BAYESIAN VISUAL SURVEILLANCE: A MODEL FOR DETECTING AND TRACKING A VARIABLE NUMBER OF MOVING OBJECTS

Carlos R. del-Blanco, Fernando Jaureguizar and Narciso García

Grupo de Tratamiento de Imágenes, Universidad Politécnica de Madrid, Madrid, 28040, Spain

ABSTRACT

An automatic detection and tracking framework for visual surveillance is proposed, which is able to handle a variable number of moving objects. Video object detectors generate an unordered set of noisy, false, missing, split, and merged measurements that make extremely complex the tracking task. Especially challenging are split detections (one object is split into several measurements) and merged detections (several objects are merged into one detection). Few approaches address this problem directly, and the existing ones use heuristics methods, or assume a known number of objects, or are not suitable for on-line applications. In this paper, a Bayesian Visual Surveillance Model is proposed that is able to manage undesirable measurements. Particularly, split and merged measurements are explicitly modeled by stochastic processes. Inference is accurately performed through a particle filtering approach that combines ancestral and MCMC sampling. Experimental results have shown a high performance of the proposed approach in real situations.

Index Terms— Split detections, merged detections, moving regions, multiple object tracking, variable number of objects.

1. INTRODUCTION

Automatic detection and tracking of a variable number of objects is the core of any surveillance system. Object detectors produce a set of unordered noisy measurements that need to be appropriately associated to the existing tracked objects to satisfactorily estimate their trajectories. In addition, some of the measurements can be false detections due to the clutter. Also, detectors can fail in detecting some objects (missing measurements). Object trackers have to be able to manage these kind of measurements to robustly estimate both the number of objects and their trajectories. This task is usually carried out by a data association stage that tries to compute the best correspondence between actual measurements and objects.

A wide range of data association techniques have been proposed [1] that can handle noisy, false and missing measurements. The majority of them impose a one-to-one mapping between objects and measurements. This restriction consists in that one measurement can be associated with at most one object and viceversa. This restriction is reasonable for point based measurements, in which an object is considered as having neither physical volume nor resolvable features. Nevertheless, in visual surveillance systems, the measurements are regions that cannot be satisfactorily modeled by a single point. In particular, visual moving object detectors produce moving regions that ideally represents moving objects. However, object occlusions, changes in illumination, varying object appearances, shadows, reflections, and complex backgrounds give rise to split MRs (one object is split into several MRs) and merged MRs (several objects are merged into one MR). Few works that have explicitly addressed this problem. An interesting approach is presented in [2] that augments the set of MRs with virtual MRs to represent possible split and merged events. A similar strategy is followed in [3] where the virtual MRs are derived from the region overlapping between prediction and detection. However, the previous methods do not provide an explicit model for split and merged measurements limiting their performance. In [4], a model for simulating split and merged MRs is introduced, which uses a Markov Chain Monte Carlo method (MCMC) to draw association samples spatial and temporally. Nonetheless, this approach performs a batch processing that uses MRs of several time steps. This fact makes it unsuitable for on-line applications that either cannot delay the detection and tracking results, or have restriction in computational cost. In [5], split and merged MRs are generated by a different MCMC method that draws association samples sequentially. However, this paper assumes a known number of targets, and therefore it can not handle the entrance and exit of objects in the scene.

In this paper, an automatic detection and tracking framework for visual surveillance applications is presented, which is able to detect and track a variable number of moving objects in complex situations. The main contribution is the Bayesian modeling of the detection and tracking tasks that takes into account specific problems that arise in video based systems. These problems are mainly related to the split, merged, noisy, false and missing MRs that generate a real video object detector. The developed Bayesian model not only manages all the previous types of MRs to efficiently perform the tracking of multiple objects, but also infers the number of existing moving objects. The inference in the proposed Bayesian Visual Surveillance Model (BVSM) is performed by means of a particle filtering method that approximates the posterior distribution of the tracked objects by a set of unweighted samples. The procedure to compute these samples is based on a combination of several techniques such as ancestral sampling and MCMC simulation. Experimental results in indoor video sequences have shown that the proposed Bayesian framework is able to reliably detect and track a variable number of moving objects in real operating conditions.

2. PROBLEM DESCRIPTION

The main goal of BVSM is to detect and track a variable number of moving objects. Moving objects are represented by a state vector $\mathbf{x}_{[t]} = \left[\mathbf{x}_{[t,i_x]} | i_x = 1, \ldots, N_{obj}\right]$, where N_{obj} is the number of moving objects at the time step t that varies along the time. Each component $\mathbf{x}_{[t,i_x]} = \left[\mathbf{r}_{[t,i_x]}, \mathbf{v}_{[t,i_x]}, \mathbf{s}_{[t,i_x]}, \mathbf{l}_{[t,i_x]}\right]$ contains the tracking information of a moving object. $\mathbf{r}_{[t,i_x]}$ is the object position over the image plane, $\mathbf{v}_{[t,i_x]}$ is the object velocity, $\mathbf{s}_{[t,i_x]}$ is the object size represented by orientable bounding box, and $\mathbf{l}_{[t,i_x]}$ is a label that univocally identifies the object along the video stream. .

This work has been partially supported by the Ministerio de Ciencia e Innovación of the Spanish Government under the project TEC2010-20412 (Enhanced 3DTV).

In order to estimate the vector state, a sequence of noisy measurements $\mathbf{z}_{[t]} = \left[\mathbf{z}_{[t,i_z]} | i_z = 1, \ldots, N_{ms}\right]$ are used, where N_{ms} is the number of measurements at the time step t that varies along the time. Every component contains the information of a measurement $\mathbf{z}_{[t,i_z]} = \left[\mathbf{r}_{[t,i_z]}, \mathbf{s}_{[t,i_z]}, \mathbf{l}_{[t,i_z]}\right]$, which is the position, size and a label of a moving region (MR), respectively. MRs represent potential moving objects and they are obtained by means of a background subtraction technique. In this paper, it has been used the background subtraction technique presented in [6]. Measurements between consecutive time steps are unordered, which means that the correspondence between MRs and tracked objects is unknown. In addition, the correspondence is not one-to-one in general. Interactions among objects, changes in illumination, varying object appearances, shadows, reflections, and complex backgrounds give rise to missing MRs (missing measurements), objects split in several MRs (split measurements), MRs merging several objects (merged measurements), and false MRs (clutter). To deal with this problem, a data association stage is introduced to compute the correspondence between MRs and tracked objects. This correspondence is represented by the variable $\mathbf{a}_{[t]} = \left[\mathbf{a}_{[t,i_a]} | i_a = 1, \ldots, N_{ms}\right]$, where each component expresses the association of a MR as

$$\mathbf{a}_{[t,i_a]} = \begin{cases} \{i_x | i_x \in \{1, \ldots, N_obj\}\} & \text{if } cond_1 \\ 0 & \text{if } cond_2. \end{cases} \quad (1)$$

$cond_1$ is a logical condition that is true if $\mathbf{z}_{[t,i_a]}$ is associated to $\{\mathbf{x}_{[t,i_x]}\}$, and $cond_2$ is another condition that is verified if $\mathbf{z}_{[t,i_a]}$ is not associated to any object.

The set of MRs not associated to any object, $\mathbf{z}_{na} = \{\mathbf{z}_{[t,i_a]} | \mathbf{a}_{[t,i_a]} = 0\}$, can be originated by clutter or by the entrance of new objects in the scene. This information is encoded by the variable $\mathbf{b}_{[t]} = \left[\mathbf{b}_{[t,i_b]} | i_b = 1, \ldots, N_{zna}\right]$, where N_{zna} is the number of components of $\mathbf{z}_{na}$. The $\mathbf{b}_{[t,i_b]}$ component indicates the origin of the i_b^{th} MR in $\mathbf{z}_{na}$ as

$$\mathbf{b}_{[t,i_b]} = \begin{cases} 1 & \text{new object,} \\ 0 & \text{clutter.} \end{cases} \quad (2)$$

On the other hand, there can be also objects without any MRs associated to them, $\mathbf{x}_{na} = \left[\mathbf{x}_{[t,i_x]} | \neg\exists i_a, \mathbf{a}_{[t,i_a]} = i_x\right]$, because either the MR is missing, or the object has exited from the scene. This information is given by the variable $\mathbf{d}_{[t]} = \left[\mathbf{d}_{[t,i_d]} | i_d = 1, \ldots, N_{xna}\right]$ where N_{xna} is the number of components of $\mathbf{x}_{na}$. The $\mathbf{d}_{[t,i_d]}$ component indicates the state of the i_d^{th} object in $\mathbf{x}_{na}$ as

$$\mathbf{d}_{[t,i_d]} = \begin{cases} 1 & \text{object has exited,} \\ 0 & \text{object still present (missing measurement).} \end{cases} \quad (3)$$

The number of tracked objects in the scene at each time step is then directly determined by the birth (entrance of new object) and dead events The variables $\mathbf{b}_{[t]}$ and $\mathbf{d}_{[t]}$ controls the size of $\mathbf{x}_{[t]}$ by adding or removing components (moving objects). Thus, the number of tracked objects in the scene at each time step is obtained.

3. BAYESIAN VIDEO SURVEILLANCE MODEL

From a Bayesian perspective, the goal is to estimate the posterior probability density function (pdf) of the state vector, $p(\mathbf{x}_[t]|\mathbf{z}_{[1:t]})$, using the prior information about the object dynamics and the sequence of available measurements (moving regions) until the current time step $\mathbf{z}_{[1:t]} = \{\mathbf{z}_{[1]}, ..., \mathbf{z}_{[t]}\}$. This probability contains all the required information to optimally compute an estimation of the number of moving objects and their trajectories.

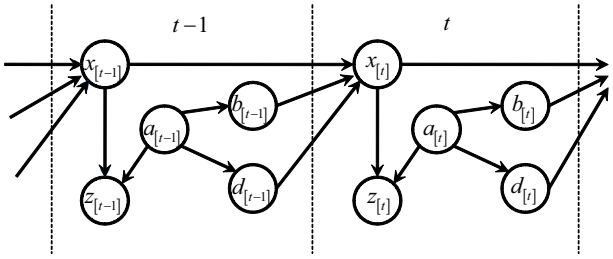

Fig. 1. Graph model showing the dependence of all the variables.

To derive a tractable mathematical expression of $p(\mathbf{x}_[t]|\mathbf{z}_{[1:t]})$, the conditional independence properties of all the variables involved in the surveillance system are used. Fig. 1 depicts these conditional dependencies according to Sec. 2. Then, the mathematical expression of $p(\mathbf{x}_[t]|\mathbf{z}_{[1:t]})$ can be expressed as

$$p(\mathbf{x}_{[t]}|\mathbf{z}_{[1:t]}) = \sum_{\mathbf{a}_{[t]}} \sum_{\mathbf{b}_{[t]}} \sum_{\mathbf{d}_{[t]}} p(\mathbf{x}_{[t]}, \mathbf{a}_{[t]}, \mathbf{b}_{[t]}, \mathbf{d}_{[t]}|\mathbf{z}_{[1:t]}), \quad (4)$$

where $p(\mathbf{x}_{[t]}, \mathbf{a}_{[t]}, \mathbf{b}_{[t]}, \mathbf{d}_{[t]}|\mathbf{z}_{[1:t]})$ is the joint posterior pdf of all the variables. Defining $\mathbf{y}_{[t]} = \{\mathbf{x}_{[t]}, \mathbf{a}_{[t]}, \mathbf{b}_{[t]}, \mathbf{d}_{[t]}\}$ and using the Bayes' theorem [7], the joint posterior pdf can be expressed in a recursive way as

$$p(\mathbf{y}_{[t]}|\mathbf{z}_{[1:t]}) = \frac{p(\mathbf{z}_{[t]}|\mathbf{z}_{[1:t-1]}, \mathbf{y}_{[t]})}{p(\mathbf{z}_{[t]}|\mathbf{z}_{[1:t-1]})} \times$$
$$\times \int p(\mathbf{y}_{[t]}|\mathbf{z}_{[1:t-1]}, \mathbf{y}_{[t-1]}) p(\mathbf{y}_{[t-1]}|\mathbf{z}_{[1:t-1]}) d\mathbf{y}_{[t-1]} \quad (5)$$

where $p(\mathbf{y}_{[t-1]}|\mathbf{z}_{[1:t-1]})$ is the joint posterior pdf at the previous time step, $p(\mathbf{y}_{[t]}|\mathbf{z}_{[1:t-1]}, \mathbf{y}_{[t-1]})$ is the joint transition probability, $p(\mathbf{z}_{[t]}|\mathbf{z}_{[1:t-1]}, \mathbf{y}_{[t]})$ is the likelihood, and $p(\mathbf{z}_{[t]}|\mathbf{z}_{[1:t-1]})$ is just a normalization constant given by

$$p(\mathbf{z}_{[t]}|\mathbf{z}_{[1:t-1]}) = \int p(\mathbf{z}_{[t]}|\mathbf{z}_{[1:t-1]}, \mathbf{y}_{[t]}) p(\mathbf{y}_{[t]}|\mathbf{z}_{[1:t-1]}) d\mathbf{y}_{[t]}. \quad (6)$$

Making use of the concept of "d-separation" [8] to analyze the conditional independence among variables, the likelihood can be simplified as

$$p(\mathbf{z}_{[t]}|\mathbf{z}_{[1:t-1]}, \mathbf{y}_{[t]}) = p(\mathbf{z}_{[t]}|\mathbf{x}_{[t]}, \mathbf{a}_{[t]}) \quad (7)$$

since $\mathbf{z}_{[1:t-1]}$ are d-separated from $\mathbf{z}_{[t]}$ by $\mathbf{x}_{[t]}$, and $\mathbf{b}_{[t]}$ and $\mathbf{d}_{[t]}$ are d-separated from $\mathbf{z}_{[t]}$ by $\mathbf{a}_{[t]}$. The likelihood evaluates how well a candidate state vector along with a specific data association accounts for the set of measurements, i.e. how well a configuration of tracked objects accounts for the set of MRs taking into account a particular correspondence between and MRs. The likelihood can be factorized according to the MRs that have not been assigned to any object $\mathbf{z}_{na}$, and the ones that do have $\mathbf{z}_a$

$$p(\mathbf{z}_{[t]}|\mathbf{x}_{[t]}, \mathbf{a}_{[t]}) = p(\mathbf{z}_{na}) p(\mathbf{z}_a|\mathbf{x}_{[t]}, \mathbf{a}_{[t]}), \quad (8)$$

where $p(\mathbf{z}_{na})$ is modeled by a multivariate uniform distribution that does not depend on the state vector, and $p(\mathbf{z}_a|\mathbf{x}_{[t]}, \mathbf{a}_{[t]})$ follows a multivariate Gaussian distribution given by

$$p(\mathbf{z}_a|\mathbf{x}_{[t]}, \mathbf{a}_{[t]}) = N(\mathbf{H}_{[t,1]}\mathbf{z}_a; \mathbf{H}_{[t,2]}\mathbf{x}_{[t]}, \mathbf{\Sigma}_{[t,lh]}). \quad (9)$$

The matrices $\mathbf{H}_{[t,1]}$ and $\mathbf{H}_{[t,1]}$ are set to satisfy the MRs-objects correspondence defined by $\mathbf{a}_{[t]}$. The matrix $\mathbf{H}_{[t,1]}$ can encode associations involving several MRs and one object (split measurements), and $\mathbf{H}_{[t,2]}$ associations involving one MR and several objects (merged measurements). The rows of $\mathbf{H}_{[t,1]}$ ($\mathbf{H}_{[t,2]}$) are normalized to reflect that an association of several MRs (objects) creates a virtual MRs (object) whose position and size parameters are the average of the involved MRs (objects).

By applying first the product rule of probability and then the conditional independence properties, the joint transition probability can be simplified as

$$p(\mathbf{y}_{[t]}|\mathbf{z}_{[1:t-1]}, \mathbf{y}_{[t-1]}) = p(\mathbf{x}_{[t]}|\mathbf{x}_{[t-1]}, \mathbf{b}_{[t-1]}, \mathbf{d}_{[t-1]}) \times \\ \times\, p(\mathbf{a}_{[t]}) p(\mathbf{b}_{[t]}|\mathbf{a}_{[t]}) \times p(\mathbf{d}_{[t]}|\mathbf{a}_{[t]}), \tag{10}$$

where it has been taking into account that $\mathbf{z}_{[1:t-2]}$ are d-separated from $\mathbf{x}_{[t]}$ by $\mathbf{x}_{[t-1]}$; $\{\mathbf{z}_{[t-1]}, \mathbf{a}_{[t-1]}\}$ are d-separated from $\mathbf{x}_{[t]}$ by $\mathbf{b}_{[t-1]}$; $\{\mathbf{z}_{[1:t-1]}, \mathbf{x}_{[t-1]}, \mathbf{a}_{[t-1]}, \mathbf{b}_{[t-1]}, \mathbf{d}_{[t-1]}\}$ are d-separated from $\mathbf{a}_{[t]}$ by $\mathbf{x}_{[t]}$; $\mathbf{x}_{[t]}$ is d-separated from $\mathbf{a}_{[t]}$ by $\mathbf{x}_{[t+1]}$; $\{\mathbf{z}_{[1:t-1]}, \mathbf{x}_{[t-1:t]}, \mathbf{a}_{[t-t]}, \mathbf{b}_{[t-1:t]}, \mathbf{d}_{[t-1]}\}$ are d-separated from $\mathbf{d}_{[t]}$ by $\{\mathbf{a}_{[t]}, \mathbf{x}_{[t+1]}\}$; and $\{\mathbf{z}_{[1:t-1]}, \mathbf{x}_{[t-1:t]}, \mathbf{a}_{[t-t]}, \mathbf{b}_{[t-1:t]}, \mathbf{d}_{[t-1]}\}$ are d-separated from $\mathbf{d}_{[t]}$ by $\{\mathbf{a}_{[t]}, \mathbf{x}_{[t+1]}\}$.

The term $p(\mathbf{a}_{[t]})$ is used to impose the following restriction over the data association: a MR cannot be associated at the same time to clutter and objects, since it makes no sense in the visual tracking. The term $p(\mathbf{b}_{[t]}|\mathbf{a}_{[t]})$ expresses the probability that the measurements that have not been assigned to any object, $\mathbf{z}_{na}$, be considered as new objects, rather than clutter. It is modeled by a multinomial distribution

$$p(\mathbf{b}_{[t]}|\mathbf{a}_{[t]}) = \prod_{i_b} \mu_{i_b}^{\mathbf{b}_{[t,i_b]}}, \tag{11}$$

where μ_{i_b} is the probability that the i_b^{th} MR in $\mathbf{z}_{na}$ be a new object. The term $p(\mathbf{d}_{[t]}|\mathbf{a}_{[t]})$ is the probability that the objects not associated with any MR, $\mathbf{x}_{na}$, leave the scene,and it is modeled by a Gamma distribution. This distribution simulates the "time to death" of one object according to the last time a MR was associated to it. Finally, the term $p(\mathbf{x}_{[t]}|\mathbf{x}_{[t-1]}, \mathbf{b}_{[t-1]}, \mathbf{d}_{[t-1]})$ predicts the number of objects and their tracking information between consecutive time steps. This prediction is accomplished in three stages. First, the tracked objects that have left the scene, indicated by $\mathbf{d}_{[t-1]}$, are removed from $\mathbf{x}_{[t-1]}$. Then, the tracking information of the alive objects is predicted using a constant model for the velocity and size with Gaussian uncertainty [5]. Lastly, the new objects according to $\mathbf{b}_{[t-1]}$ are added to $\mathbf{x}_{[t]}$, and their parameters are initialized by a Gaussian distribution, whose parameters are set in accordance with the MR parameters that gave rise to the new object.

Once the expression of the joint posterior pdf has been derived, an optimal estimation of the state vector $\widetilde{\mathbf{x}}_k$ is obtained by means of the Minimum Mean Squared Error (MMSE) estimator. However, $p(\mathbf{y}_{[t-1]}|\mathbf{z}_{[1:t]})$ can not be analytically solved due to the non-linear and non-Gaussian processes involved in the surveillance system [9]. To overcome this problem, a particle filtering strategy based on ancestral sampling is used to approximate the joint posterior pdf.

4. PARTICLE FILTERING APPROXIMATION

The joint posterior pdf can be approximated by a set of N_{sam} unweighted samples $\mathbf{y}_{[t]}^{[i]} = \{\mathbf{x}_{[t]}^{[i]}, \mathbf{a}_{[t]}^{[i]}, \mathbf{b}_{[t]}^{[i]}, \mathbf{d}_{[t]}^{[i]}\}$, also called particles, as

$$p(\mathbf{y}_{[t]}|\mathbf{z}_{[1:t]}) = \sum_{i=1}^{N_{sam}} \delta\left(\mathbf{y}_{[t]} - \mathbf{y}_{[t]}^{[i]}\right), \tag{12}$$

where $\delta(x)$ is a multivariate Dirac delta function. According to Monte Carlo simulation, samples are drawn from a function proportional to the own joint posterior pdf

$$p(\mathbf{y}_{[t]}|\mathbf{z}_{1:t}) \propto q(\mathbf{y}_{[t]}|\mathbf{z}_{1:t}) = p(\mathbf{z}_{[t]}|\mathbf{z}_{[1:t-1]}, \mathbf{y}_{[t]}) \times \\ \times \int p(\mathbf{y}_{[t]}|\mathbf{z}_{[1:t-1]}, \mathbf{y}_{[t-1]}) p(\mathbf{y}_{[t-1]}|\mathbf{z}_{[1:t-1]}) d\mathbf{y}_{[t-1]}. \tag{13}$$

Assuming that joint posterior pdf has been approximated in the previous time step by a set of unweighted samples as in Eq. 12, $q(\mathbf{y}_{[t]}|\mathbf{z}_{1:t})$ can be expressed as

$$q(\mathbf{y}_{[t]}|\mathbf{z}_{1:t}) = p(\mathbf{z}_{[t]}|\mathbf{z}_{[1:t-1]}, \mathbf{y}_{[t]}) \sum_{i=1}^{N_{sam}} p(\mathbf{y}_{[t]}|\mathbf{z}_{[1:t-1]}, \mathbf{y}_{[t-1]}^{[i]}) = \\ = p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{a}_t) p(\mathbf{a}_{[t]}) p(\mathbf{b}_{[t]}|\mathbf{a}_{[t]}) p(\mathbf{d}_{[t]}|\mathbf{a}_{[t]}) \times \\ \times \sum_{i=1}^{N_{sam}} p(\mathbf{x}_{[t]}|\mathbf{x}_{[t-1]}^{[i]}, \mathbf{b}_{[t-1]}^{[i]}, \mathbf{d}_{[t-1]}^{[i]}). \tag{14}$$

The procedure to draw samples from $q(\mathbf{y}_{[t]}|\mathbf{z}_{[1:t]})$ is based on the ancestral sampling technique [10]. First, a sample $\mathbf{y}_{[t-1]}^{[i]}$ is drawn from $p(\mathbf{y}_{[t-1]}|\mathbf{z}_{1:t-1})$, which has been approximated by a discrete probability. Second, a sample $\mathbf{x}_{[t]}^{[i]}$ is drawn from $p(\mathbf{x}_{[t]}|\mathbf{x}_{[t-1]}^{[i]}, \mathbf{b}_{[t-1]}^{[i]}, \mathbf{d}_{[t-1]}^{[i]})$, which was essentially a Gaussian distribution. Then, $\mathbf{a}_{[t]}^{[i]}$ is sampled from $q(\mathbf{a}_{[t]}^{[i]}) = p(\mathbf{z}_t|\mathbf{x}_t^{[i]}, \mathbf{a}_t) p(\mathbf{a}_{[t]})$. Since the likelihood $p(\mathbf{z}_t|\mathbf{x}_t^{[i]}, \mathbf{a}_t)$ is a discrete probability over $\mathbf{a}_t$ given $\mathbf{x}_t^{[i]}$, $q(\mathbf{a}_{[t]}^{[i]})$ is a discrete distribution from which is straightforward to draw samples. However, the fact that several MRs can be associated to several objects gives rise to a combinatorial explosion of possible associations whose computational cost is prohibitively. To deal with this problem, the MCMC based sampling approach presented in [5] is used, although lifting the restriction that a MR cannot split and merge at the same time. This approach defines a Markov Chain whose stationary distribution is just the desired $q(\mathbf{a}_{[t]}^{[i]})$. Finally, $\mathbf{b}_{[t]}^{[i]}$ and $\mathbf{d}_{[t]}^{[i]}$ are drawn from the multinomial probability $p(\mathbf{b}_{[t]}|\mathbf{a}_{[t]}^{[i]})$ and the gamma distribution $p(\mathbf{d}_{[t]}|\mathbf{a}_{[t]}^{[i]})$ respectively. As a result, a sample $\mathbf{y}_{[t]}^{[i]} = \{\mathbf{x}_{[t]}^{[i]}, \mathbf{a}_{[t]}^{[i]}, \mathbf{b}_{[t]}^{[i]}, \mathbf{d}_{[t]}^{[i]}\}$ is obtained from $p(\mathbf{y}_{[t]}|\mathbf{z}_{[1:t]})$.

5. RESULTS

The proposed Bayesian Video Surveillance Model has been tested using a dataset consisting of several indoor situations with a variable number of moving objects. The main challenge is the proper management of merged, split, missing and clutter MRs that occur in the operation of a typical moving object detector based on background subtraction.

An illustrative example of the tracking procedure is shown in Fig. 2. The first shows the detected moving regions (white regions) computed by the detector. The second row shows the position and size information relative to the vector state samples that approximate the posterior pdf $p(\mathbf{x}_{[t]}|\mathbf{z}_{[1:t]})$. Lastly, the third row shows the tracked objects marked by bounding boxes, which result from the MMSE estimation over the joint posterior pdf. On the other hand, the estimated number of objects per time step is shown in Fig. 3 for the previous sequence. The solid line indicates the number of estimated objects, and the dashed line the actual number of moving objects.

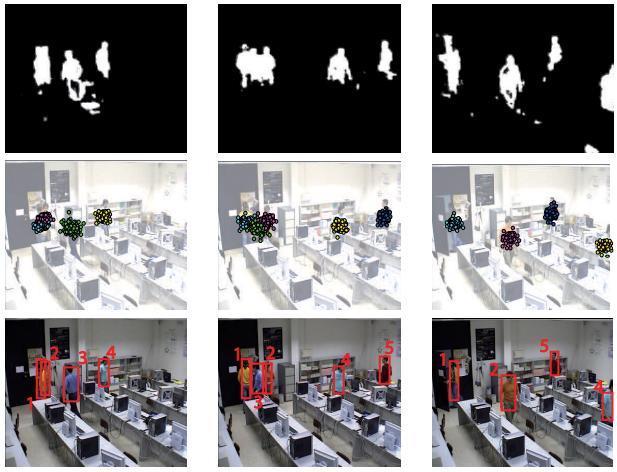

Fig. 2. Tracking a variable number of objects

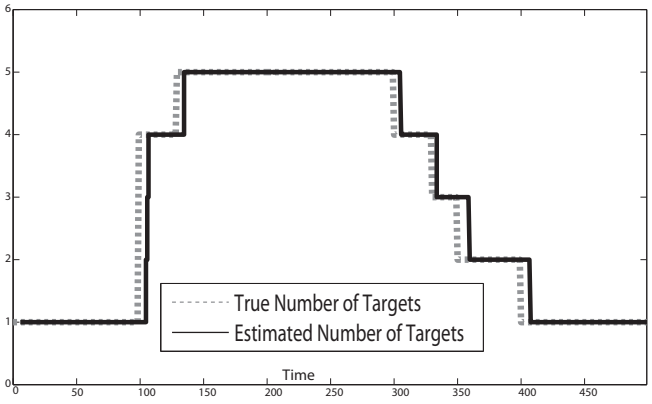

Fig. 3. Estimated number of objects across time.

The overall tracking performance has been evaluated by means of two different measures: the number of tracking errors. N_{te} and the average number of miscounted track objects, A_{mo}. Regarding N_{te}, it is considered that a tracking error occurs when the estimated bounding box of an object and the one corresponding to the ground truth do not overlap each other. The other measure is defined as $A_{mo} = \sum_t d_{o,t}/T$, where $d_{o,t}$, is the absolute value of difference between the actual and the estimated number of moving objects at the time step t, and T is the length of the sequence. Tab. 1 shows global tracking results in the aforementioned dataset using the defined performance measures. As can be observed from the low values of A_{mo}, the proposed BVSM is able to accurately infer the number of moving objects, since the model has the ability to create and remove objects in order that the estimated number of objects be consistent with the existing detections. On the other hand, the low values of N_{te} prove a great efficiency in the tracking task thanks to the robust management of false, missing, split and merged MRs.

6. CONCLUSIONS

Automatic detection and tracking of multiple moving objects for visual applications is a challenging task. The main problem arises from the fact that set of unordered detections (MRs) and the set of existing moving objects in the scene can not be mapped one to one. The reason is that visual detectors produces undesirable false, missing, split and merged MRs. In this paper, a novel Bayesian model for visual object detection and tracking has been presented, which is able to handle the complex set of detected MRs to successfully estimate the number of moving regions and their trajectories. This is accomplished by means of an reliable modeling of the causes that originate the undesirable detections, especially the split and merged ones. On the other hand, the high complexity of the proposed Bayesian model has forced the use of approximate inference. For this purpose, a particle filtering approach that combines ancestral and MCMC sampling techniques has been used, which allow to accurately simulate the data association between MRs and tracked objects, and thus to reliably estimate the number of objects and their trajectories. Experimental results have proven the efficiency of the this approach in real situations.

Video description	N_{te}	A_{mo}	T
2 people, 1 cross	0	0.013	500
3 people, 1 cross	0	0.017	450
3 people, 2 crosses	3	0.021	500
4 people, 2 crosses	3	0.025	550
5 people, 2 crosses	9	0.031	650
5 people, 4 crosses	14	0.033	600

Table 1. Overall tracking performance.

7. REFERENCES

[1] G.W. Pulford, "Taxonomy of multiple target tracking methods," *IEEE Proc. Radar, Sonar and Navigation*, vol. 152, no. 5, pp. 291–304, 2005.

[2] Auguste Genovesio and Jean-Christophe Olivo-Marin, "Split and merge data association filter for dense multi-target tracking," in *Proc. ICPR*. IEEE, 2004, vol. 4, pp. 677–680.

[3] Yunqian Ma, Qian Yu, and Isaac Cohen, "Target tracking with incomplete detection," *Comp. Vision and Image Understanding*, vol. 113, no. 4, pp. 580–587, 2009.

[4] Qian Yu and G. Medioni, "Multiple-target tracking by spatiotemporal monte carlo markov chain data association," *IEEE Trans. PAMI*, vol. 31, no. 12, pp. 2196–2210, 2009.

[5] Zia Khan, Tucker Balch, and Frank Dellaert, "Multitarget tracking with split and merged measurements," in *Proc. CVPR*. IEEE, 2005, vol. 1, pp. 605–610.

[6] Zoran Zivkovic and Ferdinand van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recogn. Lett.*, vol. 27, pp. 773–780, 2006.

[7] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[8] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, October 2007.

[9] Arnaud Doucet, Simon Godsill, and Christophe Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[10] David J. C. Mackay, *Information Theory, Inference & Learning Algorithms*, Cambridge University Press, June 2002.