Simultaneous Motion Detection and Background Reconstruction with a Mixed-State Conditional Markov Random Field

Tomás Crivelli[1,2], Gwenaelle Piriou[3], Patrick Bouthemy[2],
Bruno Cernuschi-Frías[1,2], and Jian-feng Yao[2,4]

[1] University of Buenos Aires, Buenos Aires, Argentina
[2] INRIA Rennes, Irisa, France
[3] Université de Bretagne-Sud, Vannes, France
[4] IRMAR, Rennes, France
tcrivell@irisa.fr

Abstract. We consider the problem of motion detection by background subtraction. An accurate estimation of the background is only possible if we locate the moving objects; meanwhile, a correct motion detection is achieved if we have a good available background model. This work proposes a new direction in the way such problems are considered. The main idea is to formulate this class of problem as a joint decision-estimation unique step. The goal is to exploit the way two processes interact, even if they are of a dissimilar nature (symbolic-continuous), by means of a recently introduced framework called mixed-state Markov random fields. In this paper, we will describe the theory behind such a novel statistical framework, that subsequently will allows us to formulate the specific joint problem of motion detection and background reconstruction. Experiments on real sequences and comparisons with existing methods will give a significant support to our approach. Further implications for video sequence inpainting will be also discussed.

1 Introduction

The recent advances in computer vision have been moving towards the quest for the development of systems and algorithms able to tackle complex situations where an integrated and optimal decision-estimation process is required. Efficient early vision techniques are nowadays able to feed subsequent stages of high-level information processing in a desirable way: fast, accurate and robust. Anyway, there has been always a component of sequentiality that tends to address a certain task as a succession of atomic steps. Consider the problem of foreground moving objects detection by background subtraction, where a model of the background or reference image is usually learned and motion detection is solved from differences between image observations and such a model. A "chicken-and-egg" situation arises when we want to set an optimal approach for both tasks: an accurate estimation of the background is only possible if we know which regions of the image belong to it, that is, if we locate the moving objects; meanwhile,

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part I, LNCS 5302, pp. 113–126, 2008.

a correct motion detection is achieved if we have a good available background model.

This work proposes a new direction in the way such problems are considered. The main idea is to formulate a unique and joint decision-estimation step, which means more than simply solving two (or more) problems at the same time (either sequentially, iteratively or adaptively). We emphasize that the goal is to exploit the way two processes interact, even if they are of a dissimilar nature.

Returning to the problem of motion detection and background modeling, we can redefine the problem as an example of the starting point for our proposal: let us consider that a point in the image is a single process that can take two types of values, a symbolic value (or abstract label) accounting for a positive motion detection, or a numeric value associated to the brightness intensity of the reference image at that location. Consequently, what would it mean to solve both tasks jointly in this context, is to obtain a single optimal estimate of such a process (Fig. 1).

Additionally, our method relies not only on the comparison between the current image and the reference image but explicitly integrates motion measurements obtained between consecutive images. Conditional random fields [1], extended to a mixed-state version, allow us to introduce these observations (or any other) in the model and contributes to make the overall scheme complete, accurate and powerful.

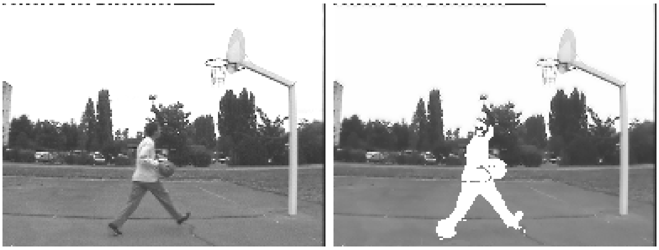


Fig. 1. Left: original image from the Basketball sequence. Right: a mixed-state field obtained with the proposed motion detection method. In white it is represented the symbolic part, accounting for a positively detected moving point. The continuous part is represented by the reconstructed background

This paper is organized as follows. In section 2, we will sufficiently describe the theory behind such a novel statistical framework. In section 3, a review of motion detection by background subtraction techniques with their advantages and drawbacks are discussed. Section 4 is devoted to the formulation of the proposed conditional mixed-state model for simultaneous motion detection and background reconstruction. In section 5 we show experiments on real sequences and comparisons with existing methods, which will give a significant support to our approach. These comparisons will show an improvement in the detection rate, diminishing the number of false positives and negatives, together with a correct reconstruction of the real background image, not a model of it. Further implications and results of this method for video sequence inpainting will be also discussed.

2 The Mixed-State Statistical Framework and Related Approaches

The concept of a random process that can take different types of values (either numerical or abstract) according to a generalized probability function, is formalized through the so-called *mixed-state random variables*. This includes diverse situations. We have formulated a mixed discrete-continuous Markov random field in [2] in the context of modeling of dynamic or motion textures. In this case, it is demonstrated that normal flow scalar motion observations arised from these types of sequences show a discretely distributed value at zero (null-motion) and a Gaussian-like continuous distribution for the rest of the values. This model was extended in [3] and applied to the problem of motion texture segmentation. Previously, Salzenstein and Pieczynski [4] (more recently in [5]), have proposed a fuzzy image segmentation model where the fuzzy labels are a particular instance of mixed-state variables with values in $[0, 1]$.

Our work takes a step further, not in the theoretical aspect of the framework, but in the exploitation of its implications in computer vision.

Let us define $\mathcal{M} = \{\omega\} \cup \mathbb{R}$, with ω a "discrete" value, called *symbolic* value. A random variable X defined on this space, called *mixed-state variable*, is constructed as follows: with probability $\rho \in (0, 1)$, set $X = \omega$, and with probability $1 - \rho$, X is continuously distributed in $\mathbb{R}$. In order to compute the probability density function of X, $\mathcal{M}$ is equipped with a "mixed" reference measure $m(dx) = \nu_\omega(dx) + \lambda(dx)$, where ν_ω is the discrete measure for the value ω and λ the Lebesgue measure on $\mathbb{R}$. Let us define the indicator function of the symbolic value ω as $\mathbf{1}_\omega(x)$ and its complementary function $\mathbf{1}^*_\omega(x) = \mathbf{1}_{\{\omega\}^c}(x) = 1 - \mathbf{1}_\omega(x)$. Then, the above random variable X has the following density function w.r.t. $m(dx)$:

$$p(x) = \rho \mathbf{1}_\omega(x) + (1 - \rho)\mathbf{1}^*_\omega(x) p^c(x), \tag{1}$$

where $p^c(x)$ is a continuous pdf w.r.t. λ, defined on $\mathbb{R}$. Hereafter, such generalized density will be called *mixed-state density*.

2.1 Mixed-State Markov Models

In the context of Markov fields, the concept of mixed-state random variables and mixed-state densities, derives in the definition of mixed-state conditional densities. Let $S = \{1....N\}$ be a lattice of points or image locations such that $\mathbf{X} = \{x_i\}_{i \in S}$. Define $\mathbf{X}_A$ as the subset of random variables restricted to $A \subset S$, i.e., $\mathbf{X}_A = \{x_i\}_{i \in A}$. Then we write:

$$p(x_i \mid \mathbf{X}_{\mathcal{N}_i}) = \rho(\mathbf{X}_{\mathcal{N}_i})\mathbf{1}_\omega(x_i) + (1 - \rho(\mathbf{X}_{\mathcal{N}_i}))\mathbf{1}^*_\omega(x_i) p^c(x_i \mid \mathbf{X}_{\mathcal{N}_i}), \tag{2}$$

where $\rho(\mathbf{X}_{\mathcal{N}_i}) = P(x_i = \omega \mid \mathbf{X}_{\mathcal{N}_i})$ and $\mathbf{X}_{\mathcal{N}_i}$ is the subset of $\mathbf{X}$ restricted to a neighborhood of locations $\mathcal{N}_i$. Equation (2) defines the local characteristics of a global random field, that will respond to a nearest neighbor Gibbs distribution, as stated by the equivalence of Hammersley-Clifford [6]. Moreover, it is the starting point for defining a model that allows to obtain a regularized symbolic-continuous field. We recall one useful result of the proposed statistical framework (see [6,7]) for the case of second order Markov random fields:

Result 1. *For a second order Markov random field that responds to a family of conditional densities given by (2) the associated joint Gibbs distribution* $Z^{-1}\exp Q(\mathbf{X}) = Z^{-1}\exp -H(\mathbf{X})$ *is given by the energy:*

$$H(\mathbf{X}) = \sum_{i\in S}\left\{V_i^d(x_i) + V_i^c(x_i)\right\} + \sum_{<i,j>\in S}\left\{V_{i,j}^d(x_i,x_j) + V_{i,j}^c(x_i,x_j)\right\}, \qquad (3)$$

where $V_i^d(x_i) = \alpha_i \mathbf{1}_\omega^*(x_i)$ *and* $V_{i,j}^d(x_i,x_j) = \beta_{ij}\mathbf{1}_\omega^*(x_i)\mathbf{1}_\omega^*(x_j)$*, that is they correspond to purely discrete potentials, and* $V_i^c(x_i)$ *and* $V_{i,j}^c(x_i,x_j)$ *are energy terms related to the continuous part* p^c*.*

Thus, we know the general shape of the potentials for a mixed-state model. In what follows, we apply this result to the formulation of the joint problem of motion detection and background reconstruction as a conditional mixed-state Markov random field estimation problem.

3 Motion Detection by Background Subtraction: Overview of Existing Methods

One of the most widely used methods for motion detection is *background subtraction*. The approach, derived initially from a thresholding process over the difference between the observed intensity (or color) at a point and a reference value representing the background, has evolved into more complex schemes where the shared idea is to consider that a foreground moving object does not respond to some representation of the background.

For existing methods, a necessary step consists in the learning of the background and this implies either the availability of training frames with no moving objects, or the assumption that a point belongs to the background most of the time. Adaptive schemes have also been proposed in order to update the model sequentially and selectively, according to the result of the detection step. Anyway, a general consensus has been to estimate a probability density for each background pixel. The simplest approach is to assume a single Gaussian per pixel (see for example [8]), whose parameters may be estimated by simple running averages or even median filters. A valid and certain criticism to this hypothesis is that the distribution of the intensity of a background pixel over time can vary considerably, but usually repeatedly. In that direction, multi-modal density models seemed to perform better. Mixtures of Gaussians [9] and non-parametric models [10,11,12] have shown good results, able to deal with the dynamic of the background distribution.

However they suffer from several drawbacks. The approach does not assume spatial correlation between pixels, nor in the model of the background, neither in the binary detection map. Aware of this, posterior morphological operations are applied in order to achieve some sort of regularization in the resulting motion detection map. No regularization is proposed for the reference model.

They need also to incorporate points detected as foreground to estimate the background model (called blind update) in order to avoid deadlock situations, where a badly estimated background value for a pixel results in a continuously and wrongly detected moving point. This leads to bad detections as intensity values that do not belong to the background are incorporated to the model. A lot of heuristic corrections are usually applied in order to correct this drawback, but unfortunately, introducing others. Finally, they are very sensitive to the initialization of the background model, particularly, when an initial image with no moving objects is not available in the video sequence.

The advantages of incorporating spatial context and regularization, in the background and also the foreground, are demonstrated for example in [13,14] by means of a Markov random field model and ARMA processes, respectively. As for another energy-based method for background subtraction, the work of Sun et al. [15] on Object Cut, models the likelihood of each pixel belonging to foreground or background along with an improved spatial contrast term. The method relies on a known (previously learned) background model and an adaptive update scheme is necessary. Finally in [16], a technique for motion detection, not based on background modeling, but on clustering and segmentation of motion and photometric features, is described, where explicit spatial regularization is introduced through a MAP-MRF approach.

3.1 Our Method

Based on these observations we propose a simultaneous motion detection and background reconstruction method with the following characteristics:

- **Reduction of false positive and false negatives.** Through a more complex regularization of the detection map, exploiting spatial priors, and the interaction between symbolic and continuous states.
- **Reconstruction of the background.** Obtaining a reconstructed reference image, not just a model of it, will allows to exploit the local information of the difference between the background and a foreground moving object, avoiding the undesirable effects of modeling noise, which is filtered out from the reconstructed image.
- **No need of training samples.** Through a temporal update strategy which can be adopted thanks to a correct regularized estimation of the motion map, the reference image is reconstructed on-the-fly on those regions temporally not occluded by the moving objects.
- **Joint decision-estimation solution.** Exploiting simultaneously the information that the reference image provides for motion detection, and vice versa.

4 A Conditional Mixed-State Model for Motion Detection

4.1 Definitions

Let us call $\mathbf{y}_t = \{y_i^t\}_{i \in S}$ the intensity image at time t, where $y_i^t \in [0, 255]$ is the brightness intensity value at location $i \in S = \{1....N\}$ of the image grid. Then $\mathbf{y} = \{\mathbf{y}_t\}_t$ is a sequence of images that we call *observations*. We will associate a positive motion detection for a single point to the abstract label ω. Then, we define a mixed-state random field $\mathbf{x}_t = \{x_i^t\}_{i \in S}$ where $x_i^t \in \mathcal{M} = \{\omega\} \cup [0, 255]$ is a mixed-state random variable.

Suppose we have an estimate of $\mathbf{x}_t$ for a given instant t, that is, the moving points and the estimated intensity value for the background at the non-moving points. We can use this information and the past estimated $\mathbf{x}_{t'}$ (for $t' < t$) to reconstruct the reference image at t, that we call $\mathbf{z}_t = \{z_i^t\}_{i \in S}$. We propose to update the background estimation as follows,

$$z_i^t = \begin{cases} x_i^t & \text{if } x_i^t \neq \omega \\ z_i^{t-1} & \text{otherwise.} \end{cases} \tag{4}$$

The rationale of this rule is that when we do not detect motion, we have a good estimation for the reference intensity value at a given point, so we can use this value as a background value; as the objects in the scene move, we can progressively estimate the background for different parts of the image. In other words, we can fill the gaps at those moments where the background is not occluded.

4.2 Energy Terms

Let us call $H(\mathbf{x}_t \mid \mathbf{y}, \mathbf{z}_{t-1})$ the energy function to be minimized, associated to a *conditional* mixed-state Markov random field as proposed in (3), given the observations $\mathbf{y}$ and the previously available background image [1]. In what follows we design the mixed-state energy terms. Considering a *conditional* Markov random field, as introduced in [1], allows us to define these energy terms in a flexible way, in particular it enables to exploit a large set of observations (e.g., a block) at each site. That is, it is able to integrate at an image location any information extracted from the input data and obtained across arbitrary spatial or temporal (or both) neighborhoods, or information from previously reconstructed variables, or even the association of both.

We will consider three types of energy terms. The *discriminative* term, which plays a role in the decision process, penalizing or favoring the presence of motion for a point given the observations; the *reconstruction* terms, involved in the estimation of the reference image, which also affects the motion detection decision process by means of background subtraction; and the *regularization* terms, related to the smoothing of the mixed-state field.

[1] The extension of the previous stated results to a mixed-state model conditioned to an observation process is straightforward.

First we propose to introduce a discriminative term related to the symbolic part of the field, that is, the motion detection map. Thus, we define a first-order potential

$$V_i^D(x_i^t \mid \mathbf{y}) = \alpha_i^D(\mathbf{y})\mathbf{1}_\omega^*(x_i^t), \tag{5}$$

where the weight $\alpha_i^D(\mathbf{y})$ depends on the observations and aims at tuning the belief of motion for a point. We propose to use $\alpha_i^D(\mathbf{y}) = -\log NFA_i(\mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{y}_{t+1})$, where $NFA_i(\cdot)$ stands for the *Number of False Alarms* obtained through an a contrario decision framework as in [17]. Its value is computed using three consecutive frames and taking the magnitude of the local normal flows, and constitutes a measure of the belief that a point belongs to the background (or conversely, to moving objects). We have implemented the simplest scheme proposed in [17], considering detection over square regions, usually of size 20x20 at each site i. A small value of NFA indicates a large belief of motion and conversely. Consequently, a low value of $\log NFA_i$ favors $x_i^t = \omega$ (Fig. 2). Thus, our method relies not only on the comparison between the current image and the reference image but explicitly introduces motion measurements. The overall scheme gains accuracy and completeness, integrating this low-level feature in the decision process. The flexibility of the conditional random field formulation [1] allows us to exploit these observations within the mixed-state model.

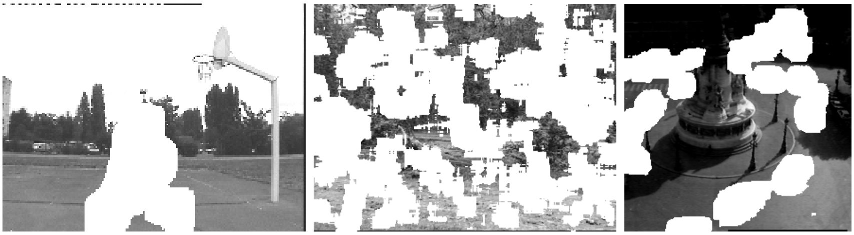


Fig. 2. Initial motion detection by computing the Number of False Alarms. The motion map is obtained by thresholding this quantity as explained in [17]. From left to right: the results are shown for the Basketball sequence (see Fig. 3), the Forest sequence (see Fig. 4) and the Traffic Circle sequence (see Fig. 5). Note that this quantity, with the basic implementation utilized here, over-regularizes the detection map, as it is a block-based detection strategy.

We elaborate now the reconstruction potential. On one side, it aims at estimating the intensity values of the background (reference) image, and exploits the flexibility of the mixed-state model in taking into account their interactions with the symbolic values. On the other side, here is where we introduce the term that compares the current image with the reconstructed reference image, which provides the basis for the decision process in a background subtraction method. We then write

$$V_i^R(x_i^t \mid \mathbf{y}, \mathbf{z}_{t-1}) = \gamma \left[\mathbf{1}_\omega^*(x_i^t) \frac{\left[x_i^t - m(z_i^{t-1}, y_i^t)\right]^2}{\sigma_i^2} + \mathbf{1}_\omega(x_i^t)\alpha_i^R(\mathbf{y}_t, \mathbf{z}_{t-1}) \right]. \tag{6}$$

First, we set $m(z_i^{t-1}, y_i^t) = cz_i^{t-1} + (1-c)y_i^t$ if we have an available previously estimated value for the reference image at that point, or $m(z_i^{t-1}, y_i^t) = y_i^t$ otherwise. Thus, the first term favors that, when there is no motion, i.e. $\mathbf{1}_\omega^*(x_i^t) = 1$, the estimated intensity value for a point is close to the previous estimated reference image, and simultaneously, penalizes the absence of motion if this difference is eventually large. Both types of values interact consequently, in order to minimize the energy. Note that this term also performs a temporal regularization of the reference estimates z_i^t by the interpolation form of the $m(\cdot)$ function. Furthermore, it is normalized by a local variance σ_i^2 estimated locally from $\mathbf{y}_t$. In the second term, we set,

$$\alpha_i^R(\mathbf{y}_t, \mathbf{z}_{t-1}) = \sigma_i^2 \left[n^{-1} \sum_{j \in \mathcal{N}_i} (z_j^{t-1} - y_j^t) \right]^{-2}, \tag{7}$$

resulting in a penalization of the presence of motion when the difference of intensity between the observation and the reference image is small. A local average of differences is introduced in order to reduce the effect of the noise present in the observations.

The potentials introduced so far, are in fact first-order terms, that relate the random variable at a point i w.r.t. the observations. Next, we introduce terms related to the regularization of the field. The objective is to have connected regions for the motion detection map, and a reconstructed background with a reduced amount of noise, but conserving edges and contrast of the image. Then, we add the following second-order mixed potential,

$$V_{ij}^S(x_{i,t}, x_{j,t} \mid \mathbf{y}) = \frac{\beta^c}{g_i(\nabla \mathbf{y}_t)} \mathbf{1}_\omega^*(x_i^t) \mathbf{1}_\omega^*(x_j^t) \left[\frac{(x_i^t - x_j^t)^2 - K}{\sigma_i^2} \right] - \frac{\beta^m}{g_i(\nabla \mathbf{y}_t)} \mathbf{1}_\omega(x_i^t) \mathbf{1}_\omega(x_j^t), \tag{8}$$

where $g_i(\nabla \mathbf{y}_t) = \max(1, \parallel \nabla y_i^t \parallel^2)$. A combined spatial regularization of both types of values is achieved through this energy potential. First, a Gaussian continuous term is introduced in order to obtain homogeneous intensity regions for the objects in the background. This regularization is only done when both points are not in motion and is stronger for those points where the image gradient is small, in such a way that we avoid the blurring of edges. Then, regarding the motion detection map[2], we observe that the amount of regularization depends as well on the continuous part, that is, is favored in homogeneous intensity regions. The constant K is set to the value $K = \frac{1}{2}(x_{\max} - x_{\min})^2 = (255)^2/2$, centering the range of values for this term and is introduced in order to indeed favor this regularization when two neighboring points tend to have similar intensities. If $K = 0$, the whole term can become null in that case, suppressing the regularization between adjacent points over non-moving regions. Another term for the smoothness of the moving points is added as well, in order to improve regularization and reduce false negative detections.

[2] More precisely, its complement, the non-motion map.

4.3 Estimation

The complete expression for the energy is finally,

$$H(\mathbf{x}_t \mid \mathbf{y}, \mathbf{z}_{t-1}) = \sum_i \left\{ V_i^D(x_i^t \mid \mathbf{y}) + V_i^R(x_i^t \mid \mathbf{y}, \mathbf{z}_{t-1}) \right\} + \sum_{i,j} V_{ij}^S(x_i^t, x_j^t \mid \mathbf{y}), \quad (9)$$

and the problem reduces to the task of estimating the field $\mathbf{x}_t$ by minimizing H. The ICM (Iterated Conditioned Modes) algorithm is applied. The concept of iteratively maximizing the conditional densities is equally applicable to generalized mixed-state densities as (2) and is equivalent to the minimization of an energy function $H(\cdot)$. Then, for each point the following rule is applied:

$$x_i^t = \begin{cases} \omega & \text{if } H(x_i^t = \omega \mid \mathbf{X}_{\mathcal{N}_i}, \mathbf{y}) < H(x_i^t = x_i^* \mid \mathbf{X}_{\mathcal{N}_i}, \mathbf{y}) \\ x_i^* & \text{otherwise} \end{cases} \quad (10)$$

where $H(x_i^t \mid \mathbf{X}_{\mathcal{N}_i}, \mathbf{y})$ is the energy associated to the conditional mixed-state density obtained from (9) and x_i^* is the continuous value that minimizes its continuous part, i.e. when $x \neq \omega$, here equals:

$$x_i^* = \frac{\frac{\beta^c}{g_i(\nabla \mathbf{y}_t)} \sum_{i,j} x_j^t \mathbf{1}_\omega^*(x_j^t) + \gamma \alpha_i^R(\mathbf{y}_t, \mathbf{z}_{t-1})}{\frac{\beta^c}{g_i(\nabla \mathbf{y}_t)} \sum_{i,j} \mathbf{1}_\omega^*(x_j^t) + \gamma}. \quad (11)$$

Note that this value is in fact, the mean of the conditional continuous density in (2), that results to be Gaussian, and is the estimated value for the reference image at point i.

5 Results and Experimental Comparisons

We have applied our method to real sequences consisting of articulated and rigid motion. As well, we compare the results with the methods of Stauffer and Grimson [9] and Elgammal et al. [10], for which we obtained an implementation from `http://www.cs.ucf.edu/~jdever/code/scode.html` and `http://cvlab.epfl.ch/~tola/source_code.html` respectively. At the same time, we compare the performance of the full mixed-state model, with two sequential implementations based on simplified (non-mixed) versions of the proposed energy potentials, in order to show the importance of the mixed-state terms. Firstly, we have implemented a sequential algorithm using equation (5) and the second term of equation (6): the first step is to estimate the moving points and then, with a fixed detection map, the background is updated. No regularization is introduced, nor in the detection or the background reconstruction. Secondly, we have implemented another sequential algorithm, now including non-mixed regularization, that is, using the potential of equation (5), the second term of equation (6) and the second term of equation (8). In other words, we take out the mixed potentials from the energy.

For our method, we use the 8-point nearest neighbor set, as the neighborhood for the mixed-state Markov random field. The parameters of the model were set as following: $\gamma = 8$, $\beta^c = 1$, $\beta_m = 5$ and $c = 0.7$. For all the sequences these same values were used. This is justified observing equation (11). Assume all neighbor points are not in motion, then the estimated value for the background intensity is a weighted average between the 8 neighbors and the previous estimated background. Setting $\beta^c = 1$ we get a total weight of 8 for the surrounding points (if the local gradient is small), and then with $\gamma = 8$, we give the same weight to the previous estimated value. This situation establishes an equilibrium working point of the algorithm, from which we derived the order of magnitude of the parameters. β^m was set empirically in order to effectively remove isolated points. Anyway, the results practically did not show variations for $\beta^c \in [0.5, 1]$, $\gamma \in [8, 12]$ and $\beta^m \in [3, 6]$. This low sensitivity allowed us to fix a unique set of parameters for all the samples.

In Fig. 3 we show the result of comparing the different algorithms applied to the Basketball sequence. The method by Stauffer and Grimson shows false positively detected moving points in the background. The method by Elgammal et al. performs better, but has some problems at correctly achieving connected regions. The mixed-state method shows an improved regularization of the motion map, visually reducing false positives and false negatives, also compared with the sequential non-mixed versions of the algorithm.

Fig. 4 shows a complex scene of two man walking through a forest. In this example the background is not completely static as there is swaying vegetation. Our method supplies the best results discarding practically all the background motion, even compared with multi-modal density models. The proposed observations (Number of False Alarms) introduced in the discriminative term are able to cope with this kind of background dynamics.

Finally in Fig. 5 we show the result for a sequence of a Traffic Circle with multiple rigid motions. In this case, never during the sequence a complete background image is available. The cars continuously pass around the square entering and leaving the scene. The method by Stauffer and Grimson shows a deadlock situation due to the lack of training samples: initially the algorithm includes in the background some of the moving cars, resulting in a continuous positive wrong detection for subsequent frames and takes too long for the model to finally remove them from the reference image. Moreover, some regions of the background are never correctly updated. The same sequence tested with the non-parametric method of Elgammal et al. failed in generating valid results, resulting in an everywhere negative detection for mostly every point and every frame. The lack of training samples for the background, on which the method relies, is likely to be the cause of the failure.

For the method proposed here these problems are not present. The cars are well detected with less false positives for the mixed-state method. The algorithm is not able to distinguish the small cars entering the scene from the street in the top, grouping all in a single connected region. In this case, the separation between the cars in that region is about 4 pixels (the image is of size 256x256), which is in the order of the size of the considered neighborhoods used in the regularization terms. Nevertheless, it results in a well segmented scene where

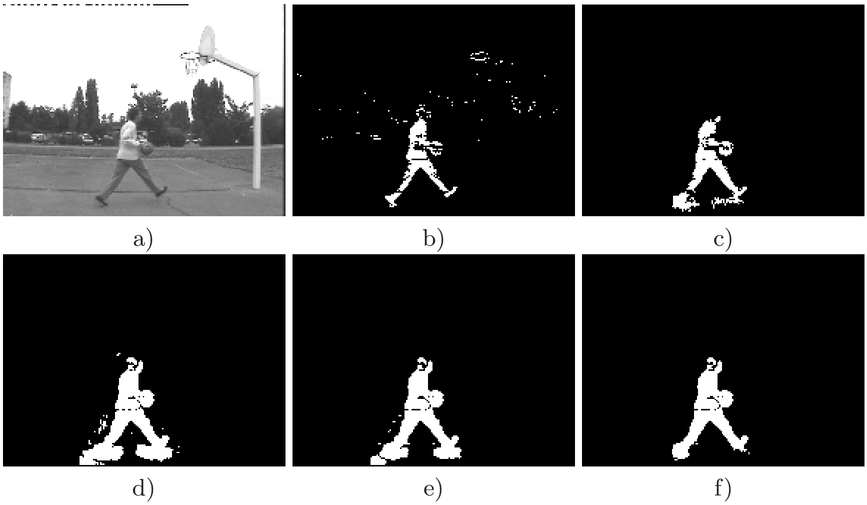


Fig. 3. Detection result for the Basketball sequence. a) Original image, b) Stauffer-Grimson method c) Elgammal et al. method, d) detection with a sequential detection-reconstruction method, without spatial regularization, e) detection using a sequential detection-reconstruction method with regularization, f) detection by our mixed-state method.

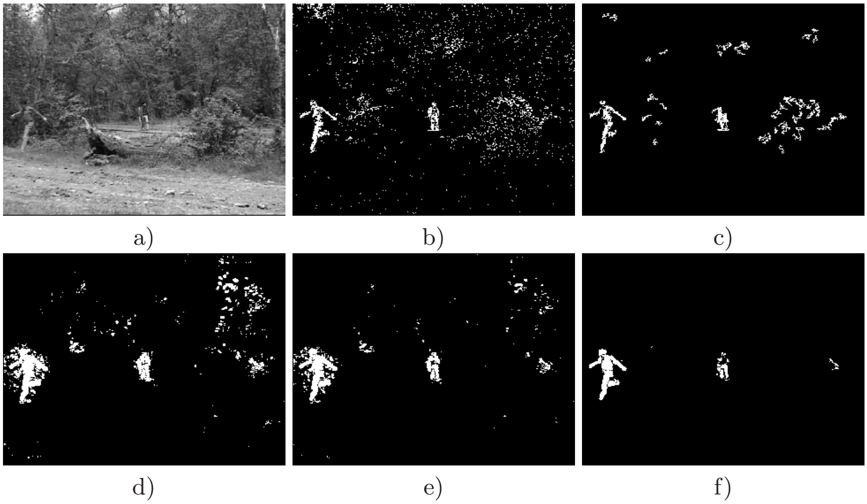
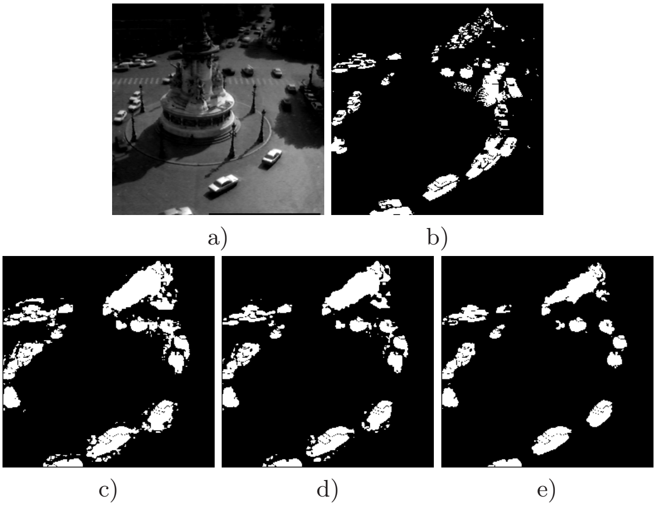


Fig. 4. Detection result for the Forest sequence. a) Original image, b) Stauffer-Grimson method c) Elgammal et al. method, d) detection with a sequential detection-reconstruction method, without spatial regularization, e) detection using a sequential detection-reconstruction method with regularization, f) detection by our mixed-state method.



a) b)

c) d) e)

Fig. 5. Detection result for the Traffic Circle sequence. a) Original image, b) Stauffer-Grimson method, c) detection with a sequential detection-reconstruction method, without spatial regularization, d) detection using a sequential detection-reconstruction method with regularization, e) detection by our mixed-state method. The method by Elgammal et al. did not generate a valid result due to the lack of training samples.

the regions occupied by the moving objects are obtained compactly. Note how most of the cars are indeed detected as uniformly connected regions.

5.1 Video Sequence Inpainting

The proposed algorithm generates estimates of the background image, not a model of it, viewed as a problem of reconstruction. The approach uses all the information about the background across time to build a complete image. The importance of this reconstruction not only has implications in the problem of motion detection, but also solves the problem of video sequence inpainting. In this case, moving objects can be removed from the scene as shown in Fig. 6. Moreover, the reconstruction implies smoothing of the background image, over homogeneous intensity regions, filtering out the observation noise, but preserving the edges. In the third row of Fig. 6 we display a small region for each sample, in order to more clearly observe the effect of the background reconstruction. In Fig. 6a), the basketball court is smoothed, and the lines are well preserved. In the Forest sequence 6b), we see how the algorithm preserves the texture of the trees and does not blur the intensity borders. In c), d) and e), the cars are correctly removed even in a complex situation where the background partially occludes the moving object, as in e), and the image noise is reduced as well.

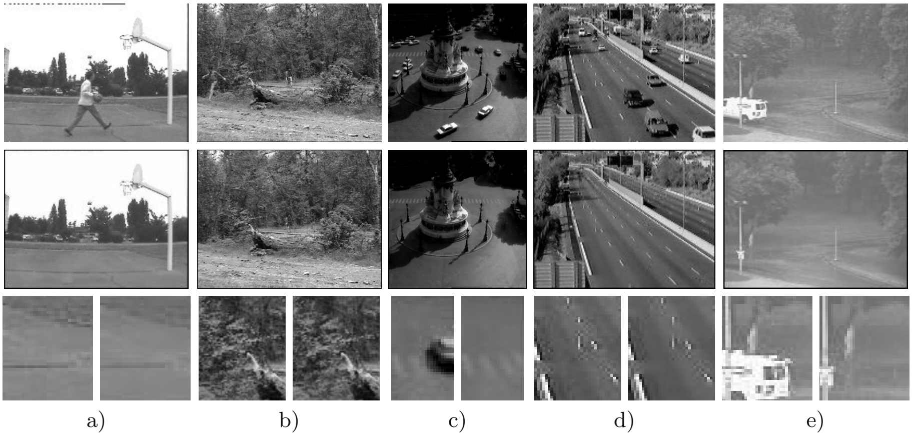


Fig. 6. Top row: original sequences. Center row: background images estimated with our method. Bottom row: a close-up over a small region of the original (left) and reconstructed (right) images. The spatio-temporal reconstruction of the background is achieved jointly with motion detection, resulting in virtually removing the moving objects from the scene. The reference image is also filtered over homogeneous intensity regions in order to reduce noise, but preserving borders.

6 Conclusions

In this paper, we have presented a new approach for addressing a complex problem as simultaneous motion detection and background reconstruction. The interaction between the two tasks was exploited, through a joint decision-estimation formulation, which reduces the problem to a unified step. This improves the regularization of the detection map w.r.t. existing background subtraction methods and against similar but sequential (non-simultaneous) strategies, resulting in more compact and well-defined detected regions. Another original contribution is the introduction of a *conditional* mixed-state random field that allows the integration of motion observations in the scheme.

The implications of considering these types of mixed-state models are enormous in computer vision, where high-level information, represented by abstract labels, can be introduced in an optimal way. Future applications include introduction of symbolic states for: borders (e.g. estimating discontinuous optical flow fields), detection of regions of interest (defined abstractly) or structural change detection (e.g., in remote sensing).

References

1. Kumar, S., Hebert, M.: Discriminative random fields. Int. J. Comput. Vision 68(2), 179–201 (2006)
2. Bouthemy, P., Hardouin, C., Piriou, G., Yao, J.F.: Mixed-state auto-models and motion texture modeling. Journal of Mathematical Imaging and Vision 25(3), 387–402 (2006)

3. Crivelli, T., Cernuschi-Frias, B., Bouthemy, P., Yao, J.F.: Mixed-state markov random fields for motion texture modeling and segmentation. In: Proc. IEEE Int. Conf. on Image Processing (ICIP 2006), Atlanta, USA, pp. 1857–1860 (2006)
4. Salzenstein, F., Pieczynski, W.: Parameter estimation in hidden fuzzy markov random fields and image segmentation. Graph. Models Image Process 59(4), 205–220 (1997)
5. Salzenstein, F., Collet, C.: Fuzzy markov random fields versus chains for multispectral image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 28(11), 1753–1767 (2006)
6. Cernuschi-Frias, B.: Mixed states markov random fields with symbolic labels and multidimensional real values. Technical Report 6255, INRIA (July 2007)
7. Hardouin, C., Yao, J.F.: Spatial modelling for mixed-state observations. Electronic Journal of Statistics (2008)
8. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.P.: Pfinder: real-time tracking of the human body. IEEE Trans. Pattern Anal. Mach. Intell. 19(7), 780–785 (1997)
9. Stauffer, C., Grimson, W.: Learning patterns of activity using real-time tracking. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 747–757 (2000)
10. Elgammal, A.M., Harwood, D., Davis, L.S.: Non-parametric model for background subtraction. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 751–767. Springer, Heidelberg (2000)
11. Parag, T., Elgammal, A., Mittal, A.: A framework for feature selection for background subtraction. In: CVPR 2006: Proc. of the 2006 IEEE Conf. on Computer Vision and Pattern Recognition, Washington, DC, USA, pp. 1916–1923 (2006)
12. Mittal, A., Paragios, N.: Motion-based background subtraction using adaptive kernel density estimation. In: CVPR 2004: Proc. of the 2004 IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp.II–302–II–309 (2004)
13. Sheikh, Y., Shah, M.: Bayesian modeling of dynamic scenes for object detection. IEEE Trans. Pattern Anal. Mach. Intell. 27(11), 1778–1792 (2005)
14. Monnet, A., Mittal, A., Paragios, N., Ramesh, V.: Background modeling and subtraction of dynamic scenes. In: Proc. of the Ninth IEEE Int. Conf. on Computer Vision, vol. 2, pp. 1305–1312 (2003)
15. Sun, J., Zhang, W., Tang, X., Shum, H.Y.: Background cut. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 628–641. Springer, Heidelberg (2006)
16. Bugeau, A., Pérez, P.: Detection and segmentation of moving objects in highly dynamic scenes. In: CVPR 2007: Proc. of the 2007 IEEE Conf. on Computer Vision and Pattern Recognition, Minneapolis, MI (2007)
17. Veit, T., Cao, F., Bouthemy, P.: An a contrario decision framework for region-based motion detection. International Journal on Computer Vision 68(2), 163–178 (2006)