Int J Comput Vis (2011) 94:295–316
DOI 10.1007/s11263-011-0429-z

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

Tomás Crivelli · Patrick Bouthemy · Bruno Cernuschi-Frías · Jian-feng Yao

Received: 29 May 2010 / Accepted: 25 February 2011 / Published online: 17 March 2011

Abstract In this work we present a new way of simultaneously solving the problems of motion detection and background image reconstruction. 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. The key of our joint approach is to define a single random process that can take two types of values, instead of defining two different processes, one symbolic (motion detection) and one numeric (background intensity estimation). It thus allows to exploit the (spatio-temporal) interaction between a decision (motion detection) and an estimation (intensity reconstruction) problem. Consequently, the meaning of solving both tasks jointly, is to obtain a single optimal estimate of such a process. The intrinsic interaction and simultaneity between both problems is shown to be better modeled within the so-called mixed-state statistical framework, which is extended here to account for symbolic states and conditional random fields. Experiments on real sequences and comparisons with existing motion detection methods support our proposal. Further implications for video sequence inpainting will be also discussed.

Keywords Motion detection · Background reconstruction · Mixed-state Markov models · Conditional random fields

T. Crivelli (✉) · B. Cernuschi-Frías
University of Buenos Aires, Buenos Aires, Argentina
e-mail: tomas.crivelli@gmail.com

B. Cernuschi-Frías
e-mail: bcf@ieee.org

T. Crivelli · P. Bouthemy
INRIA Rennes, Campus de Beaulieu, 35042 Rennes, France

P. Bouthemy
e-mail: patrick.bouthemy@inria.fr

B. Cernuschi-Frías
CONICET, Buenos Aires, Argentina

J.-f. Yao
IRMAR University of Rennes 1, Rennes, France
e-mail: jian-feng.yao@univ-rennes1.fr

1 Introduction

The problem of moving object detection from a video sequence is an open issue of great interest in image analysis. Solving it correctly is essential to computer vision systems that perform diverse and complex tasks as object tracking, sequence segmentation, object recognition, behavior analysis, and it is a crucial component in surveillance applications.

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 of the image and a reference value representing the background (Fig. 1), 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. Indeed, the simple inter-frame difference with a global threshold reveals itself as being very sensitive to usual phenomena as noise and illumination changes.

The problem consists in obtaining an accurate representation of the background or reference image and solving for motion detection by an appropriate comparison of the current and reference images. However, a "chicken-and-egg" situation arises when we want to set an optimal approach for both tasks: an accurate estimation of the background is

Fig. 1 Motion detection by background subtraction. Moving points are those where the current image and the reference image differ considerably

only possible if we know which regions of the image belong to it, that is, if we locate the moving objects; conversely, a correct motion detection is achieved if we have a relevant background representation.

A simplified approach may be considered by means of alternate decision/estimation steps, allowing one to solve each task separately and sequentially. This means to solve motion detection, assuming we know the background, and then updating the background model from points classified as background. However, there is no warranty that the scheme results in an optimal solution. Decomposing a simultaneous problem into a sequence of separate steps, and solving each of them in a sub-optimal fashion do not necessarily end up in an optimal solution for the whole problem. Moreover, the decision step involves a hidden symbolic variable to be determined. Consequently, it implies an inference process which may be complex.

We identify the problem as a *simultaneous decision-estimation problem*. One deals with a decision process (motion detection) together with an estimation process (background recovering). Consequently, we explicitly recognize two types of intervening values: *numeric* values to be estimated (the background image) and a *symbolic* value, associated to the motion detection task and represented by an abstract label. As separate (alternate) problems, they are solved in different domains: continuous vs. discrete. However, if we want to exploit the natural relation and interaction that exist between both tasks, we need to solve the problem in a unified framework involving a single domain, where symbolic (discrete) and numeric (continuous) states can be jointly modeled and/or recovered.

In this work, we present a new way of solving the aforementioned coupled problems jointly, based on the so-called *mixed-state statistical framework* (Hardouin and Yao 2008; Bouthemy et al. 2006). A preliminary version has been published before in Crivelli et al. (2008). The key of this approach is to define a single random variable that can take two types of values, instead of defining a pair of random variables for each image location, one symbolic and one numeric (Fig. 2). In view of this, we can redefine the problem of motion detection by background subtraction as the starting point for our proposal. Let us consider that a point in

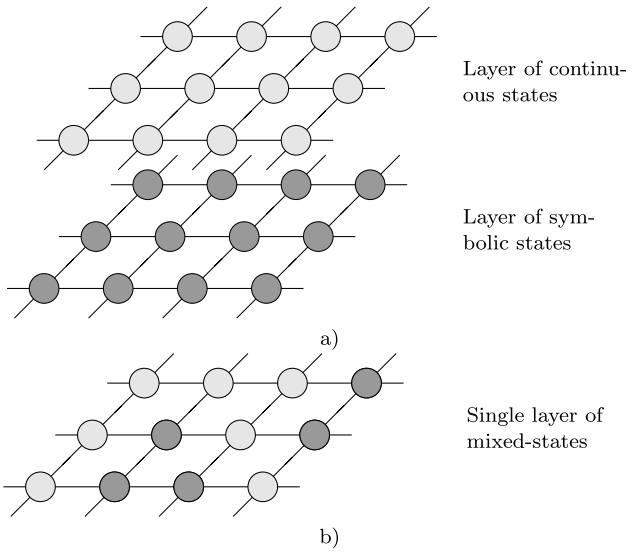


Fig. 2 (**a**) Two separate layers of continuous and symbolic states. (**b**) A single layer where both types of values are jointly modeled in a mixed-state domain

the image is a single process that can take either a symbolic value (or abstract label) accounting for the presence of motion, or a continuous numeric value associated to the brightness intensity of the reference image at that location. Consequently, the meaning of solving the motion detection and the background reconstruction jointly, is to obtain a single optimal estimate of such a process.

This paper is organized as follows. In Sect. 2 we discuss the advantages and drawbacks of state-of-the-art methods for motion detection by background subtraction. This will serve as a guide for defining our method in the following sections. In Sect. 3 the general mixed-state probabilistic framework is described and then we introduce the concept of mixed-states conditional random fields (MS-CRF). Then in Sect. 4 we specify a MS-CRF for the simultaneous problem of motion detection and background reconstruction. Results and experimental comparisons are discussed in Sect. 5.

2 Background Subtraction Techniques

For existing background subtraction methods, a necessary step consists in the learning of the background and this im-

plies 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 motion detection step. Anyway, a general consensus has been established to estimate a probability density for each background pixel. The simplest approach is to assume a single Gaussian law per pixel (see for example, Wren et al. 1997), whose parameters may be estimated by simple running averages or median filters. A valid criticism to this hypothesis is that the distribution of the intensity of a background pixel over time can vary considerably. In that direction, multi-modal density models seemed to perform better. Mixtures of Gaussians (Stauffer and Grimson 2000) and non-parametric models (Elgammal et al. 2000) have shown good results, able to deal with the variation of the background distribution. Several improvements on these ideas have been further developed (Zivkovic and van der Heijden 2006; Parag et al. 2006; Mittal and Paragios 2004) on how to efficiently update the background model. It is also worthy to mention approaches based on separating large data clusters representing the background and small clusters representing the foreground (Wright et al. 2009). Principal component analysis is used for separating low-rank approximations of the video (the reference image) from a sparse error component (the moving object).

However, they suffer several drawbacks. The approach does not assume spatial correlation between pixels, neither in the model of the background, nor in the binary detection map. To cope with this, posterior morphological operations are applied in order improve the resulting motion detection map. No regularization is proposed for the reference model. Also, points detected as foreground are incorporated to estimate the background model (called blind update, Elgammal et al. 2000) 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. Many heuristic corrections are usually applied in order to alleviate this drawback, but unfortunately, introducing others. Finally, these methods are sensitive to the initialization of the background model, particularly, when an initial image with no moving objects is not available in the video sequence.

Instead of looking at the temporal variation of point intensity statistics, region-based features, as used in Criminisi et al. (2006), Ko et al. (2008), are less sensitive to variations of the textured background, and are more robust in detecting foreground objects.

Unlike most approaches to moving object detection which detect objects by building adaptive models of the background, in Criminisi et al. (2006), Sheikh and Shah (2005) the foreground is also modeled. This permits to exploit the *temporal persistence* of a moving object. True foreground objects, as opposed to spurious noise, tend to maintain consistent colors and remain in the same spatial area. Thus, previous foreground information contains substantial evidence to be used at the current instant. Other approaches have proposed to model and infer multiple layers of moving objects combining deformable masks and foreground appearance maps (Jojic and Frey 2001).

The advantages of incorporating spatio-temporal context and regularization, in the background modeling and also the foreground one, are demonstrated for example in Sheikh and Shah (2005), Monnet et al. (2003), Migdal and Grimson (2005) by means of a Markov random field model and ARMA processes. In Bouthemy and Lalande (1993), a Markovian approach for motion detection exploiting temporal regularization between consecutive frames is proposed. In Bugeau and Pérez (2007), 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. Related to the class of energy-based methods for background subtraction, the work by Sun et al. (2006) on Object Cut, models the likelihood of each pixel belonging to foreground or background along with an improved spatial contrast term. This term is a penalty term when adjacent pixels are assigned with different labels (background or foreground), and the amount of penalization depends on how similar are the colors of the pixels. The method relies on a known (previously learned) background model and an adaptive update scheme is necessary. Finally, conditional random fields have been used before for background-foreground segmentation in Criminisi et al. (2006), integrating color and motion cues, and a temporal dependency model in the detection process.

Our review of background subtraction techniques has led us to make the following observations:

- Pointwise motion detection (Elgammal et al. 2000; Stauffer and Grimson 2000) is not enough for a correct segmentation. Spatial coherency and contextual information is needed (Sheikh and Shah 2005).
- Region-based image features are more robust to local variations (Migdal and Grimson 2005; Ko et al. 2008).
- Motion cues combined with intensity cues is better than considering only one or the other (Mittal and Paragios 2004).
- Regularization on the motion detection (symbolic) map should be enforced on neighboring points with similar intensity (numeric) values (Sun et al. 2006; Criminisi et al. 2006). Thus, it was pointed out in the literature that there exists a close interaction between the symbolic and the numeric processes.

- Results of existing methods depend strongly on a background-foreground learning stage (Elgammal et al. 2000; Sun et al. 2006; Criminisi et al. 2006)

In the following sections we propose to deal with each of these issues by combining the mixed-state framework with the conditional random fields (CRF) approach (Lafferty et al. 2001).

3 The Mixed-State Approach

3.1 Related Work and Connections

The concept of a random process that can take different types of values (either continuous or abstract) includes diverse situations. A mixed discrete-continuous Markov random field is formulated in Bouthemy et al. (2006) for the modeling of dynamic or motion textures. It is demonstrated that the normal flow scalar motion observations extracted from these video sequences, show a discrete value at zero (null-motion) and a Gaussian continuous distribution for the rest of the values. This model was extended in Crivelli et al. (2006, 2009) and applied to the problems of motion texture segmentation, recognition and tracking. For these applications, the issue is different than for simultaneous decision-estimation problems. The mixed nature operates on the observation itself.

In previous works on fuzzy pixels classification as Salzenstein and Pieczynski (1997) and Salzenstein and Collet (2006), the authors introduce a class of fuzzy MRF's or fuzzy Markov chains where each state variable, or classification variable, $x_i \in [0, 1]$ represents a classification rate. The fuzzy principle implies that the two hard classification states $x_i = 0$ or $x_i = 1$ have a positive probability while all the soft classification states, i.e. $x_i \in (0, 1)$ follow a continuous distribution. Indeed, fuzzy random fields, concept originally introduced in Caillol et al. (1993), are instances of spatial mixed-state models with numeric discrete part. Also a class of Markov chains with mixed states appeared in Carincotte et al. (2004, 2006), allowing the coexistence of a hard and fuzzy segmentation.

We can mention other models that exploit the interaction between symbolic and numeric values in computer vision decision problems. The *line process* introduced by Geman and Geman (1984) is an unobservable binary process $\mathbf{L}$ for edge elements. These authors regard the original image $\mathbf{I}$ as a marginal process from an extended joint field $\mathbf{X} = (\mathbf{I}, \mathbf{L})$ which is recovered from image observations. The idea behind this formulation is that the presence of an edge between two image locations, breaks the link between them and accounts for a discontinuity. Later in Black and Rangarajan (1996), the line process is viewed as a way of rejecting outliers giving an equivalence with robust estimators. In Wu and Chung (2007) a more sophisticated line process is used for image segmentation.

Finally, in hybrid Bayesian networks (Murphy 1999; Koller et al. 1999) the generalization is that a discrete node can have continuous parents and a continuous node can have discrete parents. The first case is useful for modeling threshold phenomena, while the second is associated to a sort of model selection state. The nature of the values taken by each random variable (parent or child) associated to a node is fixed to be continuous or discrete. What this formulation permits is to model discrete-continuous interaction, but not to infer if a node is discrete or continuous. Other so-called hybrid approaches follow a similar formulation as the original line process proposed in Geman and Geman (1984), by performing joint inference of two coupled fields, one discrete and one continuous, by means of an EM-like strategy (Lu et al. 2009).

3.2 Mixed-State Random Fields

In this work we extend the original idea of Bouthemy et al. (2006) to more general random fields where the discrete part may take abstract labels or values related to a decision problem. Our proposal is different from previous symbolic-numeric approaches described in the last sub-section. We deal with a single random field on a lattice, where each point may display either symbolic or a numeric value. First, this avoids defining and modeling two different processes as it was done with the line process in Geman and Geman (1984). No marginalization is needed to obtain the desired field, no complementary hidden states have to be introduced. On the other side, the nature (symbolic or numeric) of each site variable is not fixed as it occurs in hybrid Bayesian networks (Murphy 1999; Koller et al. 1999). In contrast, determining the optimum state is what allows us to solve two coupled problems in a single estimation process.

Definition 1 (Mixed-state random variable) Let $\{l\}$ be a symbolic state or label and let $\mathcal{I} \subset \mathbb{R}$ be an interval of the real line. A mixed-state random variable x is defined as taking values in a mixed-state space $\mathbb{M} = \{l\} \cup \mathcal{I}$ and is constructed as follows. With probability $\rho \in [0, 1]$, set $x = l$, and with probability $1 - \rho$, x is continuously distributed in $\mathcal{I}$.

Since symbolic labels such as l do not have any algebraic structure, a probability distribution function cannot be defined to characterize the random variable. One proceeds directly to define a probability measure for a mixed-state random variable, resorting on the theory of measure and integration (Cernuschi-Frias 2007). We can then construct a probability density for x, defined as

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

with $\rho \in [0,1]$, $\rho^* = 1-\rho$ and where we define the characteristic functions

$$\mathbf{1}_l(x) = \begin{cases} 1 & \text{if } x = l, \\ 0 & \text{if } x \neq l, \end{cases} \qquad \mathbf{1}_l^*(x) = 1 - \mathbf{1}_l(x) \tag{2}$$

and $p^c(x)$ is a continuous probability density function. The density $p(x)$ in (1) is given with respect to a reference measure $m(dx) = m_l(dx) + \lambda(dx)$, where $m_l(dx)$ is a counting measure for the value l and $\lambda(dx)$ is the usual Lebesgue measure, i.e. the length of the interval in the real line. Interpret this equation as follows: the density function $p(x)$ assigns a probability mass ρ to the discrete value, and acts as a continuous density function $p^c(x)$ for the continuous values.

Let us pin things down and consider a first approach to the problem of motion detection and background reconstruction with a mixed-state model. Define a mixed-state random variable x_i for each location i of the image plane. Define l as the symbolic state that indicates a detected moving point, and consider the interval of the real line $\mathcal{I} = [0, 255]$, i.e., the range of gray level intensity values for the background image.

We are now ready to propose a first very simple mixed-state model. Following equation (1) we write:

$$p(x_i) = \rho_i \mathbf{1}_l(x_i) + \rho_i^* \mathbf{1}_l^*(x_i) \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i - m_i)^2}{2\sigma_i^2}}. \tag{3}$$

With probability ρ_i, $x_i = l$, i.e., the location corresponds to a moving point, and with probability $\rho_i^* = 1 - \rho_i$, the location corresponds to a background intensity value perturbed by Gaussian noise. We may thus estimate the value of x_i by point-wise maximization of (3). Note that assigning an intensity value implies considering the point as background, so that (background) estimation is performed simultaneously with (motion) detection.

It should be clear that this simple model will be far from performing well in most of the situations and that we need to incorporate a more complex scheme that allows us to introduce spatial interaction, and to enforce correlation within the random field and between continuous and symbolic values, as we describe in the next section.

3.3 Mixed-State Auto-Models with Symbolic Values

Markov random field models have been applied successfully to estimation problems (e.g. texture modeling and analysis, Lorette et al. 2000, optical flow estimation, Heitz and Bouthemy 1993; Li and Huttenlocher 2008, image restoration and denoising, Chen and Tang 2007; Geman and Geman 1984) as well as decision problems (e.g. image segmentation, Collet and Murtagh 2004; Salzenstein and Collet 2006; Crivelli et al. 2006; Benboudjema and Pieczynski 2007; Blanchet and Forbes 2008, motion detection, Benedek et al. 2007; Bouthemy and Lalande 1993, edge detection, Wu and Chung 2007, structural change detection, Kasetkasem and Varshney 2002). Our motivation have been to exploit the power of mixed-state MRF's for simultaneous decision-estimation problems.

How can we formulate a mixed-state Markov random field in order to include continuous and symbolic states within a single random field model?

Let $S = \{1 \dots N\}$ be a lattice of points or image locations such that $\mathbf{x} = \{x_i\}_{i \in S}$. Define $\mathbf{x}_{\mathcal{N}_i}$ as the set of random variables in a neighborhood $\mathcal{N}_i$ of location i, i.e., $\mathbf{x}_{\mathcal{N}_i} = \{x_i\}_{i \in \mathcal{N}_i}$. Then the Markovian property is expressed in the mixed-state conditional densities:

$$\begin{aligned} &p(x_i \mid \mathbf{x}_{S \setminus \{i\}}) \\ &\quad = p(x_i \mid \mathbf{x}_{\mathcal{N}_i}) \\ &\quad = \rho(\mathbf{x}_{\mathcal{N}_i}) \mathbf{1}_l(x_i) + \rho^*(\mathbf{x}_{\mathcal{N}_i}) \mathbf{1}_l^*(x_i) p^c(x_i \mid \mathbf{x}_{\mathcal{N}_i}), \end{aligned} \tag{4}$$

where $\rho(\mathbf{x}_{\mathcal{N}_i}) = P(x_i = l \mid \mathbf{x}_{\mathcal{N}_i})$. Equation (4) defines the local characteristics of a mixed-state random field with a symbolic discrete state. However, they cannot be chosen arbitrarily for every point as they must comply with a well-defined joint distribution $p(\mathbf{x})$.

We adopt the formulation introduced by Hardouin and Yao (2008), who generalize the auto-models of Besag (1974) to the so-called *multiparameter auto-models*. According to these authors, if $p^c(x_i \mid \mathbf{x}_{\mathcal{N}_i})$ belongs to the d-parameter exponential family of distributions, the mixed-state conditional density (4) belongs to the $(d+1)$-parameter exponential family. This leads to:

$$\log p^c(x \mid \mathbf{x}_{\mathcal{N}_i}) = -\{\tilde{\boldsymbol{\Theta}}_i^T(\mathbf{x}_{\mathcal{N}_i}) \tilde{\mathbf{S}}_i(x) + \tilde{C}_i(x) + \tilde{D}_i(\mathbf{x}_{\mathcal{N}_i})\} \tag{5}$$

with $\tilde{\mathbf{S}}_i(x_i) \in \mathbb{R}^d$, $\tilde{\boldsymbol{\Theta}}_i(\mathbf{x}_{\mathcal{N}_i}) \in \mathbb{R}^d$, $\tilde{C}_i(x_i)$ and $\tilde{D}_i(\mathbf{x}_{\mathcal{N}_i}) \in \mathbb{R}$. It then results that

$$\log p(x_i \mid \mathbf{x}_{\mathcal{N}_i}) = -\{\boldsymbol{\Theta}_i^T(\mathbf{x}_{\mathcal{N}_i}) \mathbf{S}_i(x_i) + C_i(x_i) + D_i(\mathbf{x}_{\mathcal{N}_i})\}$$

with

$$\begin{aligned} &\mathbf{S}_i(x) = [\mathbf{1}_l^*(x_i), \mathbf{1}_l^*(x_i) \tilde{\mathbf{S}}_i(x_i)]^T, \\ &\boldsymbol{\Theta}_i(\mathbf{x}_{\mathcal{N}_i}) = \left[\log \frac{\rho_i^*(\mathbf{x}_{\mathcal{N}_i})}{\rho_i(\mathbf{x}_{\mathcal{N}_i})} + \tilde{D}_i(\mathbf{x}_{\mathcal{N}_i}), \tilde{\boldsymbol{\Theta}}_i^T(\mathbf{x}_{\mathcal{N}_i}) \right]^T, \\ &C_i(x) = \mathbf{1}_l^*(x_i) \tilde{C}_i(x_i), \\ &D_i(\mathbf{x}_{\mathcal{N}_i}) = \log \rho_i(\mathbf{x}_{\mathcal{N}_i}). \end{aligned} \tag{6}$$

The second assumption is that the family of conditional densities in (4) correspond to a second-order mixed-state Markov random field, i.e. $p(\mathbf{x}) = \exp -Q(\mathbf{x})/Z$ with

$$Q(\mathbf{x}) = \sum_i V_i(x_i) + \sum_{i,j} V_{ij}(x_i, x_j). \tag{7}$$

With the aforementioned hypothesis it was shown in (Bouthemy et al. 2006; Hardouin and Yao 2008) that

$$V_i(x_i) = \boldsymbol{\alpha}_i^T \cdot \mathbf{S}_i(x_i) + C_i(x_i), \tag{8}$$

$$V_{ij}(x_i, x_j) = \mathbf{S}_i(x_i)^T \boldsymbol{\beta}_{ij} \mathbf{S}_j(x_j) \tag{9}$$

with $\boldsymbol{\beta}_{ij} \in \mathbb{R}^{(d+1)\times(d+1)}$ and $\boldsymbol{\alpha}_i = [\alpha_1 \dots \alpha_{d+1}]^T \in \mathbb{R}^{d+1}$.

Discussion A mixed-state auto-model is defined by either the conditional densities (4) or by the potential functions (8) and (9). Both are in turn defined by the parameters $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_{ij}$. Now, let us decompose these parameters by writing

$$\boldsymbol{\beta}_{ij} = \begin{pmatrix} d_{ij} & \mathbf{Y}_{ij}^T \\ \mathbf{Y}_{ij} & \widetilde{\boldsymbol{\beta}}_{ij} \end{pmatrix}, \qquad \boldsymbol{\alpha}_i = [\alpha_i^D \quad \widetilde{\boldsymbol{\alpha}}_i^T]^T, \tag{10}$$

where $\mathbf{Y}_{ij}^T$ is the first row of $\boldsymbol{\beta}_{ij}$ minus the first element, d_{ij}, and $\mathbf{Y}_{ij}$ is the first column of $\boldsymbol{\beta}_{ij}$ minus the first element. $\widetilde{\boldsymbol{\beta}}_{ij}$ is the lower-right $d \times d$ submatrix of $\boldsymbol{\beta}_{ij}$. Equivalently, $\widetilde{\boldsymbol{\alpha}}_i$ is a d-dimensional vector, and α_i^D the first element of $\boldsymbol{\alpha}_i$.

In view of this, we can write the shape of the mixed-state Gibbs energy as follows:

$$\begin{aligned} Q(\mathbf{x}) = & \sum_i \alpha_i^D \mathbf{1}_l^*(x_i) + \widetilde{\boldsymbol{\alpha}}_i^T \mathbf{1}_l^*(x_i) \tilde{\mathbf{S}}_i(x_i) + C_i(x_i) \\ & + \sum_{\langle i,j \rangle} d_{ij} \mathbf{1}_l^*(x_i) \mathbf{1}_l^*(x_j) \\ & + \sum_{\langle i,j \rangle} \mathbf{Y}_{ij}^T \mathbf{1}_l^*(x_i) \mathbf{1}_l^*(x_j) \tilde{\mathbf{S}}_j(x_j) \\ & + \sum_{\langle i,j \rangle} \mathbf{Y}_{ij}^T \mathbf{1}_l^*(x_j) \mathbf{1}_l^*(x_i) \tilde{\mathbf{S}}_i(x_i) \\ & + \sum_{\langle i,j \rangle} \mathbf{1}_l^*(x_i) \mathbf{1}_l^*(x_j) \tilde{\mathbf{S}}_i(x_i)^T \tilde{\boldsymbol{\beta}}_{ij} \tilde{\mathbf{S}}_j(x_j). \end{aligned} \tag{11}$$

Note that this model allows us to introduce different types of terms in the mixed-state Gibbs energy function. On one side we have purely discrete terms of the form $\alpha_i^D \mathbf{1}_l^*(x_i)$ or $d_{ij} \mathbf{1}_l^*(x_i) \mathbf{1}_l^*(x_j)$ as in a discrete Markov random field. On the other side, we can include unary continuous terms $\widetilde{\boldsymbol{\alpha}}_i^T \mathbf{1}_l^*(x_i) \tilde{\mathbf{S}}_i(x_i)$ or second-order terms as $\mathbf{1}_l^*(x_i) \mathbf{1}_l^*(x_j) \tilde{\mathbf{S}}_i(x_i)^T \tilde{\boldsymbol{\beta}}_{ij} \tilde{\mathbf{S}}_j(x_j)$. $\tilde{\mathbf{S}}_i(x_i)$ is a function of the continuous values of x_i. Finally, we are able to include mixed-state second-order terms as $\mathbf{Y}_{ij}^T \mathbf{1}_l^*(x_i) \mathbf{1}_l^*(x_j) \tilde{\mathbf{S}}_j(x_j)$. In this latter case, the model is able to exploit the interaction between continuous and symbolic states of neighboring points.

Indeed, many applications in computer vision are formulated directly as an energy-maximization problem where the energy terms are Gibbs potentials, usually up to second order cliques (Fablet and Bouthemy 2003; Wu and Chung 2007; Heitz and Bouthemy 1993; Elfadel and Picard 1994; Kumar and Hebert 2006; Lafferty et al. 2001). In these cases, the model is completely designed through the energy function, although the conditional densities can be eventually obtained, for example when applying certain optimization methods (e.g. ICM, Besag 1974). For us, (11) is the basis for designing a mixed-state energy that corresponds to a mixed-state Markov random field.

3.4 Conditional Random Fields with Mixed-States

In the MRF framework, the problem of estimating a random field $\mathbf{x}$ from a set of (image) observations $\mathbf{y}$ is expressed using the Bayes rule as

$$\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}) \propto \max_{\mathbf{x}} p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}). \tag{12}$$

For classical MRFs models (Besag 1986; Geman and Geman 1984; Chellappa 1985), the prior knowledge on $\mathbf{x}$ is modeled as a Markov random field and $p(\mathbf{y} \mid \mathbf{x})$ is the observation model. In order to obtain a computationally tractable (also Markovian) posterior distribution, some restrictive assumptions need to be imposed on defining $p(\mathbf{y} \mid \mathbf{x})$. For example, assuming a factorized form $p(\mathbf{y} \mid \mathbf{x}) = \prod_i p(y_i \mid x_i)$ the markovianity is assured. However, this is a strong restriction that may not be able, for example, to account for textured patterns.

As pointed out in Pieczynski and Tebbache (2000), what one usually seeks is the markovianity of $p(\mathbf{x} \mid \mathbf{y})$. This can be guaranteed by directly assuming the markovianity of $p(\mathbf{x}, \mathbf{y})$ in the form of pairwise Markov random fields (PMRF) (Pieczynski and Tebbache 2000). Thus, $p(\mathbf{x})$ need not be Markovian. This relaxes the restrictions of classical MRFs and permits to build more complex, and yet tractable, models. This approach was later extended in the triplet Markov field (TMF) model (Benboudjema and Pieczynski 2007; Blanchet and Forbes 2008) introducing a third (auxiliary) process $\mathbf{u}$ and assuming $(\mathbf{x}, \mathbf{y}, \mathbf{u})$ is now Markovian. This allows having a non-Markovian $(\mathbf{x}, \mathbf{y})$ and thus, a more general setting.

The latter approaches require modeling the observation process $\mathbf{y}$, either in the form of $p(\mathbf{y} \mid \mathbf{x})$ or through the joint distributions $p(\mathbf{x}, \mathbf{y})$ or $p(\mathbf{x}, \mathbf{y}, \mathbf{u})$. This can sometimes be viewed as a limitation if one wants to introduce arbitrary observations in the inference process of $\mathbf{x}$.

A different approach which avoids modeling $\mathbf{y}$ is given by the so-called *conditional random fields* (*CRFs*) framework (Lafferty et al. 2001; Kumar and Hebert 2006; Wang et al. 2006) which has gained interest in the last years. It considers a different point of view in estimating the posterior probabilities over $\mathbf{x}$ given the observations. The idea is to directly model the posterior $p(\mathbf{x} \mid \mathbf{y})$ and directly imposing its Markovian form. As a consequence, this conditional probability can depend on arbitrary features $\mathbf{y}$ without making any model approximations as one does not have to take

care of its distribution (Kumar and Hebert 2006). This of course permits to relax any independence assumption. Furthermore, it allows us to define these models in a flexible way, in particular it enables to exploit a large set of observations (e.g., a block) at each site, something that in the classical MRFs notably increases the complexity of the model. That is, it is able to integrate at an image location any information extracted from the input data and obtained across any spatial or temporal (or both) neighborhoods, or information from previously reconstructed variables, or even the association of both.

It is not our intention to extensively describe the conditional random fields theory but to exploit its advantages within the mixed-state framework. Then, it is enough to give the following extension to the definition given in Lafferty et al. (2001):

Definition 2 (Mixed-state Conditional Random Field (MS-CRF)) Let $\mathbf{x}$ be a mixed state random field and $\mathbf{y}$ an observation process. Then $(\mathbf{x}, \mathbf{y})$ is said to be a *mixed-state conditional random field* if $\mathbf{x}$ conditioned on $\mathbf{y}$ is a mixed-state Markov random field.

In addition, this framework will permit to involve not only the comparison between the current image and the reference image but to explicitly integrate motion measurements obtained between consecutive images, contributing to make the overall scheme complete, accurate and powerful.

Introducing the observations $\mathbf{y}$ in the mixed-state automodel is straightforward, by making the parameters depend on $\mathbf{y}$, i.e. $\boldsymbol{\alpha}_i^T(\mathbf{y})$ and $\boldsymbol{\beta}_{ij}(\mathbf{y})$, and in turn, $\alpha_i^D(\mathbf{y})$, $\widetilde{\boldsymbol{\alpha}}_i^T(\mathbf{y})$, $d_{ij}(\mathbf{y})$, $\mathbf{Y}_{ij}^T(\mathbf{y})$, $\tilde{\boldsymbol{\beta}}_{ij}(\mathbf{y})$.

4 A MS-CRF for Simultaneous Motion Detection and Background Reconstruction

4.1 Our Method

Recall Sect. 2 where we have discussed several aspects that the method has to take into account in order to solve the problems of motion detection and background estimation. Now, we are ready to deal with each of these issues:

- We introduce spatial context and correlation in both types of values by exploiting a random field model with second order potentials as in (11).
- We exploit both image intensity and motion observations as input for the inference process. This can be done thanks to the Conditional Random Fields framework.
- We exploit the interaction between estimation (background intensity) and detection (moving points). The mixed-state approach allows us to achieve this joint modeling by designing the continuous, discrete and mixed potentials involved in (11).
- We solve the two problems in a single inference step. Optimization of mixed-state fields implies obtaining both types of values at the same time and in a unified way.

We now specify the MS-CRF that is able to handle simultaneously the problems of motion detection and background reconstruction. As mentioned before, there is a strong coupling between the two tasks.

4.2 Definitions

Let us call $I(t) = \{I_i(t)\}_{i \in S}$ the intensity image at time t, where $I_i(t) \in [0, 255]$ is the brightness intensity value at location $i \in S = \{1 \dots N\}$ of the image grid. Then $\mathbf{I}_t = \{I(t)\}_t$ is a sequence of images that we call *observations*. We define a mixed-state random field $\mathbf{x}_t = \{x_i^t\}_{i \in S}$ for time instant t, where $x_i^t \in \mathbb{M} = \{l\} \cup [0, 255]$ is a mixed-state random variable.

4.3 Background Update Strategy

Suppose we have an estimate of the mixed-state field $\mathbf{x}_t$ for a given instant t, that is, the location of the moving points and the estimated intensity values 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 image as follows:

$$z_i^t = \begin{cases} x_i^t & \text{if } x_i^t \neq l, \\ z_i^{t-1} & \text{otherwise.} \end{cases} \tag{13}$$

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 keep this value as a background intensity value. As the objects in the scene move, we can progressively reconstruct 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.

4.4 Design of the Energy Terms

Let us call $Q(\mathbf{x}_t \mid \mathbf{I}_{t+1}, \mathbf{z}_{t-1})$ the energy function associated to a *conditional* mixed-state Markov random field, given the observations $\mathbf{I}_{t+1}$[1] and the previously available background image $\mathbf{z}_{t-1}$. In the sequel we define the mixed-state energy terms.

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 at a point

[1] We will see shortly why we use images up to $t + 1$.

Table 1 Energy potentials of the conditional mixed-state model for the motion detection and background reconstruction method

(a)	$V_i^D(x_i^t \mid \mathbf{I}_{t+1}) = \alpha_i^D(\mathbf{I}_{t+1})\mathbf{1}_l^*(x_i^t)$
	$\alpha_i^D(\mathbf{I}_{t+1}) = -\log NFA(R_i)$ (see (17))
(b)	$V_i^R(x_i^t \mid \mathbf{I}_{t+1}, \mathbf{z}_{t-1}) = \gamma\left[\mathbf{1}_l^*(x_i^t)\frac{[x_i^t - m(z_i^{t-1}, I_i(t))]^2}{\sigma_i^2} + \mathbf{1}_l(x_i^t)\alpha_i^R(I(t), \mathbf{z}_{t-1})\right]$
	$\alpha_i^R(I(t), \mathbf{z}_{t-1}) = \sigma_i^2[n^{-1}\sum_{j\in\mathcal{N}_i}(z_j^{t-1} - I_j(t))]^{-2}$
	$m(z_i^{t-1}, I_i(t)) = cz_i^{t-1} + (1-c)I_i(t)$
(c)	$V_{ij}^S(x_i^t, x_j^t \mid \mathbf{I}_{t+1}) = \frac{\beta^c}{g_i(\nabla I(t))}\mathbf{1}_l^*(x_i^t)\mathbf{1}_l^*(x_j^t)\left[\frac{(x_i^t - x_j^t)^2 - K}{\sigma_i^2}\right] - \frac{\beta^m}{g_i(\nabla I(t))}\mathbf{1}_l(x_i^t)\mathbf{1}_l(x_j^t)$
	$g_i(\nabla I(t)) = \max(1, \parallel \nabla I_i(t) \parallel^2)$

given the observations; the *reconstruction* terms, involved in the estimation of the background intensity values, 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. In Table 1 we give the complete expressions. The mixed-state energy is therefore given by:

$$Q(\mathbf{x}_t \mid \mathbf{I}_{t+1}, \mathbf{z}_{t-1}) = \sum_i V_i^D(x_i^t \mid \mathbf{I}_{t+1}) + \sum_i V_i^R(x_i^t \mid \mathbf{I}_{t+1}, \mathbf{z}_{t-1}) + \sum_{\langle i,j\rangle} V_{ij}^S(x_i^t, x_j^t \mid \mathbf{I}_{t+1}). \quad (14)$$

The objective is to minimize this expression with respect to the mixed-state field $\mathbf{x}_t$ at each time instant. This implies minimizing the contribution of the potentials $V_i^D(x_i^t \mid \mathbf{I}_{t+1})$, $V_i^R(x_i^t \mid \mathbf{I}_{t+1}, \mathbf{z}_{t-1})$ and $V_{ij}^S(x_i^t, x_j^t \mid \mathbf{I}_{t+1})$.

Discriminative Term The discriminative term $V_i^D(x_i^t \mid \mathbf{I}_{t+1})$ (Table 1a) is related to the symbolic part of the field, which can be associated to the motion detection map. The weight $\alpha_i^D(\mathbf{I}_{t+1})$ depends on the observations and aims at tuning the belief of presence of motion at a point. The idea is that, when motion is present, $\alpha_i^D(\mathbf{I}_{t+1})$ should take a large value so that we penalize that $\mathbf{1}_l^*(x_i^t) = 1$ (or equivalently, we favor $x_i^t = l$). Conversely, a low value of $\alpha_i^D(\mathbf{I}_{t+1})$ favors $x_i^t \neq l$.

Here we adopt the *a-contrario* decision framework (Veit et al. 2006) for obtaining $\alpha_i^D(\mathbf{I}_{t+1})$. In this method moving regions appear as low probability events in a model corresponding to the absence of moving objects in the scene, namely a model of the background.

In general terms, a point in the image is likely to correspond to a moving object if its local normal flow magnitude is important, which is defined as

$$\upsilon_i^{(n)}(t) = \frac{|\frac{\partial I_i(t)}{\partial t}|}{\|\nabla I_i(t)\|}. \quad (15)$$

This quantity is computed between two consecutive frames of the sequence: $I(t-1)$ and $I(t)$. In order to deal with occlusion and disocclusion of the scene background by moving objects, a three-image scheme is considered. Taking $I(t)$ as the central image, the normal flow magnitude map is obtained for the pair $I(t-1), I(t)$ and for the pair $I(t), I(t+1)$. Then, the minimum value $\upsilon_i^{(n,min)}(t) = \min[\upsilon_i^{(n)}(t), \upsilon_i^{(n)}(t+1)]$ is kept as the considered measure. Looking forward and backward in time ensures that a meaningful motion observation is obtained, since the two pairs of images cannot be simultaneously affected by an occluding situation at the same time. Taking the minimum avoids assigning high motion values to the static background.

In an image with no moving objects present, the local motion measures can be assumed to derive from an independent and identically distributed temporal noise. The background is assumed to dominate the foreground, and thus, the inverse cumulative distribution function $F(\mu) = P(\upsilon_i^{(n,min)}(t) > \mu)$ of the normal flow magnitude for the background can be learned empirically from the whole image. Now, consider a region R_i around image location i and let k_μ denote the observed number of pixels at which the motion measure exceeds the threshold μ. According to the learned background distribution, the probability that k_μ or more motion values of a total of n, exceed μ is the tail of a binomial distribution:

$$B(k_\mu, n, F(\mu)) = \sum_{j=k_\mu}^{n} \binom{n}{j} F(\mu)^j (1 - F(\mu))^{n-j}. \quad (16)$$

This probability measures how likely the background model is for displaying an observation of at least k_μ exceeding motion values. It corresponds to the probability of rejecting the hypothesis of no-motion although it is true, when k_μ is viewed as a threshold for detecting a moving point. Then, it can be interpreted as a false alarm rate for region R_i.

Setting the threshold μ arbitrarily may be problematic as a suitable value may depend on the image and the region. To avoid this, a set of N_μ thresholds μ_j are tested and the minimum false alarm probability $PFA(R_i) =$

Fig. 3 Initial motion detection by computing the Number of False Alarms. The motion map is obtained by thresholding this quantity as explained in Veit et al. (2006). From *left* to *right*: the results are shown for the sequences Basketball, Forest, Traffic Circle, Route and Van. Note that this quantity, with the basic implementation used here, over-regularizes the motion detection map at that stage, as it is a block-based detection strategy

$\min_{j=1..N_\mu} B(k_{\mu_j}, n, F(\mu_j))$ is computed. Taking the minimum means that it is sufficient that one of the probabilities $B(k_{\mu_j}, n, F(\mu_j))$ is low to consider that the region R_i does not correspond to the background model.

Instead of considering the false alarm probability as in usual hypothesis testing, the method proposes to compute the average number of occurrences of the motion detection event under the hypothesis of the background model, termed Number of False Alarms (NFA) and defined as

$$\begin{aligned} NFA(R_i) &= N_{R_i} \cdot N_\mu \cdot PFA(R_i) \\ &= N_R \cdot N_\mu \cdot \min_{j=1..N_\mu} B(k_{\mu_j}, n, F(\mu_j)) \end{aligned} \tag{17}$$

where N_R is the number of tested regions in the image and N_μ the number of tested thresholds. In Veit et al. (2006) the number of candidate regions N_R can vary across the sequence by applying a meaningful region extraction algorithm. In our case, we have implemented the simplest scheme where we compute the value of $NFA(R_i)$ for each image location over square regions of a fixed size and thus N_R and N_μ are constants so that $NFA(R_i)$ depends essentially on $PFA(R_i)$. We fix $N_\mu = 10$ where the tested μ_j are those corresponding to regularly spaced probabilities $F(\mu_j) = p\frac{N_\mu - j + 1}{N_\mu}$, $j = 1 \ldots N_\mu$ with $p = F(\mu_1)$ the probability associated to a minimal threshold $\mu_1 = 0.2$. The sequence μ_j is thus increasing and k_{μ_j} is decreasing. In other words, the method tests $N_\mu = 10$ different thresholds starting from $\mu_1 = 0.2$. This results in a non-parametric and unsupervised approach.

Note that the value of $NFA_i(R_i)$ constitutes a measure of the belief that a point belongs to the background (or conversely, to moving objects). As explained in Veit et al. (2006) one can apply a detection test specified as follows: accept the motion hypothesis for region R_i if $NFA_i(R_i) < 1$ and reject it otherwise. This results in less than one false detection on average. In Fig. 3 this rule was applied for computing an initial motion detection map. Note that the discriminative term alone is not able to correctly detect the moving regions, nonetheless providing valuable information. We then set

$$\alpha_i^D(\mathbf{I}_{t+1}) = -\log NFA(R_i) \tag{18}$$

where a low value of $\log NFA_i$ favors $x_i^t = l$.

Our method does not rely only on the comparison between the current image and the reference image but explicitly introduces (normal flow) motion measurements as explained above. The overall scheme gains accuracy and completeness, integrating this low-level feature in the decision process.

Reconstruction Terms We elaborate now the potential $V_i^R(x_i^t \mid \mathbf{I}_{t+1}, \mathbf{z}_{t-1})$ in Table 1b. On one side, it aims at estimating the intensity values of the background (reference) image, taking into account their interactions with the symbolic values. On the other side, it exploits the information of the intensity difference between the current image and the reconstructed reference image, which provides the basis for the decision process in a background subtraction method.

The first term of Table 1b,

$$\mathbf{1}_l^*(x_i^t)\frac{[x_i^t - m(z_i^{t-1}, I_i(t))]^2}{\sigma_i^2}$$

favors that, when there is no motion, i.e. $\mathbf{1}_l^*(x_i^t) = 1$, the estimated intensity value for a point is close to the previous

estimated reference intensity value. Simultaneously, it penalizes the absence of motion if this difference is eventually large.[2] Both types of values interact consequently, in order to minimize the energy. Note that this term also performs a temporal smoothing 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 over a 9×9 window centered at location i in I_t. The second term of Table 1b,

$$\mathbf{1}_l(x_i^t)\alpha_i^R(I(t), \mathbf{z}_{t-1})$$

results 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 intensity differences is introduced in order to reduce the effect of the observation noise. The parameter γ controls the influence of the reconstruction term in the total energy.

Regularization Terms The potentials introduced so far are 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 preserving edges and contrast of the image.

A combined spatial regularization of both types of values is achieved through the energy potential in Tab. 1c. First, a Gaussian term,

$$\frac{\beta^c}{g_i(\nabla I(t))}\mathbf{1}_l^*(x_i^t)\mathbf{1}_l^*(x_j^t)\left[\frac{(x_i^t - x_j^t)^2 - K}{\sigma_i^2}\right]$$

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,[3] 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 to 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 Table 1c, in order to improve regularization and reduce false negative detections. The parameters involved in Table 1 are set to achieve a correct regularization in both the motion detection map and the background intensity values. Their influence is analyzed in Sect. 5.4.

[2] We set $m(z_i^{t-1}, I_i(t)) = I_i(t)$ if we do not have an available previously estimated value for the reference image at that point.

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

4.5 Estimation

The problem reduces to the task of estimating the field $\mathbf{x}_t$ by minimizing $Q(\mathbf{x}_t \mid \mathbf{I}_{t+1}, \mathbf{z}_{t-1})$. The ICM (Iterated Conditioned Modes) algorithm (Besag 1974) is used for this task which is an iterative procedure for maximizing $p(\mathbf{x}_t \mid \mathbf{I}_{t+1}, \mathbf{z}_{t-1})$. By choosing the value of x_i^t at site i that maximizes the conditional probability $p(x_i^t \mid \mathbf{x}_{t,\mathcal{N}_i}, \mathbf{I}_{t+1}, \mathbf{z}_{t-1})$, it results that $p(\mathbf{x}_t \mid \mathbf{I}_{t+1}, \mathbf{z}_{t-1})$ increases (Guyon 1995). Passing by each point a sufficient number of times, an optimal solution is obtained. Then, we only have to compute the conditional mixed-state density at each location, which can be derived directly from (14).

Defining $H(x_i^t) = V_i^D(x_i^t \mid \mathbf{I}_{t+1}) + V_i^R(x_i^t \mid \mathbf{I}_{t+1}, \mathbf{z}_{t-1}) + \sum_{j\in\mathcal{N}_i} V_{ij}^S(x_i^t, x_j^t \mid \mathbf{I}_{t+1})$ this conditional density is given by:

$$p(x_i^t \mid \mathbf{x}_{t,\mathcal{N}_i}, \mathbf{I}_{t+1}, \mathbf{z}_{t-1}) = \frac{\exp -H(x_i^t)}{Z_i}, \tag{19}$$

where Z_i is a normalization factor that does not depend on x_i^t. Then, for each point the following rule is applied:

$$x_i^t = \begin{cases} l & \text{if } H(x_i^t = l) < H(x_i^t = x_i^*), \\ x_i^* & \text{otherwise,} \end{cases} \tag{20}$$

where x_i^* is the continuous value that maximizes the continuous part of (19), i.e. when $x \neq l$:

$$x_i^* = \frac{\frac{\beta^c}{g_i(\nabla I(t))}\sum_{j\in\mathcal{N}_i} x_j^t \mathbf{1}_l^*(x_j^t) + \gamma m(z_i^{t-1}, I_i(t))}{\frac{\beta^c}{g_i(\nabla I(t))}\sum_{j\in\mathcal{N}_i} \mathbf{1}_l^*(x_j^t) + \gamma}. \tag{21}$$

Note that when $x_i^t \neq l$, the conditional distribution of x_i^t given its neighbors is Gaussian as one can infer from the quadratic terms in $V_{ij}^S(x_i^t, x_j^t \mid \mathbf{I}_{t+1})$ and $V_i^R(x_i^t \mid \mathbf{I}_{t+1}, \mathbf{z}_{t-1})$. Thus the maximizing value x_i^* coincides with the mean of this conditional continuous density, and is the estimated value for the reference image at point i.

4.6 Discussion

As said before, the proposed mixed-state formulation introduces a new way of dealing with simultaneous decision-estimation problems and provides key advantages with respect to more conventional approaches.

Classical robust estimation methods could also be applied for estimating the background image, while considering foreground moving points as outliers. It is somehow

true that they would permit to reject the outliers while recovering the background at the same time. But this approach only considers that most of the time, the background is not occluded. In essence, the nature of the outlier is not taken into account: it is just a point that does not respond to an assumed model as, for example, also a noisy intensity value can be. By nature, robust approaches are not able to detect the moving points as belonging to a particular class.

If we want to introduce specific information for what we want to detect (the class of moving points), a CRF has shown to be a useful discriminative model. The presence (not the intensity) of a moving point is explicitly modeled using (spatio-temporal) motion and intensity information. Consequently, there is no assumption of a predominant model, but a joint process of detection and estimation. The result is that there is no need to estimate a model assuming that the number of outliers is small. If a point is in motion, the discriminative terms of our approach will detect it "just because it moves", not as a rejected outlier.

Other authors apart from us have successfully applied conditional random fields for background/foreground segmentation (and many other problems) as the method by Criminisi et al. (2006), here tested and compared to our method. But CRFs used before were applied for solving a decision (discrete) problem alone or an estimation problem alone, but separately. We thus have adopted a CRF with its strengths, and combined it with a mixed-state field. Now we were able to solve both problems jointly, determining the moving points and the intensity of the non-moving (background) points through the inference of a single random (mixed-state) field.

Summarizing: robust methods can perform a simultaneous labeling and estimation, but are not able to introduce useful information for the detection process; meanwhile, CRFs are able to introduce arbitrary information for a decision or an estimation process, but they had never been defined before within a truly simultaneous scenario. A MS-CRF possesses both attributes.

5 Results and Experimental Comparisons

5.1 Mixed-State Field

For our method, we use the 8-point nearest neighbor set as the neighborhood $\mathcal{N}_i$ for the mixed-state Markov random field. The parameters of the model were set as follows: $\gamma = 8$, $\beta^c = 1$, $\beta_m = 5$ and $c = 0.7$. For all the sequences these same values were used. This is justified observing (21). 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. A complete analysis of the parameter values is left for Sect. 5.4.

Let us first present the result of applying our method to the sequences Parking and Tennis as depicted in Fig. 4. In the figure we observe the process of joint motion detection and background reconstruction at different frames. These examples illustrate how the algorithm works. Figures 4b and 4e contain the estimated mixed-state fields where for some points the mixed-state variable x_i takes the symbolic value (red in the figure) indicating a detected moving point and for the rest, an intensity value is assigned as the background intensity estimate. This is the single output of the estimation process in (20). The background update rule (13) is then applied to recover the reference image at those points where motion is absent. Observe in Fig. 4c how the moving car is detected and the background is gradually reconstructed. At the first frame, on the region where the car is detected, there is no information of the background image. This is shown as a black hole in the estimated image.

5.2 Focus on the Motion Detection Performance and Comparisons

We have applied our motion detection method to real sequences consisting of rigid and articulated motion. We compare the results with the standard methods of Stauffer and Grimson (2000) and Elgammal et al. (2000). We also consider two more recent methods. The one by Zivkovic and van der Heijden (2006) which exploits the unsupervised learning method introduced in Zivkovic and van der Heijden (2004). The other by Criminisi et al. (2006) employs, as in our approach, a conditional random field which includes a temporal persistency model of the labels and a contrast-dependent regularization term. However, the temporal model has to be learned from ground truth data for the processed sequences and the background model is learned adaptively using color histograms by processing an initial extended observation of the background. Then, its distribution is static over time.

Additionally, we compare the performance of the full mixed-state model, with two sequential implementations based on non-mixed versions of the proposed energy potentials (Algorithms 1 and 2), in order to show the importance of the mixed-state terms and the simultaneous approach. In both latter cases, the first step is to estimate the moving points and then, with a fixed detection map, the background

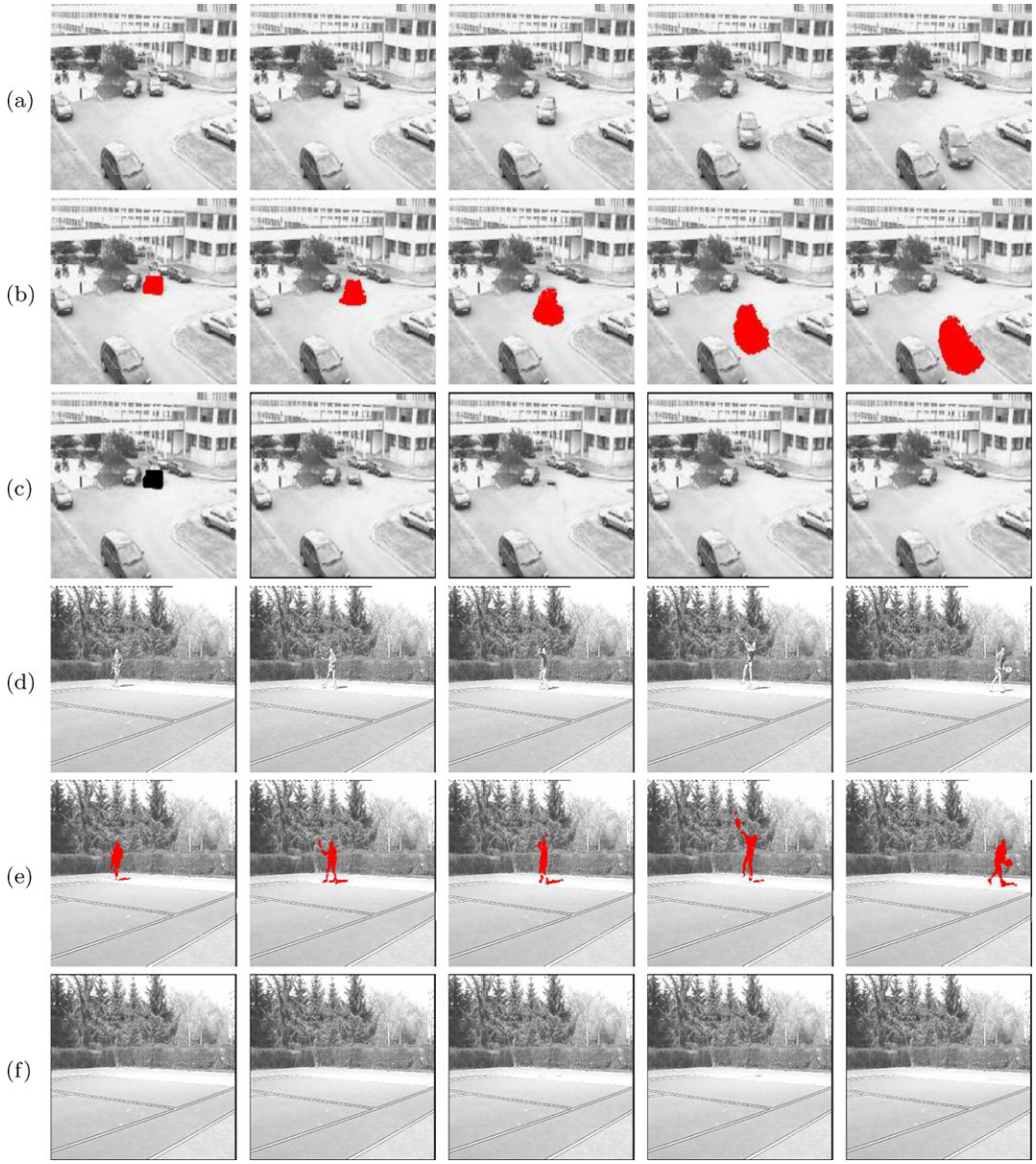


Fig. 4 (Color online) Simultaneous motion detection and background reconstruction with our MS-CRF method. (**a**) Frames 2, 22, 42, 62, 72 of the Parking sequence. (**b**) Mixed-state field estimated for each frame. *Red* indicates a detected moving point ($x_i^t = l$). (**c**) Background reconstruction process (image $\{z_i^t\}$). As the sequence advances, reconstruction of the non-moving regions is performed to obtain the complete background image. Note how the car virtually disappears. (**d**) Frames 20, 30, 40, 46, 77 of the Tennis sequence. (**e**) Mixed-state field and (**f**) reconstructed background. The player is replaced along the sequence by the reference image estimates though the background is never completely uncovered

is reconstructed and updated. In Algorithm 1 we have only left the unary and purely discrete terms, not including any type of spatial regularization. In Algorithm 2, we add the spatial regularization terms for the discrete states, and for the background reconstruction as well. In other words, we take out the mixed potentials from the energy.

Next, we compare the motion detection performance of each of the six methods considered here (Stauffer-Grimson, Elgammal et al., Zivkovic, Criminisi et al., Seq1, Seq2 and MS-CRF) and display the values for: Precision, Recall and the so-called F-score. The latter is computed as the harmonic mean of precision and recall, and is a global measure of the method accuracy. These quantities are defined as

$$\begin{aligned} \mathit{precision} &= \frac{\#\mathit{true\ positives}}{\#\mathit{true\ positives} + \#\mathit{false\ positives}}, \\ \mathit{recall} &= \frac{\#\mathit{true\ positives}}{\#\mathit{true\ positives} + \#\mathit{false\ negatives}}, \\ \text{F-}\mathit{score} &= 2 \cdot \frac{\mathit{precision} \cdot \mathit{recall}}{\mathit{precision} + \mathit{recall}}. \end{aligned} \tag{22}$$

Algorithm 1 Sequential without spatial regularization (Seq1)

for each t **do**
  minimize w.r.t. all $w_i \in \{0, 1\}$

$$\sum_i \alpha_i^D (1 - w_i) + \gamma w_i \alpha_i^R$$

  for each i **do**
    if $w_i = 0$ **then**
      $z_i^t \leftarrow m(z_i^{t-1}, I_i(t))$
    else
      $z_i^t \leftarrow z_i^{t-1}$
    end if
  end for
end for

Algorithm 2 Sequential with spatial regularization (Seq2)

for each t **do**
  minimize w.r.t. all $w_i \in \{0, 1\}$

$$\sum_i \alpha_i^D (1 - w_i) + \gamma w_i \alpha_i^R - \sum_{i,j} \frac{\beta^m}{g_i(\nabla I(t))} w_i w_j$$

  for each i **do**
    if $w_i = 0$ **then**
      $z_i^t \leftarrow (21)$
    else
      $z_i^t \leftarrow z_i^{t-1}$
    end if
  end for
end for

They were computed with respect to the ground-truth detection map, which we have determined by manual segmentation of the video sequences. We have tested the video sequences Basket, Forest, Tennis, Van and Traffic Circle.

Basketball Sequence In Fig. 5 we present the results for the Basketball sequence. The method by Stauffer and Grimson (Fig. 5b) yields wrongly detected moving points in the background. The method by Elgammal et al. (Fig. 5c) performs better, but has some problems to correctly recover connected regions. The result applying the method of Zivkovic (Fig. 5d) shows less false positives but the segmentation is not that smooth. The approach of Criminisi et al. (Fig. 5e) yields a good segmentation, though it seems oversmoothed. It is important to point out that for these last methods, there are images available without moving objects for estimating the background model. Finally, the mixed-state method (Fig. 5h) shows an improved regularization of the motion map, reducing false positives and false negatives, also compared with the sequential non-mixed versions of the algorithm (Figs. 5f and 5g). In the comparative table given in Fig. 5a) we observe that the methods by Stauffer-Grimson, Elgammal et al. and Zivcovic show a better Precision but at a cost of numerous false negatives. This is reflected in the Recall rate which is poor compared with MS-CRF, Seq1 and Seq2. At the same time, MS-CRF shows less false positives than Seq1 and Seq2, with a similar Recall value. The method by Criminisi et al. also shows a high F-score due to a high Recall, but diminished by a lower Precision. Overall, our method has the best F-score.

Forest Sequence The Forest sequence (Fig. 6) depicts a complex scene of two men walking through the woods. In this example the background is not completely static as there is swaying vegetation. Our method (Fig. 6h) supplies very good results discarding practically all the background motion, even compared with multi-modal density models (Figs. 6b and 6c). The proposed motion-based measures $NFA(R_i)$ (17) introduced in the discriminative term are in theory able to cope with this kind of background dynamics. However, by themselves they generate many false detections as shown in Fig. 3 in the case of the Forest sequence. Embedding these observations in the mixed-state conditional random field notably improves the overall motion detection.

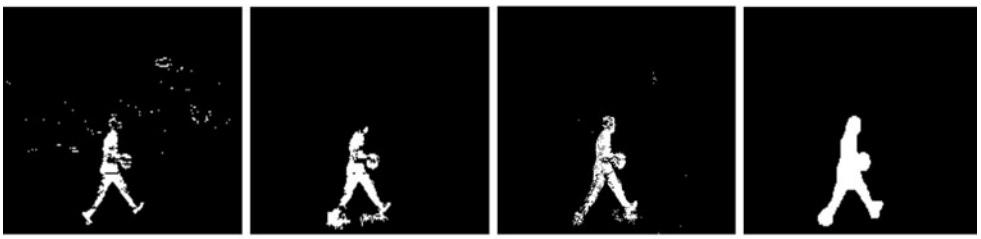
The performance of MS-CRF is clearly better in Precision (Fig. 6a) w.r.t. the other methods. This is a consequence of a large reduction of false positives, as one can confirm visually. Seq1 and Seq2 show a high Recall, but are not able to correctly segment the two men from the background. Criminisi et al. again oversmoothes the detection map but performs very well giving the best F-score.

Van Sequence In the Van sequence (Fig. 7), the video is shot on a rainy day and thereby the background contains again some variation. In this case, the variation is more uniform and weaker than for the Forest sequence, so that the methods by Elgammal et al. (Fig. 7c) and Stauffer-Grimson (Fig. 7b) gave satisfactory results in this sense. However, our method (Fig. 7h) delivers a reduced amount of false negatives (note the windows of the van) and more compact detected moving regions. Meanwhile, the method by Criminisi et al. is not able to achieve a good Precision with many false positives. The algorithms Seq1 and Seq2 show many artifacts around the Van which is also reflected in a low Precision (Fig. 7a).

Tennis Sequence For the Tennis example (Fig. 8), the algorithms Seq1, Seq2 and MS-CRF have shown a similar performance yielding the best results compared to the other methods. This can be observed in both the motion detection map and the values of Precision, Recall and F-score (Fig. 8a). As for Elgammal et al. (Fig. 8c), the background model is wrongly estimated since it includes the player at

Method	Precision (%)	Recall (%)	F-Score (%)
Elgammal et al.	80.7	65.2	72.1
Stauffer-Grimson	83.1	55.3	66.4
Zivcovic	82.5	66.3	73.5
Criminisi et al.	71.2	100	83.1
Seq. 1 (No Reg)	57.4	89.8	70.1
Seq. 2	64.6	88.7	74.8
MS-CRF	**79.2**	**88.8**	**83.7**

a) Performance of each method



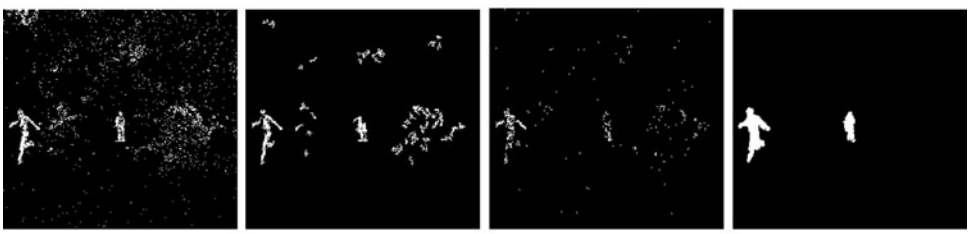
b) Stauffer-Grimson c) Elgammal et al. d) Zivkovic e) Criminisi et al.

f) Seq1 g) Seq2 h) MS-CRF

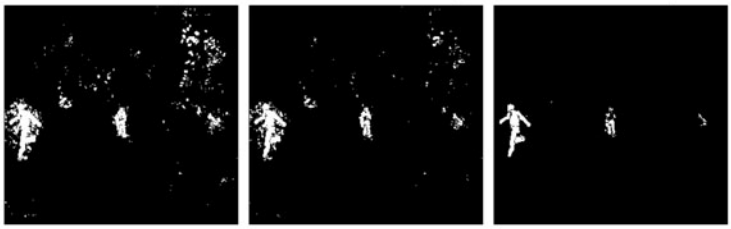
Fig. 5 Basketball sequence: motion detection results for different algorithms compared to our method

Method	Precision (%)	Recall (%)	F-Score (%)
Elgammal et al.	34.7	50.8	41.2
Stauffer-Grimson	22.2	57.6	32.1
Zivcovic	30.1	46.2	36.4
Criminisi et al.	75.9	90.5	82.4
Seq. 1 (No Reg)	35.1	91.6	50.8
Seq. 2	50.1	86.1	63.3
MS-CRF	**85.6**	**72.9**	**78.7**

a) Performance of each method



b) Stauffer-Grimson c) Elgammal et al. d) Zivkovic e) Criminisi et al.

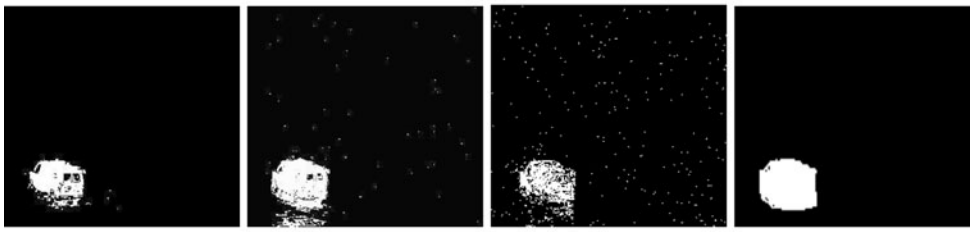


f) Seq1 g) Seq2 h) MS-CRF

Fig. 6 Forest sequence: motion detection results for different algorithms compared to our method

Method	Precision (%)	Recall (%)	F-Score (%)
Elgammal et al.	76.1	85.6	80.5
Stauffer-Grimson	91.0	65.1	75.9
Zivcovic	69.7	64.0	66.7
Criminisi et al.	68.1	96.3	79.8
Seq. 1 (No Reg)	66.1	88.9	75.8
Seq. 2	61.6	91.9	73.7
MS-CRF	**84.5**	**90.2**	**87.3**

a) Performance of each method



b) Stauffer-Grimson c) Elgammal et al. d) Zivkovic e) Criminisi et al.

f) Seq1 g) Seq2 h) MS-CRF

Fig. 7 Van sequence: motion detection results for different algorithms compared to our method

different frames as part of it, resulting in a ghost effect. The method by Stauffer-Grimson (Fig. 8b) gave a satisfactory result but with a lower Recall rate, which is related to its inability of obtaining compact and smooth segments. Notice that the missed detections in the segmentation obtained with the method by Criminisi et al. (Fig. 8e) is a consequence of a behavior observed also in the previous examples. Basically, it is unable to segment small moving structures, as for example the tennis ball, due to an oversmoothing effect.

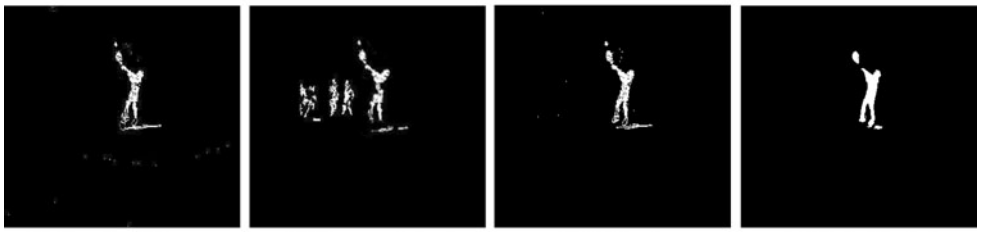
Traffic Circle Sequence Finally, in the Traffic Circle sequence (Fig. 9) we have multiple rigid motions. In this case, a complete background image is never available during the sequence. The cars continuously pass around the square entering and leaving the scene. The method by Stauffer and Grimson (Fig. 9b) is affected by 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 continuously wrong detection for subsequent frames It takes too long for the model to remove them from the reference image. Moreover, some regions of the background are never correctly updated. For the same sequence the non-parametric method of Elgammal et al. (Fig. 9c) failed in generating valid results, yielding absence of motion 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. Also for Criminisi et al. (Fig. 9e) this is a problem as the color likelihoods for the background cannot be learned and thus computed correctly.

For our method (Fig. 9h), 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 256×256), 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 the regions occupied by the moving objects are obtained compactly. Note how most of the cars are indeed detected as uniformly connected regions. From the table in Fig. 9a) we deduce that MS-CRF gave the best F-score, basically due to a notably better Precision. Meanwhile, for Seq1 and Seq2 this value is lower as a consequence of the many artifacts that appear around the car.

Regarding the computation time for processing the tested sequences, the algorithm solves the motion detection and the background reconstruction at a rate of 1 frame/sec on average, for 320×240 gray-scale images. This was obtained with a non-optimized implementation in C++, running on a standard desktop PC.

Method	Precision (%)	Recall (%)	F-Score (%)
Elgammal et al.	39.9	51.2	44.9
Stauffer-Grimson	86.2	66.1	74.9
Zivcovic	88.1	67.9	76.7
Criminisi et al.	89.0	72.6	79.9
Seq. 1 (No Reg)	84.4	84.0	84.2
Seq. 2	86.3	79.9	83.0
MS-CRF	**90.7**	**76.7**	**83.1**

a) Performance of each method



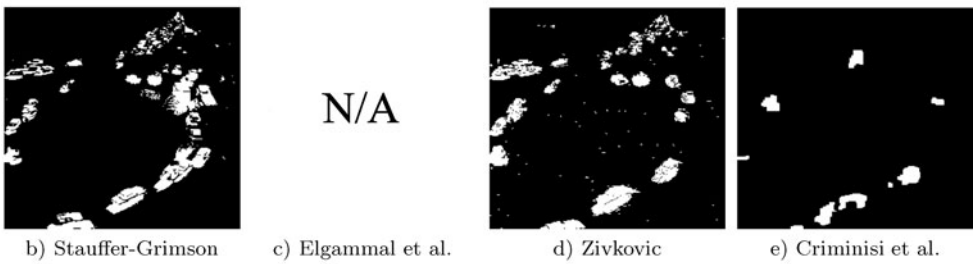
b) Stauffer-Grimson c) Elgammal et al. d) Zivkovic e) Criminisi et al.

f) Seq1 g) Seq2 h) MS-CRF

Fig. 8 Tennis sequence: motion detection results for different algorithms compared to our method

Method	Precision (%)	Recall (%)	F-Score (%)
Elgammal et al.	-	-	-
Stauffer-Grimson	54.8	63.9	59.0
Zivcovic	64.8	85.1	73.6
Criminisi et al.	38.0	26.5	31.2
Seq. 1 (No Reg)	49.2	88.5	63.3
Seq. 2	57.1	86.3	68.7
MS-CRF	**68.9**	**85.2**	**76.2**

a) Performance of each method



N/A

b) Stauffer-Grimson c) Elgammal et al. d) Zivkovic e) Criminisi et al.

f) Seq1 g) Seq2 h) MS-CRF

Fig. 9 Traffic Circle sequence: motion detection results for different algorithms compared to our method. The method by Elgammal et al. (**c**) did not give valid results

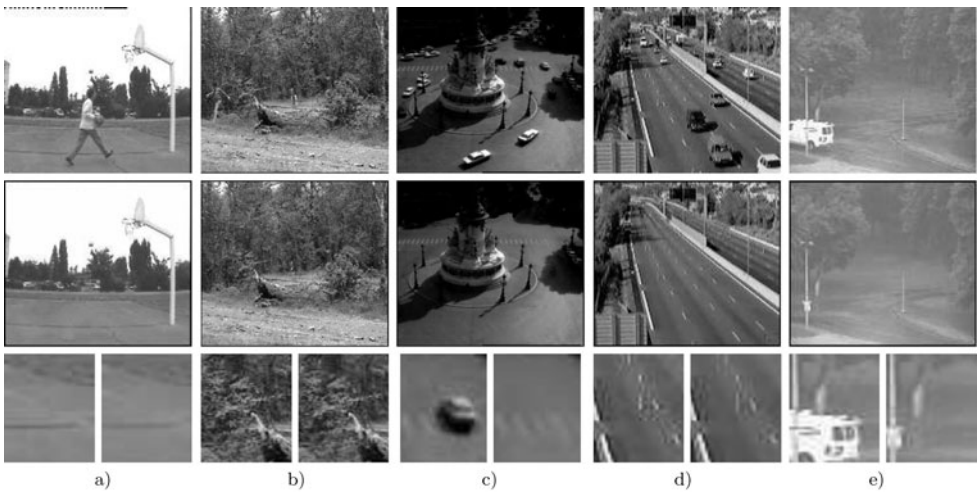


Fig. 10 *Top row*: original sequences. *Center row*: background images estimated with our method. *Bottom row*: a close-up over a small region of the original (*left*) and reconstructed (*right*) images. The spatio-temporal reconstruction of the background is achieved jointly with motion detection, resulting in virtually removing the moving objects from the scene. The reference image is also filtered over homogeneous intensity regions in order to reduce noise, while preserving borders

5.3 Focus on Background Reconstruction

The proposed algorithm generates, at each time instant, estimates of the background image, not a model of it. We have really tackled a problem of reconstruction. The approach uses all the information about the background across time to build a complete image. In this case, moving objects can be removed from the scene as shown in Fig. 10. Moreover, this reconstruction also involves smoothing of the background image, over homogeneous intensity regions, filtering out the observation noise, but preserving the edges. In the third row of Fig. 10 we display a small region for each sample, in order to more clearly observe the effect of the background reconstruction. In Fig. 10a, the basketball court is smoothed, and the lines are well preserved. In the Forest sequence 10b, 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.

Finally, in order to assess the efficiency of the algorithm in obtaining the background, we have computed the percentage of the reference image left to be reconstructed until each frame in the video sequence. We compare the full MS-CRF algorithm with the sequential algorithms Seq1 and Seq2 on the sequences Traffic Circle and Highway (Fig. 11). Observe that for the MS-CRF method it takes less video frames to perform the reconstruction, that is, at a particular instant t it has estimated a larger part of the background. It means that our method for reconstructing the background image can also be viewed as properly addressing the video inpainting issue.

5.4 Experimental Parameter Analysis

The parameters involved in the MS-CRF model were set to fixed values for all the experiments. They were obtained by an experimental analysis of the motion detection and background reconstruction results, which is presented in what follows. One could say that it would be more appropriate to learn or estimate them from ground truth data. However, our methodology permits us to sweep a range of values and observe the performance of the method in order to establish how sensitive it is to their values.

The first parameter we analyze is the size of the regions R_i where $NFA(R_i)$ is computed in (17). This determines the motion likelihood for the discriminative term (Table 1a) in our mixed-state energy function. As in Fig. 3, we can threshold $\log NFA_i$ computed using different region sizes to obtain the detection maps depicted in Fig. 12a. Note that this is done only for visualization in order to clearly distinguish where the motion likelihood is high or low, but it is not the result of the MS-CRF method. For a small region size as 4×4, we can see that the discriminative term is not sufficiently reliable, giving a low likelihood to a big proportion of the moving region. Consequently, the final result of the MS-CRF method in Fig. 12b is poor. As the region size increases, the discriminative term overestimates the moving regions but thanks to the reconstruction and smoothing terms (Table 1a–b), the result of the detection improves notably. On the other side, taking bigger regions R_i implies that the learning of the background is slower (Fig. 12c) and the initial detection performance is lower (Fig. 12d). As the background is learned, the F-score values grow up to a steady state. Indeed, at the beginning of the sequence the detection relies mostly on the motion likelihoods, as there is no background information. Of course, for regions of size 4×4, though the learning process is faster, the background is wrongly estimated. At the same time, if the regions are too big, this affects the detection precision as we can see in the F-score curves (Fig. 12d). A value of 20×20 have shown to be the best choice, and it was applied to all the tested sequences. Note that between 12×12 and 32×32 the performance does not vary drastically.

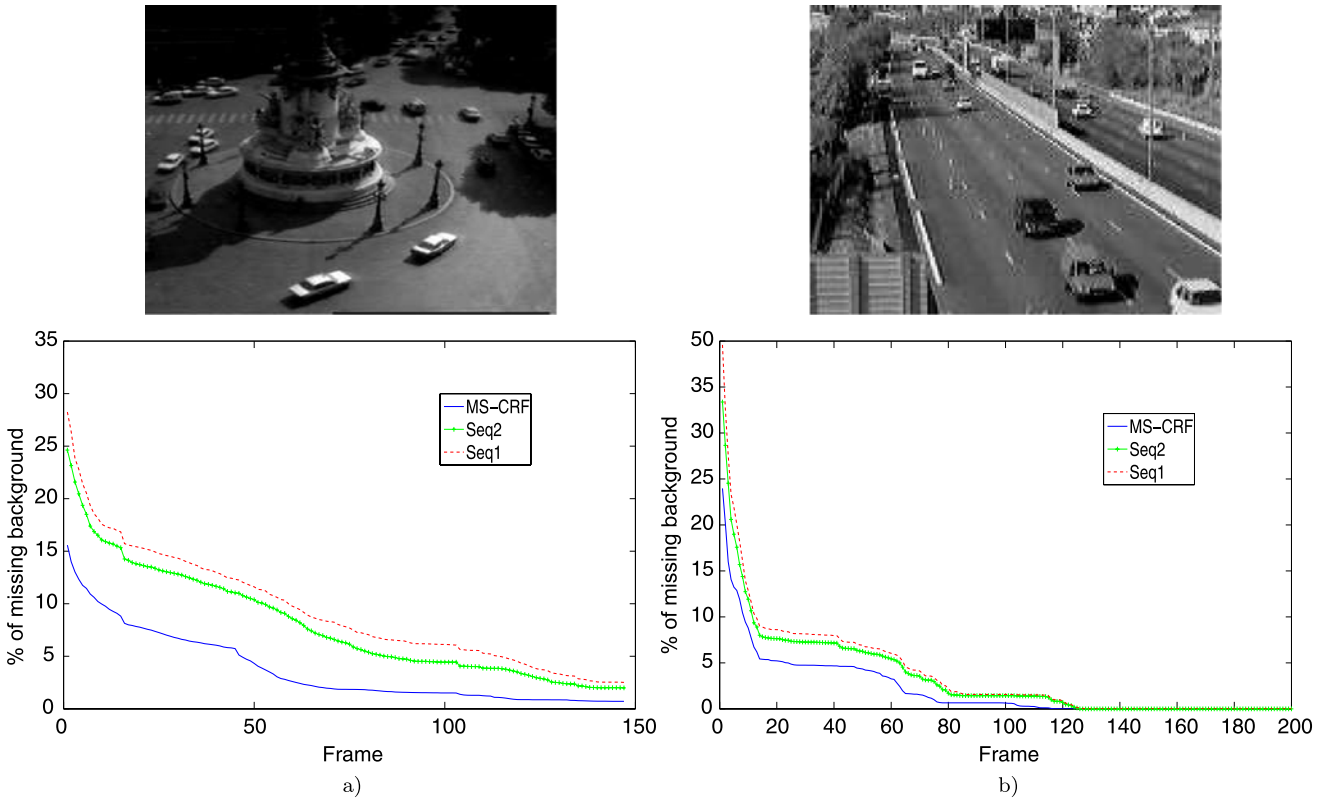


Fig. 11 The plots show the percentage of the background image that remains to be reconstructed for the algorithms MS-CRF, Seq1 and Seq2 and for the sequences (**a**) Traffic Circle and (**b**) Highway. The values represent the proportion of the reference image that each algorithm was not able to reconstruct during the elapsed time

Next, we analyze the effect of the parameters involved in the reconstruction and smoothing terms (Table 1a–b), that is, γ, β^c and β^m. The values chosen for all the processed sequences in the previous sections were $\gamma = 8$, $\beta^c = 1$ and $\beta_m = 5$. As mentioned in Sect. 5, this sets an equilibrium working point for the algorithm.

In order to observe the sensitivity of the method to variations of the parameters, we have swept their values around the working point. We have tested $\gamma \in \{2, 6, 10, 14\}$, $\beta^c \in \{0.2, 0.6, 1, 1.4\}$ and $\beta^m \in \{2, 4, 6, 8\}$. The results are shown in Fig. 13. We have chosen the Forest sequence given its complexity due to the presence of a noisy and highly dynamic background. This will permit us to have a better view of the performance variations.

In the first row we observe the effect of a varying γ. This parameter weights the reconstruction potential. Thus, for low values the information given by the reconstructed background image is underestimated and the motion likelihood governs the energy. As a result the detection is incorrect, similarly to what is shown in Fig. 3. With increasing γ the improvement is clear. However, taking a high value may mask the effect of the smoothing terms and one obtains a noisy detection map, as seen in the last image of the first row.

In the second row we vary β^c, related to the mixed-state term, and which affects the joint spatial regularization of the background estimates and the non-moving regions. Increasing this value permits to obtain more compact regions. However, if this value is too big, the number of false negatives may also increase as observed in the last figure of the row.

Finally, in the third row we observe the effect of β^m. This parameters is involved in the regularization of the motion detection map and, as we can see in the figures, a high value gives more compact regions at the risk of an increased number of false positives.

This same behavior, exemplified here for the Forest sequence, was observed in all the cases. As said before, we have obtained good results for $\gamma = 8$, $\beta^c = 1$ and $\beta_m = 5$ and in general, the performance did not decreased considerably for $\gamma \in [8, 12]$, $\beta^c \in [0.6, 1.2]$ and $\beta^m \in [4, 6.5]$.

5.5 Discussion

The ability of incorporating arbitrary information (here spatio-temporal motion and intensity information) is a general characteristic of conditional random fields, not only of mixed-state models. We could have chosen different energy terms or used CRF terms designed in other approaches, embedding them into our decision-estimation framework. What

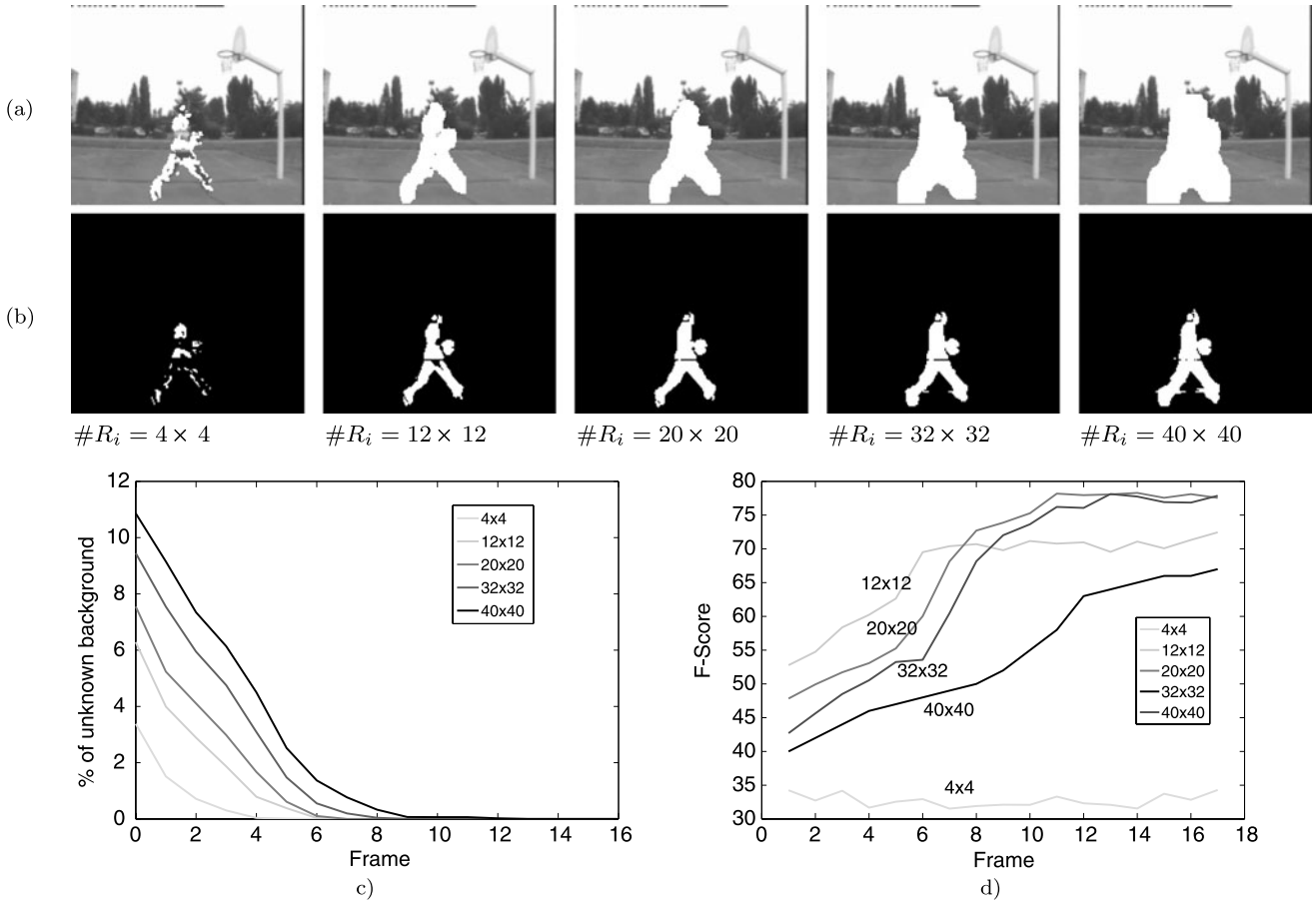


Fig. 12 Detection and reconstruction performance as a function of the region size in (17). (**a**) Motion map obtained by thresholding $\log NFA_i$. (**b**) Result of the MS-CRF method for different region sizes. (**c**) Percentage of the background image remaining to be learned as a function of the frame number. (**d**) F-score for the motion detection result as a function of the frame number

we claim is that both the discrete (discriminative) part of the problem, and the estimation part of the problem, can better be solved simultaneously (besides the particular energy design) inferring a mixed-state field. And for showing this, we have considered sequential versions of the method (algorithms Seq1 and Seq2) that apply a sequential decision-estimation strategy. Note that the difference between the mixed-state method and the sequential implementations is not the design of the energy terms, but the assumption of a mixed-state field which is inferred at a single step instead of solving two problems in two steps (estimation after decision). Other previous related methods assume the background is known (normally from a training stage, then adapted in time) and then they solve a discrete field for the problem of foreground segmentation (see Criminisi's CRFs, Criminisi et al. 2006, Grimson's parametric models, Stauffer and Grimson 2000, Elgammal's non-parametric models, Elgammal et al. 2000, and Zivcovic's improved learning strategy, Zivkovic and van der Heijden 2006). We have provided several experimental results, showing that the truly simultaneous mixed-state performs better, both for motion detection and background reconstruction.

6 Conclusion

We have proposed a simultaneous motion detection and background reconstruction method using a mixed-state conditional random field. The algorithm outperforms state-of-the-art motion detection methods, as confirmed by the experiments. As well, it improves the performance compared with algorithms that follow a sequential strategy, both for the motion detection map and the reconstructed background.

We have adopted the conditional random field framework given its flexibility and its demonstrated good performance against other statistical methods. We have combined CRFs with mixed-state fields as a new way of investigating simultaneous decision-estimation problems. A formal demonstration of the optimality properties of the mixed-state approach will be part of a more theoretical study. It is not within the scope of this work.

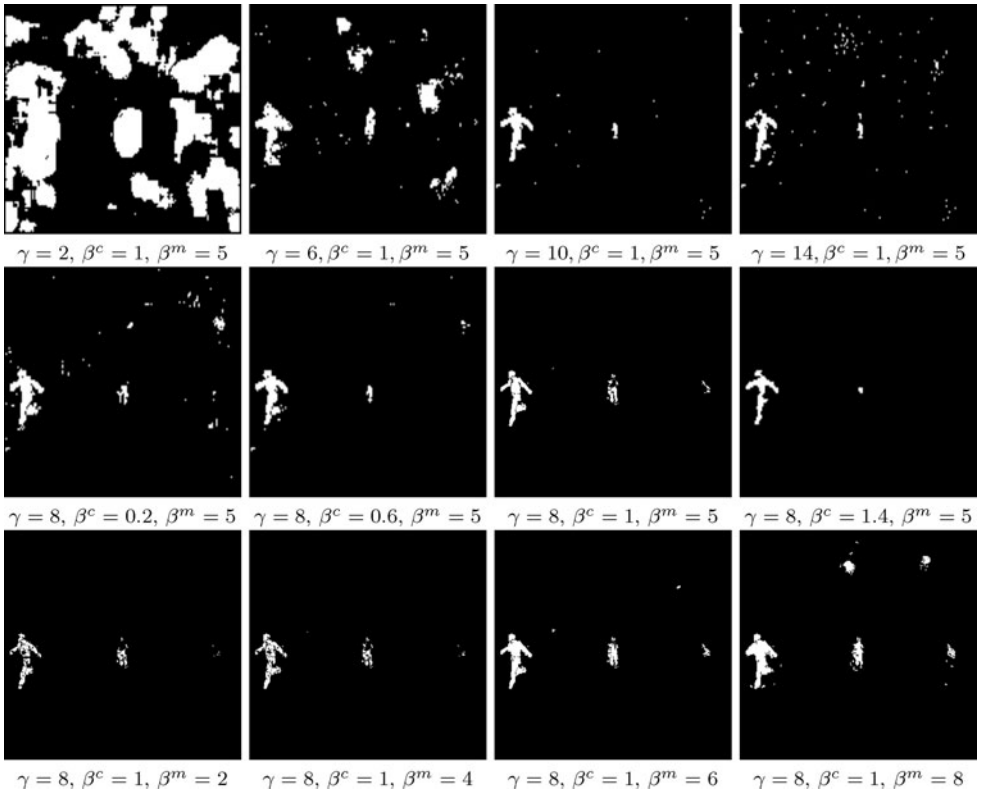


Fig. 13 Motion detection maps obtained with varying γ (*top row*), β^c (*middle row*) and β^m (*bottom row*)

It is worthy to say that the parameters involved in the energy terms were set empirically, in order to obtain a correct motion detection and background estimation. The values were the same for all the experiments, though it is fair to emphasize the necessity of studying the problem of online optimal parameter estimation, making the method fully unsupervised. This will be studied in a future work.

In summary the method has the following characteristics:

- *Reduction of false positive and false negatives* Through a more complex regularization of the motion 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, allowing us to exploit the local information of the intensity difference between the true background and a foreground moving object.
- *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 in the regions 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.

References

Benboudjema, D., & Pieczynski, W. (2007). Unsupervised statistical segmentation of nonstationary images using triplet Markov fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*, 1367–1378.

Benedek, C., Sziranyi, T., Kato, Z., & Zerubia, J. (2007). A multi-layer Mrf model for object-motion detection in unregistered airborne image-pairs. In *IEEE international conference on image processing, 2007 (ICIP07)*, 16 2007–Oct. 19 2007 (Vol. 6, pp. 141–144).

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, *36*, 192–236.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B*, *48*(3), 259–302.

Black, M. J., & Rangarajan, A. (1996). On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, *19*(1), 57–91.

Blanchet, J., & Forbes, F. (2008). Triplet Markov fields for the classification of complex structure data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(6), 1055–1067.

Bouthemy, P., & Lalande, P. (1993). Recovery of moving object masks in an image sequence using local spatiotemporal contextual information. *Optical Engineering*, *32*(6), 1205–1212.

Bouthemy, P., Hardouin, Ch., Piriou, G., & Yao, J.-F. (2006). Mixed-state auto-models and motion texture modeling. *Journal of Mathematical Imaging and Vision*, *25*(3), 387–402.

Bugeau, A., & Pérez, P. (2007). Detection and segmentation of moving objects in highly dynamic scenes. In *CVPR '07: proc. of the 2007 IEEE conf. on computer vision and pattern recognition*, Minneapolis, MI.

Caillol, H., Hillion, A., & Pieczynski, W. (1993). Fuzzy random fields and unsupervised image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, *31*, 801–810.

Carincotte, C., Derrode, S., Sicot, G., & Boucher, J. M. (2004). Unsupervised image segmentation based on a new fuzzy hmc model. In *ICASSP'04* (pp. 17–21).

Carincotte, C., Derrode, S., & Bourennane, S. (2006). Unsupervised change detection on sar images using fuzzy hidden Markov chains. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(2), 432–441.

Cernuschi-Frias, B. (2007). Mixed states Markov random fields with symbolic labels and multidimensional real values. Technical Report 6255, INRIA, July 2007.

Chellappa, R. (1985). Two-dimensional discrete Gaussian Markov random field models for image processing. In *Progress in pattern recognition 2* (Vol. 85, pp. 79–112).

Chen, J., & Tang, C. (2007). Spatio-temporal Markov random field for video denoising. In *Proc. IEEE conf. on comp. vision and pattern recognition (CVPR'07)*, June (pp. 1–8).

Collet, C., & Murtagh, F. (2004). Segmentation based on a hierarchical Markov model. *Pattern Recognition*, *37*(12), 2337–2347.

Criminisi, A., Cross, G., Blake, A., & Kolmogorov, V. (2006). Bilayer segmentation of live video. In *CVPR '06: proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 53–60). Washington: IEEE Computer Society.

Crivelli, T., Cernuschi-Frias, B., Bouthemy, P., & Yao, J.-F. (2006). Mixed-state Markov random fields for motion texture modeling and segmentation. In *Proc. IEEE int. conf. on image processing (ICIP'06)*, Atlanta, USA (pp. 1857–1860).

Crivelli, T., Piriou, G., Bouthemy, P., Cernuschi-Frías, B., & Yao, J.-F. (2008). Simultaneous motion detection and background reconstruction with a mixed-state conditional Markov random field. In *ECCV '08: proceedings of the 10th European conference on computer vision*, Marseille, France (pp. 113–126).

Crivelli, T., Bouthemy, P., Cernuschi-Frías, B., & Yao, J.-F. (2009). Learning mixed-state Markov models for statistical motion texture tracking. In *MLVMA'09: 2nd IEEE int. workshop on machine learning for vision-based motion analysis*, Kyoto, Japan.

Elfadel, I., & Picard, R. (1994). Gibbs random fields, cooccurrences, and texture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(1), 24–37.

Elgammal, A. M., Harwood, D., & Davis, L. S. (2000). Non-parametric model for background subtraction. In *ECCV '00: proc. of the 6th European conf. on comp. vision-part II*, London, UK (pp. 751–767).

Fablet, R., & Bouthemy, P. (2003). Motion recognition using nonparametric image motion models estimated from temporal and multiscale co-occurrence statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(12), 1619–1624.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Guyon, X. (1995). *Random fields on a network: modeling, statistics and applications*. New York: Springer.

Hardouin, C., & Yao, J. (2008). Multi-parameter auto-models and their applications. *Biometrika*, *95*(2), 335–349.

Heitz, F., & Bouthemy, P. (1993). Multimodal estimation of discontinuous optical flow using Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(12), 1217–1232.

Jojic, N., & Frey, B. J. (2001). Learning flexible sprites in video layers. In *IEEE int. conference on computer vision and pattern recognition 2001* (pp. 199–206).

Kasetkasem, T., & Varshney, P. K. (2002). An image change detection algorithm based on Markov random field models. *IEEE Transactions on Geoscience and Remote Sensing*, *40*(8), 1815–1823.

Ko, T., Soatto, S., & Estrin, D. (2008). Background subtraction on distributions. In *ECCV08* (pp. 276–289).

Koller, D., Lerner, U., & Angelov, D. (1999). A general algorithm for approximate inference and its application to hybrid Bayes nets. In *Proc. of the fifteenth conference on uncertainty in artificial intelligence*, Stockholm, Sweden (pp. 324–333).

Kumar, S., & Hebert, M. (2006). Discriminative random fields. *International Journal of Computer Vision*, *68*(2), 179–201.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th international conf. on machine learning*, Williamstown, MA, USA (pp. 282–289).

Li, Y., & Huttenlocher, D. P. (2008). Learning for optical flow using stochastic optimization. In *ECCV '08: proceedings of the 10th European conference on computer vision*, Marseille, France (pp. 379–391).

Lorette, A., Descombes, X., & Zerubia, J. (2000). Texture analysis through a Markovian modelling and fuzzy classification: Application to urban area extraction from satellite images. *International Journal of Computer Vision*, *36*(3), 221–236.

Lu, W. L., Murphy, K. P., Little, J. J., Sheffer, A., & Fu, H. B. (2009). A hybrid conditional random field for estimating the underlying ground surface from airborne lidar data. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(8), 2913–2922.

Migdal, J., & Grimson, W. E. (2005). Background subtraction using Markov thresholds. In *WACV-MOTION '05: proceedings of the IEEE workshop on motion and video computing* (pp. 58–65). Washington: IEEE Computer Society.

Mittal, A., & Paragios, N. (2004). Motion-based background subtraction using adaptive kernel density estimation. In *CVPR '04: proc. of the 2004 IEEE conf. on computer vision and pattern recognition* (Vol. 2, pp. 302–309).

Monnet, A., Mittal, A., Paragios, N., & Visvanathan, R. (2003). Background modeling and subtraction of dynamic scenes. In *Proc. of the ninth IEEE int. conf. on computer vision* (Vol. 2, pp. 1305–1312).

Murphy, K. P. (1999). A variational approximation for Bayesian networks with discrete and continuous latent variables. In *Uncertainty in artificial intelligence* (Vol. 15, pp. 457–466). San Mateo: Morgan Kaufmann.

Parag, T., Elgammal, A., & Mittal, A. (2006). A framework for feature selection for background subtraction. In *CVPR '06: proc. of the IEEE conf. on computer vision and pattern recognition*, Washington, DC, USA (pp. 1916–1923).

Pieczynski, W., & Tebbache, A. (2000). Pairwise Markov random fields and segmentation of textured images. In *Machine graphics and vision* (pp. 705–718).

Salzenstein, F., & Collet, C. (2006). Fuzzy Markov random fields versus chains for multispectral image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(11), 1753–1767.

Salzenstein, F., & Pieczynski, W. (1997). Parameter estimation in hidden fuzzy Markov random fields and image segmentation. *Graphical Models and Image Processing*, *59*(4), 205–220.

Sheikh, Y., & Shah, M. (2005). Bayesian modeling of dynamic scenes for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(11), 1778–1792.

Stauffer, C., & Grimson, W. E. L. (2000). Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 747–757.

Sun, J., Zhang, W., Tang, X., & Shum, H. (2006). Background cut. In *Proc. European conf. comp. vision, ECCV 2006* (pp. 628–641).

Veit, Th., Cao, F., & Bouthemy, P. (2006). An a contrario decision framework for region-based motion detection. *International Journal of Computer Vision*, *68*(2), 163–178.

Wang, T., Li, J., Diao, Q., Hu, W., Zhang, Y., & Dulong, C. (2006). Semantic event detection using conditional random fields. In *CVPRW '06: proceedings of the conference on computer vision and pattern recognition workshop*, Washington, DC, USA.

Wren, C. R., Azarbayejani, A., Darrell, T., & Pentland, A. P. (1997). Pfinder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 780–785.

Wright, J., Ganesh, A., Rao, S., & Ma, Y. (2009). Robust principal component analysis: exact recovery of corrupted low-rank matrices. In *NIPS 2009*. 0905.0233.

Wu, J., & Chung, A. C. S. (2007). A segmentation model using compound Markov random fields based on a boundary model. *IEEE Transactions on Image Processing*, *16*(1), 241–252.

Zivkovic, Z., & van der Heijden, F. (2004). Recursive unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(5), 651–656.

Zivkovic, Z., & van der Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, *27*(7), 773–780.