while the three-step spatio-temporal motion detection procedure is discussed in Section 3. In Section 4, experimental results are presented and a brief conclusion is provided in section 5.

2. UNSUPERVISED BAYESIAN SEGMENTATION

The motion detection method presented in this work performs two Markovian segmentations in order to obtain the CDM. The first one is an *M*-class *spatial* segmentation applied to the two input images, i.e f_t (frame at time t) and f_{t+1} (frame at time $t+1$). The second one is a three-class *temporal* segmentation with spatial constraints applied to the denoised inter-frame gradient image $\bar{o}_t$derived from f_t and f_{t+1}. In both cases, the segmentation model and the parameter estimation procedure use the same strategy. Only the likelihood model is different and will be defined in Section 3.

2.1. Markovian Segmentation

Let $Z = \{X, Y\}$ a pair of random fields where $X = \{x_s, s \in S\}$ and $Y = \{y_s, s \in S\}$, represent respectively the label field (related to the spatial or temporal segmented image) and observation field (associated with f_t or f_{t+1} in the first step or associated with $\overline{o}_t$ in the third step), both defined on $S = \{s = (i,j)\}$, a 2D lattice of N sites. Each y_s takes a value in $\{0, ..., 255\}$ and x_s takes a value in $\{1, \ldots, m\}$, where m corresponds to the number of classes of the segmentation map. The distribution of $\{X, Y\}$ is defined by the Markovian prior distribution $P(X)$ combined with the conditional data likelihood $P(Y/X)$ depending on a parameter vector Φ.

In this framework, the segmentation problem can be viewed as a statistical labeling problem according to a global Bayesian formulation in which the posterior distribution $P(X/Y) \propto \exp -U(X,Y)$ has to be maximized [13]. By assuming independence between each random variable Y_s given X_s (i.e., $P(Y/X) = \prod_{s \in S} P(Y_s/X_s)$), and an isotropic spatial Potts model with a second-order neighborhood for both models, the corresponding posterior energy to be minimized is

$$U(X,Y) = \underbrace{\sum_{s \in S} \Psi_s(x_s, y_s)}_{U_{\text{data}}} + \underbrace{\sum_{<s,t>} \beta\big[1 - 2\,\delta(x_s, x_t)\big]}_{U_{\text{smooth}}}, \quad (1)$$

where δ is the Kronecker function, β is a constant and $\Psi_s(x_s, y_s) = -\ln P(y_s/x_s)$. The conditional distribution $P(y_s/x_s)$ of each class x_s is modeled by a Normal law depending on a set of parameters Φ. To perform an unsupervised segmentation of Y, $\Phi = [(\mu_1, \sigma_1), \ldots, (\mu_m, \sigma_m)]$ has to be estimated. To this end, we resort to an iterative method called Iterated Conditional Estimation (ICE) [14].

2.2. Mixture Parameter Estimation

Assuming that the "complete data" $Z = \{X, Y\}$ is known, the parameters of the gaussian mixture can be computed with the Maximum Likelihood (ML) estimator on each class $\{1, \ldots, m\}$. The random field X being unobservable, the ICE procedure defines $\Phi^{[p+1]}$ at iteration $p+1$ as the conditional expectation of $\hat{\Phi}$ given $Y = y$, computed according to the current value $\Phi^{[p]}$. This gives the best approximation in terms of the mean square error [14]. Denoting E_p the expectation relative to parameter vector $\Phi^{[p]}$, $\Phi^{[p+1]}$ is computed from $\Phi^{[p]}$ and $Y = y$, as a fixed point, by $E_p[\hat{\Phi}(X,Y)|Y = y]$. This expectation is impossible to compute in practice, but it can be approximated in the following way, thanks to the law of large numbers :

$$E_p[\hat{\Phi}(X,Y)|Y = y] \approx \frac{1}{n}\big(\hat{\Phi}^{[p]}(x_{(1)}, y) + \ldots + \hat{\Phi}^{[p]}(x_{(n)}, y)\big),$$

where $x_{(i)}, i = 1, \ldots, n$ are realizations of X drawn from the *posterior* distribution $P(X/Y, \hat{\Phi}^{[p]})$. As it turns out, $n = 1$ is sometimes found sufficient to get good estimates when convergence is reached [14]. It is also the case in our unsupervised segmentation models and we actually choose $n = 1$ in our experiments. A good initialization is important and has a significant impact on the speed of convergence of this iterative procedure. In our application, we use a K-mean clustering algorithm to roughly obtain an m-class partition of the input image. ML estimators on these partitions then allow $\Phi^{[0]}$ to be obtained.

3. THREE-STEP MOTION DETECTION PROCEDURE

3.1. Step One : Markovian Spatial Segmentation

This first step intends to efficiently detect, from two successive frames, small uniform regions that are likely to be either moving of stationary. This is done by defining a partition $\hat{\mathcal{R}} \triangleq \{r_n, n = 1, \ldots, N_{r_{\max}}\}$ of f_t and f_{t+1} into a set of disjoint and uniform regions. These regions will be exploited as spatial constraints with the Markovian temporal detection model (in step three).

To calculate $\hat{\mathcal{R}}$, we over-segment f_t and f_{t+1} by using a Gaussian law as degradation model to describe the gray-level luminance within each spatial region $r \in \hat{\mathcal{R}}$. In this context, x is the spatial label field and the data likelihood energy term $U_{\text{data}}(.)$ of Eq. (1) can be written as

$$\Psi_s\big(x_s, g(s)\big) = \ln(\sqrt{2\pi}\,\sigma_{x_s}) + \frac{(g(s) - \mu_{x_s})^2}{2\sigma_{x_s}^2}, \quad (2)$$

where $g(s)$ is either $f_t(s)$ or $f_{t+1}(s)$ and $\hat{\Phi} = \{(\mu_{x_s}, \sigma_{x_s}), x_s \in \{1, \ldots, m\}\}$ is estimated by the ICE procedure on f_t (see Section 2.2). Segmentation results of f_t and f_{t+1} (i.e., label field images f_t^{seg} and f_{t+1}^{seg}) are then combined together by the following linear operation

$$I_t^{\text{seg}}(s) = f_t^{\text{seg}}(s) + m f_{t+1}^{\text{seg}}(s)$$

such that I_t^{seg} contains the desired set of uniform disjoint regions $\hat{\mathcal{R}}$.

3.2. Step Two : Restoration of the Temporal Gradient Field

The inter-frame gradient image $o_t = |f_{t+1} - f_t|$ is often very noisy and induces unacceptable errors when used directly as an observation field in a temporal segmentation model. In order to reduce the effect of inter-frame noise, o_t is restored with an edge-preserving denoising procedure. To this end, we have implemented a filtering algorithm called *mean shift*. Mean shift is a simple iterative nonparametric estimator of density gradient that was first

introduced by Fukunaga and Hosteler [15] and adapted to image filtering by Comaniciu and Meer [16]. As shown in Fig. 2b, mean shift efficiently reduces undesired noise while preserving edges well.

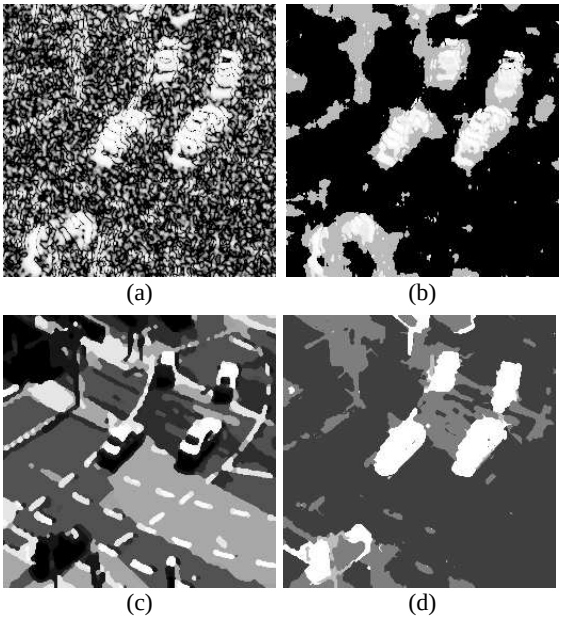

FIG. 2 – *(a) Noisy gradient image* $o_t = |f_t - f_{t+1}|$ *taken from the* Karlsruhe *sequence. (b)* o_t *after a mean-shift restoration. (c)* 8*-class spatial segmentation* f_t^{seg}*. (d)* $\hat{o}(t)$*, the filtered gradient field.*

3.3. Step Three : Region-Constrained Temporal Segmentation Model

This last step segments $\overline{o}_t$ into three classes in order to seperate regions with low gradient norm from the ones with medium and high gradient norm. In this context, the set of disjoint regions $\hat{\mathcal{R}}$ is exploited for two reasons. The first reason is to provide a hierarchical (spatial) filtered version of $\overline{o}_t$ by averaging its values over each region $r \in \hat{\mathcal{R}}$,

$$\forall\, r \in \hat{\mathcal{R}} \quad \hat{o}_t(s) = \frac{1}{N_r} \sum_{\forall s \in r} \overline{o}_t(s), \quad \forall\, s \in r, \tag{3}$$

where N_r is the number of sites within region $r \in \hat{\mathcal{R}}$. As shown in Fig. 2d, the filtered gradient field $\hat{o}_t$ better represents the moving scene than o_t or $\overline{o}_t$ since it contains more precise boundaries around moving regions.

The second reason to exploit $\hat{\mathcal{R}}$ is to globally constrain the temporal segmentation model. In this perspective, we associate the same motion label C_r^{st} with each site s belonging to a detected region $r \in \hat{\mathcal{R}}$ at time t. This helps preserve the integrity of the different object shapes present in the scene. In this context, x is the temporal label field and the data likelihood energy term $U_{\text{data}}(.)$ of Eq. (1) can be written as

$$U_{\text{data}}(x, \hat{o}_t) = \sum_{r \in \hat{\mathcal{R}}} \sum_{s \in r} \ln(\sqrt{2\pi}\,\sigma_{x_s}) + \frac{(\hat{o}_t(s) - \mu_{x_s})^2}{2\sigma_{x_s}^2}, \tag{4}$$

with the constraint that $\forall r \in \hat{\mathcal{R}}, \forall s \in r, x_s = C_r^{\text{st}}$. In other words, this segmentation statistically assigns labels to regions (computed in step 1) instead of pixels. Once again, the parameters $\hat{\Phi} = \{(\mu_{x_s}, \sigma_{x_s}), x_s \in \{1, \ldots, m\}\}$ is estimated with the ICE procedure on $\overline{o}_t(s)$. The resulting label field $\hat{x}_t$ of the segmentation is the CDM we were looking for.

4. EXPERIMENTAL RESULTS

We have tested our detection method on several image sequences. The spatial segmentation map (defined in Subsection 3.1) and the CDM $\hat{x}_t$ (defined in Subsection 3.3) are both inferred by the Iterative Conditional Mode (ICM) algorithm [13]. Experiments have shown that ICM provides results as good as a stochastic estimation procedure such as simulated annealing [17], mainly because of the restoration procedure (in second step) and the global spatial constraints exploited in the temporal detection procedure (in third step). For every sequence, the spatial segmentation was made with 12 classes and β was set to 1 for every segmentation.

We have compared our results to the closest Bayesian method proposed in the literature (to our knowledge), namely the method developed by Paragios and Tziritas in [10]. This method works over an inter-frame difference feature space but has no restoration procedure and no global spatial constraint. Their Bayesian strategy includes a spatial and a temporal regularization energy term. More specifically, their method seeks to minimize a global energy function of the form

$$\begin{aligned} U(x_t, x_{t-1}, o_t) \quad = \quad & U_{\text{data}}(x_t, o_t) + U_{\text{smooth}}(x_t) \\ & + U(x_t, x_{t-1}) \end{aligned}$$

where $U_{\text{smooth}}(x_t)$ is the prior energy term, $U_{\text{data}}(x_t, o_t)$ is the likelihood energy term and $U(x_t, x_{t-1})$ is the regularization term that expresses a temporal coherence with respect to the label field at the preceding time $t-1$. However, since in our framework only two time frames are available, x_{t-1} is unknown. For this reason, $U(x_t, x_{t-1})$ was set to zero and the other two terms were kept as is. We used a simulated annealing algorithm [17] to minimize this energy function.

Our method converged in an average of 10 seconds on a 2 GHz Pentium IV with 512 MB of memory. Note that our program could be further optimized to reduce processing time. Results over the sequences *Man moving*, *Trevor white*, *Mom and daughter* and *Karlsruhe* are presented in Figs. 3 and 4. Notice that our method works well for noisy image sequences such as *Karlsruhe* and *Man moving* as well as for the others. It preserves the different boundary discontinuities well (i.e., the integrity of the different object shapes present in the dynamic scene) without leaving holes within the moving objets.

5. CONCLUSION

In this work, we have presented a novel three-step unsupervised Bayesian solution to the problem of motion detection between two time frames. Our approach estimates the motion label field by minimizing a global temporal energy function involving a restoration process and strong intra-frame global spatial constraints. Experimental results reported here seem very promising. This strategy appears to be robust with noisy image sequences and preserves well the integrity of the objects' boundaries present in the

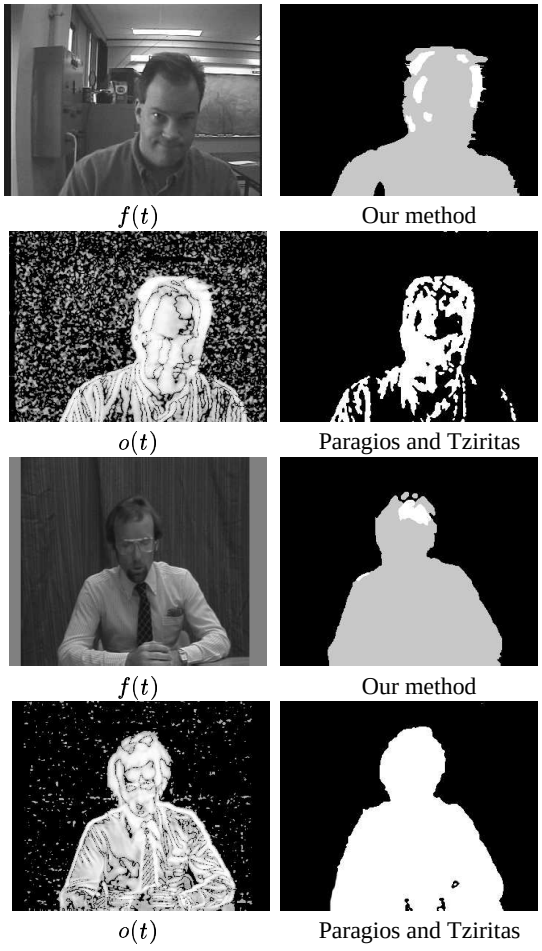

FIG. 3 – *Results taken from sequences* Man moving *and* Trevor white*. Bottom right results were obtained with a slightly modified version of Paragios and Tziritas's Bayesian method [10].*

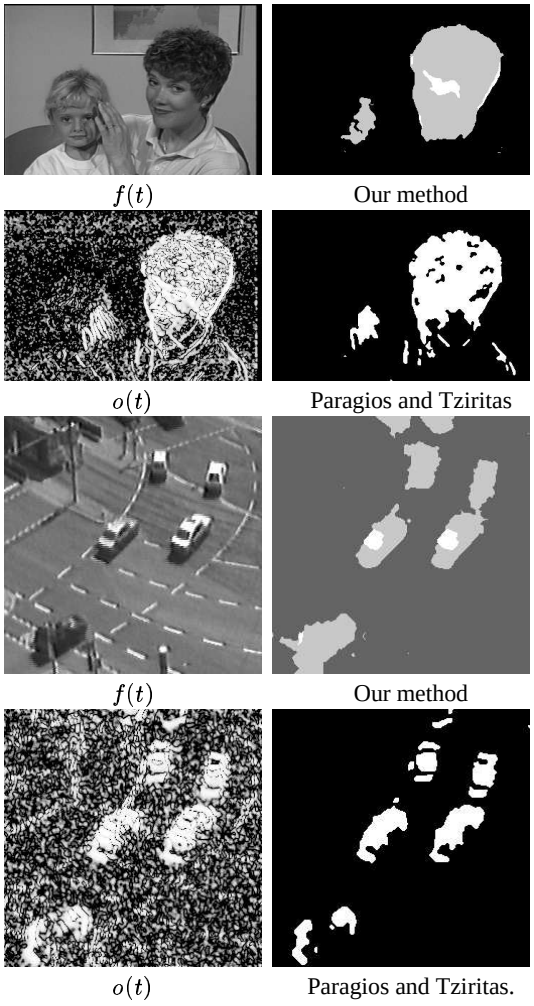

FIG. 4 – *Results from sequences* Mom and daughter *and* Karlsruhe.

scene. As future work, we aim to adapt our method to object tracking and motion localization over a sequence of more than two time frames.

6. REFERENCES

[1] Zhang D. and Lu G. Segmentation of moving objects in image sequence : A review. *Circuits, Systems and Signal Processing*, 20(2) :143–183, 2001.

[2] Mech R.and Wollborn M. A noise robust method for 2d shape estimation of moving objects in video sequences considering a moving camera. *Signal Processing*, 66(2) :203–217, 1998.

[3] Odobez J. and Bouthemy P. Separation of moving regions from background in an image sequence acquired with a mobile camera. In H.H. Li, S. Sun, and H. Derin, editors, *Video Data Compression for Multimedia Computing*, chapter 8, pages 283–311. 1997.

[4] Vasconcelos N. and Lippman A. Empirical bayesian motion segmentation. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(2) :217–221, 2001.

[5] Murray D. and Buxton B. Scene segmentation from visual motion using global optimization. *IEEE Trans. Pattern Anal. Machine Intell.*, 9(2) :220–228, 1987.

[6] Chang M., Tekalp M., and Sezan I. Simultaneous motion estimation and segmentation. *IEEE Trans. on Image Process.*, 6(9) :1326–1333, 1997.

[7] Jain R.and Nagel H. On the analysis of accumulative difference pictures from image sequences of real world scenes. *IEEE Trans. Pattern Anal. Machine Intell.*, 1(2) :206–214, 1979.

[8] Aach T., Kaup A., and Mester R. Statistical model-based change detection in moving video. *Signal Processing*, 31(2) :165–180, 1993.

[9] Neri A., Colonnese S., Russo G., and Talone P. Automatic moving object and background separation. *Signal Processing*, 66(2) :219–232, April 1998.

[10] Paragios N. and Tziritas G. Adaptive detection and localization of moving objects in image sequences. *Signal Processing : Image Communication*, 4 :277–296, 1999.

[11] Meier M. *Segmentation for Video Object Plane Extraction and Reduction of Coding Artifacts*. PhD thesis, University of Western Australia, 1998.

[12] Choi J., Lee S, and Kim S. Spatio-temporal video segmentation using a joint similarity measure. *IEEE Trans. Circ. and Sys. For Vid. Tech.*, 7(2) :279–286, 1997.

[13] Besag J. On the statistical analysis of dirty pictures. *J. Roy. Stat. Soc.*, 48(3) :259–302, 1986.

[14] Pieczynski W. Statistical image segmentation. *Machine Graphics and Vision*, 1(1) :261–268, 1992.

[15] Fukunaga K. and Hostetler L. The estimation of the gradient of a density function. *IEEE Trans. on Info. Theory*, 21 :32–40, 1975.

[16] Comaniciu D. and Meer P. Mean shift : A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(5) :603–619, 2002.

[17] Geman S. and Geman D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6(6) :721–741, 1984.