Multi-level Background Initialization using Hidden Markov Models

Marco Cristani
University of Verona
Dipartimento di Informatica
cristanm@sci.univr.it

Manuele Bicego
University of Verona
Dipartimento di Informatica
bicego@sci.univr.it

Vittorio Murino
University of Verona
Dipartimento di Informatica
vittorio.murino@univr.it

ABSTRACT

Most of the automated video-surveillance applications are based on the process of background modelling, aimed at discriminating motion patterns of interest at pixel, region or frame level in a nearly static scene. The issues characterizing an ordinary background modelling process are typically three: the background model representation, the initialization, and the adaptation. This paper proposes a novel initialization algorithm, able to bootstrap an integrated pixel- and region-based background modelling algorithm. The input is an uncontrolled video sequence in which moving objects are present, the output is a pixel- and region-level statistical background model describing the static information of a scene. At the pixel level, multiple hypotheses of the background values are generated by modelling the intensity of each pixel with a Hidden Markov Model (HMM), also capturing the sequentiality of the different color (or gray-level) intensities. At the region level, the resulting HMMs are clustered with a novel similarity measure, able to remove moving objects from a sequence, and obtaining a segmented image of the observed scene, in which each region is characterized by a similar spatio-temporal evolution. Experimental trials on synthetic and real sequences have shown the effectiveness of the proposed approach.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*video analysis*; I.5.1 [**Pattern Recognition**]: Models—*statistical*; I.5.3 [**Pattern Recognition**]: Clustering—*similarity measures*

General Terms

Design, Performance

Keywords

Video Surveillance, pixel-region background initialization, Hidden Markov Model

IWVS'03, November 7, 2003, Berkeley, California, USA.

1. INTRODUCTION

Analysis and understanding of video sequences is an active research field, whose importance is rapidly increased in the last years, due to the availability of more and more powerful hardware, to the development of effective real-time techniques, and to the potential vastity of the involved applications [30, 6, 28]. Video surveillance is undoubtedly one of the most interesting applications of sequence analysis: human action recognition [31], semantic indexing of video [21], and, more generally, on-line discovering of unusual activities [12] are all tasks under investigations to partially or fully automate the surveillance.

Typically, a video-surveillance system contemplates the monitoring of a site for long periods, using a static camera whose goal is to distinguish (and possibly classify) unusual behaviors from typical ones. To this end, the basic operation needed is the separation of the moving objects, the so-called *foreground* (FG), from the static information [7], the *background* (BG). This process is usually called background modelling.

The issues characterizing a background modelling process are usually three: model representation, model initialization, and model adaptation. The first describes the kind of model (e.g., mixture of Gaussians) used to represent the background; the second one regards the initialization of this model, and the third one relies to the mechanism used for adapting the model to the background changes (e.g., illumination changes). Recently, several techniques have been proposed in order to address the representation and the adaption issues, whereas the model initialization has received poor attention. In the background model initialization problem, also called *bootstrapping* [29], the input is a short uncontrolled video sequence in which a number of moving objects may be present. The purpose is then to produce a background model describing the observed scene. Actually, most of the background models are built on a set of initial parameters that comes out from a short sequence, in which no foregrounds objects are present [10]. This is a too strong assumption, because in some situations it is difficult or impossible to control the area being monitored (e.g., public zones), which are characterized by a continuous presence of moving objects, or other disturbing effects.

In the literature, the initialization problem is typically disregarded, and only few methods are present. All of these methods discard the solution of computing a simple mean over all the frames, because it produces an image that exhibits blending pixel values in areas of foreground presence. A general analysis regarding the blending rate and how it

may be computed is present in [8]. In [9], the background initial values are estimated by calculating the median value of all the pixels in the training sequence, assuming that the background value in every pixel location is visible more than 50% of the time during the training sequence. Even if this method avoids the blending effects of the mean, the output of the median will contains large error when this assumption is false. Another proposed work [18], called adaptive smoothness method, avoids the problem of blending finding intervals of stable intensity in the sequence. Then, using some heuristics, the longest stable value for each pixel is selected and used as the value that most likely represents the background. This method is similar to the recent Local Image Flow algorithm [11], which generates background values' hypotheses by locating intervals of relatively constant intensity, and weighting these hypotheses by using local motion information. Unlike most of the approaches, this method does not treat each pixel value sequence as an i.i.d. (independent identically distributed) process, but it considers also information generated by the neighboring locations.

To the best of our knowledge, all the proposed methods are devoted to the initialization of algorithms working at the pixel level, disregarding higher-level information. Indeed, background analysis could be carried out at different data-abstraction levels: pixel, region, and frame levels [29]. The pixel-level analysis processes independently each pixel, classifying it as foreground or background, and managing adaptation to changing background [25]. In this modality, the analysis is performed at a very low level, and many problems of the background subtraction remain unsolved, such as local or global sudden illumination changes [5]. The region-level analysis considers a higher level representation, modelling also inter-pixel relationships, so allowing a refinement of the modelling obtained at the pixel level. For instance, in [29], the spatial motion of the foreground is detected by segmenting the foreground patterns, and intersecting successive segmentations in order to improve the region-level dynamics and to avoid the problem of foreground aperture. Finally, the frame-level analysis looks for changes in large parts of the image, and eventually swaps in more expressive background models [27, 19].

Recently proposed background models [29, 5, 14, 13] try to integrate these different kinds of data, producing beneficial effects on the the effectiveness of the background modelling. Initialization methods for this kind of integrated background models are almost missing in the literature, only some ideas are reported in [11] exploiting neighborhood information.

The aim of this paper is to propose some contributions in this context. A novel bootstrapping method is developed, able to initialize a background model that considers both pixel and region information [5]. This method [5] integrates the information obtained from a standard Time-Adaptive, Per-Pixel, Mixture Of Gaussian (TAPPMOG) [25, 26, 13] background model with region information, obtained with a spatial segmentation of the background. This integration permits to recover from sudden non uniform illumination changes, which represents one of the most serious problem in video surveillance applications. In order to initialize this method, we need a bootstrapping procedure that operates at two levels: at the pixel level, we need to know the most probable components of the background in each scene location, and, at the region level, we need a meaningful spatial partition of the scene. The method proposed in this paper takes in input a short arbitrary video sequence and generates a probabilistic representation of that sequence from which we could derive both representations.

Spatial scene segmentation of a scene may appear, at a glance, as easily obtainable by merely segmenting the first frame of the sequence, or the average frame. Nevertheless, like in the background initialization case described below, this is a too simplistic assumption for two reasons. First, moving objects can be present in the scene, and they could not be removed without a negative impact. Second, especially in case of illumination changes, we are interested in a spatio-temporal segmentation, in which the spatial gray-level data are augmented with the temporal information in order to obtain connected regions that present a chromatic and temporal similar meaningful behavior.

The approach proposed in this paper is based on the use of a forest of Hidden Markov Models (HMMs) [23], which representins the scene observed by a static camera by modelling the temporal gray-level evolution of each pixel. This representation is then used to initialize the integrated pixel- and region-based model in a twofold manner. First, by looking at the model parameters of each HMM we could infer which values of intensity of each pixel are most stable and most probably belonging to the background. This information can be used to initialize the pixel-level part of the background model.

The second information extracted is a chromatic and temporal segmentation of the background, obtained by clustering the HMMs. It is important to note that HMM-based clustering has been poorly addressed in the past, and only few papers are present in the literature [24, 16, 3, 20]. Typically, these models are used to devise a distance between sequences, which is subsequently used to perform standard clustering. In this paper, a new measure is proposed, able to remove non-stationary components of a sequence. Using this measure and a region-growing segmentation approach, we are able to process the set of pixel sequences in order to segment the scene in groups of pixels showing an homogeneous color with a similar temporal evolution. In this case, the resulting segmentation is a spatial partition of the scene, obtained by using all available information: chromatic (different regions have an homogeneous gray level value), spatial (each region is connected), and temporal (each region varies its color similarly along time). In conclusion, our method has two great advantages: first, the spatial information is augmented with temporal data that captures also the frequency of the color variation occurring in a single region, so allowing a more detailed and informative partitioning; second, moving objects have not to be removed from the sequence as this operation is accomplished by the similarity measure devised.

In the experimental session, the proposed initialization algorithm is tested using synthetic and real sequences. We will show that the proposed approach represents an useful tool able to initialize the pixel- and region-level background estimation processes.

The rest of the paper is organized as follows. In Section 2, the basic theory of the Hidden Markov Models and the description of the stationary probability distribution are reported, the approaches for HMM-based clustering of sequences are reviewed, and the integrated pixel- and region-based background modelling scheme [5] is shortly presented. The proposed approach is then described in Section 3: after

describing the probabilistic modelling of the video sequence, the methods used to initialize the pixel- and the region-levels background modelling are detailed. Experimental results on the proposed approach are presented in Section 4 and, finally, Section 5 contains conclusions and future perspectives.

2. BASIC THEORY AND METHODOLOGY

In this section the fundamental instruments of the proposed initialization approach are described. In particular, in Section 2.1 the definition of the Hidden Markov Model approach is given; Section 2.2 introduces the concept of stationary probability of a HMM, representing the key entity of the approach proposed in this paper. Section 2.3 contains the description of the HMM-based clustering approach; finally, in Section 2.4, the integrated pixel- and region-based methodology to background modelling presented in [5] is briefly summarized.

2.1 Hidden Markov Models

A discrete-time Hidden Markov Model $\boldsymbol{\lambda}$ can be viewed as a Markov model whose states are not directly observable: instead, each state is characterized by a probability distribution function, modelling the observations corresponding to that state. More formally, a HMM is defined by the following entities [23]:

- $S = \{S_1, S_2, \cdots, S_N\}$ the finite set of (hidden) states;
- the transition matrix $\mathbf{A} = \{a_{kj}\}$, $1 \leq k, j \leq N$ representing the probability of moving from state S_k to state S_j,

  $$a_{kj} = P[Q_{t+1} = S_j | Q_t = S_k], \quad 1 \leq k, j \leq N,$$

  with $a_{kj} \geq 0$, $\sum_{j=1}^{N} a_{kj} = 1$, and where Q_t denotes the state occupied by the model at time t.
- the emission matrix $\mathbf{B} = \{b(o|S_k)\}$, indicating the probability of emission of symbol $o \in V$ when system state is S_k; V can be a discrete alphabet or a continuous set (e.g. $V = I\!R$), in which case $b(o|S_k)$ is a probability density function. In this paper we used continuous Gaussian HMM, *i.e.*

  $$b(o|S_k) = \mathcal{N}(o|\mu_k, \Sigma_k).$$

  where $\mathcal{N}(o|\mu, \Sigma)$ denotes a Gaussian density of mean μ and covariance Σ, evaluated at o;
- $\boldsymbol{\pi} = \{\pi_k\}$, the initial state probability distribution,

  $$\pi_k = P[Q_1 = S_k], \quad 1 \leq k \leq N$$

  with $\pi_k \geq 0$ and $\sum_{k=1}^{N} \pi_k = 1$.

For convenience, we represent an HMM by a triplet $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$.

Learning the HMM parameters, given a set of observed sequences $\{\mathbf{O}_i\}$, is usually performed using the well-known Baum-Welch algorithm [23], which is able to determine the parameters maximizing the likelihood $P(\{\mathbf{O}_i\}|\boldsymbol{\lambda})$. One of the steps of the Baum-Welch algorithm is an evaluation step, where it is required to compute $P(\mathbf{O}|\boldsymbol{\lambda})$, given a model $\boldsymbol{\lambda}$ and a sequence $\mathbf{O}$. This is computed using the *forward-backward procedure* [23].

2.2 The stationary probability distribution

This section defines the stationary probability distribution of a HMM, which represents the core of our approach.

Given an HMM $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$, consider the Markov chain $\mathbf{Q} = Q_1, Q_2, Q_3 ...$ with state set $S = \{S_1, ..., S_N\}$, stochastic transition matrix $\mathbf{A}$, and initial state probability $\boldsymbol{\pi}$. We can define the vector of state probabilities at time t as

$$\begin{aligned} \mathbf{p}_t &= [\mathbf{p}_t(1), ..., \mathbf{p}_t(k), ...\mathbf{p}_t(N)] \\ &= [P(Q_t = S_1), ..., P(Q_t = S_k), ..., P(Q_t = S_N)] \end{aligned}$$

where $\mathbf{p}_t(k)$ represents the probability of being in state S_k at time t. Of course, $\mathbf{p}_t$ can be computed recursively from $\mathbf{p}_1 = \boldsymbol{\pi}\mathbf{A}$, $\mathbf{p}_2 = \mathbf{p}_1\mathbf{A} = \boldsymbol{\pi}\mathbf{A}\mathbf{A}$, and so on. In short, $\mathbf{p}_t = \boldsymbol{\pi}\mathbf{A}^t$.

We are interested in the *stationary probability distribution* $\mathbf{p}_\infty$, which characterizes the equilibrium behavior of the Markov chain, *i.e.*, when we let it evolve indefinitely. This vector represents the probability that the system is in a particular state after an infinity number of iterations. Since it is a stationary distribution, $\mathbf{p}_\infty$ has to be a solution of

$$\mathbf{p}_\infty = \mathbf{p}_\infty\mathbf{A}$$

or, in other words, it has to be a left eigenvector of $\mathbf{A}$ associated with the unit eigenvalue. Under some conditions (see [4] for details), the Perron-Frobenius theorem states that matrix $\mathbf{A}$ has a unit (left) eigenvalue and the corresponding left eigenvector is $\mathbf{p}_\infty$. All other eigenvalues of $\mathbf{A}$ are strictly less than 1, in absolute value. Therefore, finding $\mathbf{p}_\infty$ for a given $\mathbf{A}$ amounts to solve the corresponding eigenvalue/eigenvector problem.

2.3 HMM-based clustering of sequences

HMMs have not been extensively employed for clustering sequences, only few papers exploring this direction have been published. Even if some alternative approaches to HMM-based clustering have been proposed (e.g., [3, 15]), the typical employed method is the so-called proximity-based strategy, which uses the HMM modelling to compute distances between sequences, and standard pairwise distance matrix-based method (as hierarchical agglomerative) to obtain clustering [24, 16, 17, 20].

More in detail, given a set of R sequences $\{\mathbf{O}_1 ... \mathbf{O}_R\}$ to be clustered, the algorithm performs the following steps:

1. Train one HMM $\boldsymbol{\lambda}_i$ for each sequence $\mathbf{O}_i$.
2. Compute the distance matrix $D = \{D(\mathbf{O}_i, \mathbf{O}_j)\}$, where $D(\mathbf{O}_i, \mathbf{O}_j)$ represents a dissimilarity (or similarity) measure between the sequences $\mathbf{O}_i$ and $\mathbf{O}_j$); this is typically obtained from the forward probability $L_{ij} = P(\mathbf{O}_j|\boldsymbol{\lambda}_i)$, or by devising a measure of distances between models. In the past, some approaches to compute these distances have been proposed (for example, see [20, 1, 24]): early approaches were based on the Euclidean distance of the discrete observation probability, others on entropy, or on co-emission probability of two models, or, very recently, on the Bayes probability of error (see [1] and the references therein). The simplest example has been proposed in [24], and is defined as

   $$D(i, j) = \frac{1}{2}(L_{ij} + L_{ji}) \tag{1}$$

A more complex one, proposed in [20], is defined as

$$D(i,j) = \frac{1}{2}\left\{\frac{L_{ij}-L_{jj}}{L_{jj}} + \frac{L_{ji}-L_{ii}}{L_{ii}}\right\} \qquad (2)$$

3. Given the distance matrix, use a pairwise distance-matrix-based method (*e.g.*, an agglomerative method) to perform the clustering.

In Section 3.3, we will see how this standard method could be extended in order to deal with spatio-temporal segmentation, which represents a particular kind of clustering.

2.4 The integrated pixel- and region-based approach to background modelling

This section briefly presents the integrated pixel- and region-based approach to background modelling proposed in [5]. All the details are in the paper.

The method starts from a standard Time-Adaptive, Per-Pixel, Mixture Of Gaussian (TAPPMOG) technique [25, 13], a widely employed tool for background modelling with several attractive characteristics: adaptiveness, robustness, and real-time implementation to quote a few. This approach models the temporal evolution of each pixel as an i.i.d. process, using a mixture of Gaussians, with an on line training process that permits the adaption to the background changes. Nevertheless, this approach has some drawbacks: first it considers each pixel as an independent process without any use of spatial information or, more generically, higher-level information; second, the choice of the learning rate, that determines the "speed" of the self adaption of TAPPMOGS methods to variations of the background, is critical. The method proposed in [5] introduced an integrated region- and pixel-based approach to background modelling, able to integrate higher-level information into the per-pixel processes. Using region information obtained from a spatial segmentation of a scene, the learning rate of each pixel process could vary, in order to increase the speed of the adaption if the case. As shown in the paper, this integration permits to recover from sudden non-uniform illumination changes, one of the most severe issues in surveillance problems.

In order to initialize this method, we have to provide two kinds of data: an initialization of the pixel-level background, and a spatial segmentation of the scene, so as to identify semantically informative regions.

3. THE INITIALIZATION APPROACH

The proposed approach performs a two-step processing: first, it builds a probabilistic model of the video sequence, and, second, it derives from this model the initialization of both pixel- and region-levels processes.

3.1 The probabilistic modelling of a video sequence

The approach models the training video sequence as a set of independent per-pixel processes (x,y,t), each one describing the temporal gray-level evolution of the location (x,y) of the observed scene (using a fixed camera). Starting from this set of sequences, we need a model able to capture the most important characteristics in order to produce a probabilistic representation of a scene. In particular, we need a model able to determine: 1) the most important gray-level components measured in the whole sequence; 2) the chromatic-temporal variation of those components; 3) the sequentiality with which such components vary. Actually, an adequate computational framework showing these features is constituted by the Hidden Markov Model (HMM) [23]. Using this model, all the above requirements can be accomplished: using HMMs with continuous Gaussian emission probability, the most important gray-level components are modelled by the means $\boldsymbol{\mu}_k$ of the Gaussian functions associated to the states, the variability of those components are encoded in the covariance matrices $\boldsymbol{\Sigma}_k$, and the sequentiality is encoded in the transition matrix $\mathbf{A}$. The whole scene sequence is therefore modelled using a forest of HMMs, one for each pixel. In the experimental session, the training has been carried out using a standard Baum-Welch procedure [23], stopping the training after likelihood convergence. The number of states of each HMM has been fixed to three, which corresponds to the usual number of Gaussian components in a standard pixel-level background subtraction scheme [25, 26]. In this context, this parameter is fixed a priori using heuristic criteria guided by the video complexity, but it can be estimated in a more rigorous fashion adopting an adequate model selection technique [2].

3.2 Pixel-level initialization

In this section, we describe how the probabilistic representation of the video sequence could be used for the pixel-level bootstrapping process. In this case, we want to initialize the pixel process, *i.e.*, the mixture of Gaussians associated to each pixel. This mixture defines the probability of observing the gray level of the current pixel as

$$p(x_t) = \sum_{j=1}^{M} c_j \, \mathcal{N}\left(x_t | \, \mu_j, \sigma_j\right), \qquad (3)$$

where $\mathcal{N}\left(x_t | \, \mu_j, \sigma_j\right)$ denotes a Gaussian density with mean μ_j and variance σ_j, M is the number of the components of the mixture, and c_j is the mixing coefficient (also called weight) of the component j.

This representation puts in correspondence the Gaussians with the main components of the gray-level evolution of a pixel. The mixing coefficients denote the importance of the components with respect to the aim of background modelling, in the sense that the higher the mixing coefficient, the larger the probability that the corresponding Gaussian is associated to an important component, that is, the background. In this case, the initialization relies in the identification of these important components of the signal, without particular care for other components, which have a little impact in the background modelling scheme. It is important to note that there is not a straightforward method to initialize the pixel-level background model, since the input sequence could contain moving objects which do not permit a simple analysis (e.g., averaging).

In our approach, we use the proposed probabilistic representation in order to find the "most important components" of the evolution of each pixel, *i.e.*, to find the most probable background. A similar goal was achieved in [11], where the stability were found by using locally temporal filtering and motion estimations.

The key idea consists in the assumption that the temporal evolution of a pixel could be considered as formed by different components. Therefore, the HMM training is intended as a probabilistic assignment of each of these components

to one different state of the model, and the probability of switching between the components is driven by the transition matrix. We are interested in the significance of these components, as we could assume that the most significant components of the signal will correspond to the background with high probability. It is also important to note that this choice permits to remove possible moving objects in the scene, which are considered unstable, hence, not important components of the signal.

Since there is a correspondence between the components of the signal and the states of the HMM model, the significance of a component can be measured by the weight or the importance of the correspondent HMM state. Given a HMM $\boldsymbol{\lambda}_i$, the "importance" of a state S_k can be naturally associated to its stationary probability $\mathbf{p}_\infty(k)$. This assumption is not new in the literature, and it has been already used in the context of the HMM model selection [2].

Once identified the most stable components (and the corresponding states), we can initialize the mixture associated to the pixel using the parameters of the corresponding HMM. In particular, we initialize the mixture of each pixel by associating each state to a different Gaussian, and the parameters are then defined as

- the parameters (mean and variance) of the Gaussian k of the mixture are initialized with the parameters of the Gaussian of the state S_k;
- the mixing coefficient of the Gaussian k is the stationary probability $\mathbf{p}_\infty(k)$ of the state S_k;

3.3 Region-level initialization

In this section, we describe how the probabilistic representation of the video sequence can be used to initialize the region-level background modelling process. In this case, we need to find a spatial segmentation of the background which individuates the semantically different components of the scene. At first glance, it seems that this segmentation could be easily obtained by segmenting the first frame, or the averaged frame. This is infeasible for two reasons: first, there could be some moving objects in the scene, which are not straightforwardly removable; second, especially in the case of illumination changes, there could be regions in the scene that are spatially homogeneous but differ temporally. We are therefore interested in regions showing both spatial and temporal homogeneity. The goal is to obtain a spatio-temporal segmentation of the scene considering the temporal evolution of the gray-level pixels, *i.e.*, the spatial information is augmented with temporal information, so allowing a more detailed and informative partitioning.

In the literature, spatial-temporal segmentation assumes a slight different meaning in dependence of the application considered. In video-surveillance, it is typically defined as the partition of the video sequence into spatial regions of motion homogeneity (motion segmentation), whereas, in video indexing problems, it is linked to the subdivision of a video in representative shots. Our definition goes beyond these typical descriptions as our spatial-temporal segmentation allows the detection of regions that homogeneously vary in both the spatial and temporal domains.

Given the HMM representation proposed in this paper, it is necessary to define a similarity measure to decide when a group (at least, a couple) of neighboring pixels must be labelled as belonging to the same region. The similarity measure should exhibit some precise characteristics: two sequences have to be considered similar if they share a comparable main chromatic and temporal behavior, independently from the values assumed by the less stable components. By using the measure proposed in equations (1) or in (2), we have that the Gaussian of each state contributes in the same way to the computation of the probability because of the forward-backward procedure. For our target, nevertheless, we need that the Gaussian of each state can contribute differently to the probability computation, depending on the importance of the corresponding state. The idea is then to "flatten" or "spread out" the Gaussians of those states that are not really important, by increasing their variance. In such a way, their contribution to the computation of the probability results decreased. As explained in Section 3.2, the concept of "state importance" could be measured using the stationary probability distribution of the Markov Chain associated with the HMM.

The operation of "flattening" is performed by transforming each model $\boldsymbol{\lambda}_i$ in a new model $\boldsymbol{\lambda}'_i$, where all components remain unchanged, except variances σ_k of state S_k, for each state $k = 1, .., N$, that becomes

$$\sigma'_k = \frac{\sigma_k}{\mathbf{p}_\infty(k)}. \tag{4}$$

The new distance, called $D_{\mathrm{ES}}(i, j)$ (*Enhanced Stationary*), is then computed using the equation (2) using the modified HMM models $\boldsymbol{\lambda}'_i$ ($i = 1, .., L$, L number of image pixels). The increase of the variance σ_k, corresponding to the flattening of the Gaussian $\mathcal{N}(\mu_k, \sigma_k)$ has two beneficial effects: 1) the possibility of matching between Gaussians of important states of different models is increased; 2) Gaussians of not important states are very flattened, reducing their contributions to the probability computation. It is worthwhile to notice that such a metric is able to remove moving objects from the video sequence, as they are considered non stationary components of the temporal evolution of the pixel.

Assumed this kind of similarity measure between sequences, the spatio-temporal segmentation is developed as a typical segmentation process of static images, and a simple region-growing algorithm has been adopted. The first step presumes to rank the pixel locations by their importance. For each pixel i, the importance is measured by $\max_{1 \le k \le N} \mathbf{p}_\infty(k)$, where $\mathbf{p}_\infty$ is the stationary probability distribution of the HMM associated to the pixel i. Starting from the most important pixel, a simple region growing process is applied, using a threshold θ on the distance $D_{\mathrm{ES}}(i, j)$ to estimate when two adjacent sequences are similar. When the growing process stops, we choose as new seed the subsequent most important pixel, not belonging to an existent segmented region. The process ends when all the pixels have been labelled. We will see in the experimental session that the modification of the metric in eq. (2) together with the integration of the spatial-temporal information of the video sequence, lead to a visible improvement of the segmentation results in both synthetic experiments and real sequences.

4. EXPERIMENTAL TRIALS

In this section, some experimental results are presented. As first step, a subsequence of 1/4 of the length of the original video sequence is considered. This fraction is modelled by the proposed HMM-based initialization method.

From this probabilistic model, we could infer both the pixel- and the region-level initialization of the background modelling. After presenting few results on the pixel-level initialization, we will show some synthetic and real examples for the region-level initialization (spatio-temporal segmentation), which represents the most innovative part of our work.

4.1 Pixel-level initialization

The proposed approach was tested in a real case, concerning an indoor sequence, with a corridor in which several doors are present. A person is present in the scene, walking around the corridor. Some frames of the sequence are presented in Fig. 1. In Fig. 2 the median frame of the sequence is displayed. This image does not contain the moving objects, so the initialization could be considered correct. We apply our algorithm to the video sequence, obtaining a pixel level initialization. In the training step we have fixed the number of states of each HMM to 3, with the aim of permitting a sufficient expressive explanation of the chromatic behavior of each pixel of the sequence.
After the training process, for each pixel i, we obtain the HMM model $\boldsymbol{\lambda}_i$ from which we consider the 3 different Gaussian parameters (mean and variance), i.e. the HMM states, and the 3 corresponding mixing coefficients, which are the different stationary probabilities. In order to graphically show the representation obtained, we build an image in which for each pixel we consider the sum of all the 3 means, weighted by the related stationary probabilities. The resulting representation is displayed in Fig. 3. One can notice that also our representation is correct: by considering only the stationary part of the background signal, all the moving objects have been removed, so our approach is able to correctly initialize the pixel-level background. Actually, the real advantage of our representation is indeed expressed in the region-level initialization, in which the spatial and temporal information becomes crucial in order to correctly identify all the semantically different regions of the scene.

4.2 Region-level initialization

The proposed region level initialization is tested, using synthetic and real experiments. The input for the testing is a video sequence, and the wanted output is a segmentation of the background, in which all the semantically different regions of the scene are captured. The proposed approach returns a spatio-temporal segmentation of the background, representing the region-level initialization of the background model. It is important to note that, like the pixel-level initialization, also the region-level initialization is recovered from the unique probabilistic modelling of the sequence.

The testing was performed using both synthetic and real sequences. In the former case, the synthetic sequence, containing blocks flickering with the same color palette but with different frequency, is shown in Fig. 4 (the central region is stable). In Fig. 5(a), the resulting segmentation is presented, showing that all 9 regions are correctly identified by our algorithm. In order to explicitly assess the advantage owned by the use of temporal information of our spatio-temporal segmentation, we also present results obtained by a simpler classic segmentation algorithm, applied on the averaged image. In this case, after obtaining the mean image by averaging the gray level values of all frames of the sequence, we applied a region-growing algorithm similar to

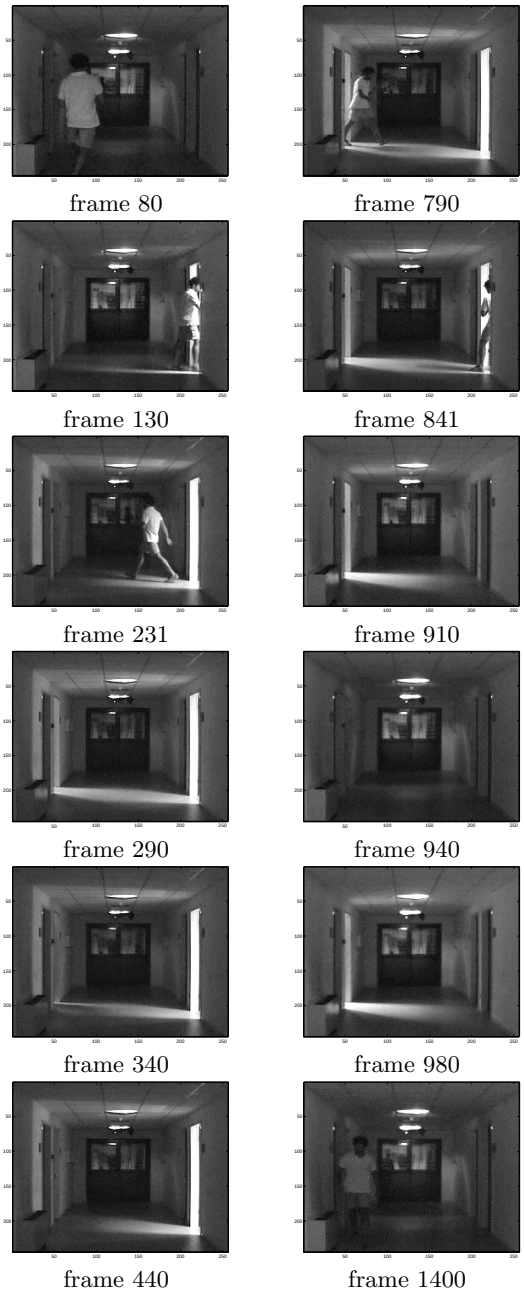


Figure 1: Frames of the corridor sequence.

that used in our algorithm. Results are shown in Fig. 5(b), in which it is evident that this method is not able to capture the temporal diversity between the pixels of the regions. In other words, not all the semantically different regions of the scene have been discovered, and only five regions have been detected. The above comparison is in some way improper as it evident that a segmentation method which does not take into account temporal information is likely to fail with respect to our proposed algorithm. In fact, to the best of our knowledge, this is the first segmentation method that exploits the temporal gray-level behavior in order to obtain a single spatio-temporal segmentation of a scene. To

Figure 2: Median Frame

Figure 3: Pixel level initialization using the proposed approach.

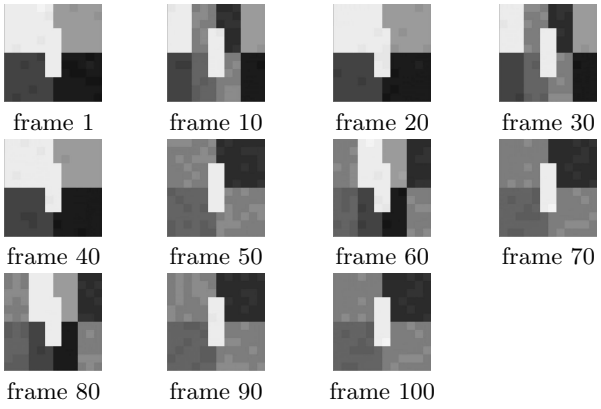


Figure 4: Synthetic sequence used for testing spatio-temporal segmentation.

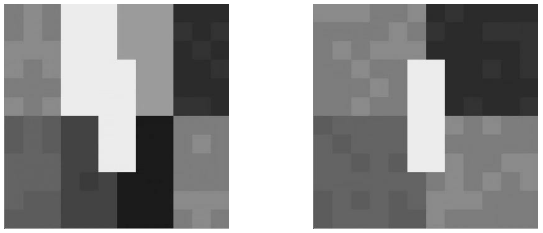


Figure 5: Segmented sequence obtained by (a) the proposed approach (b) a region growing method onto the averaged image.

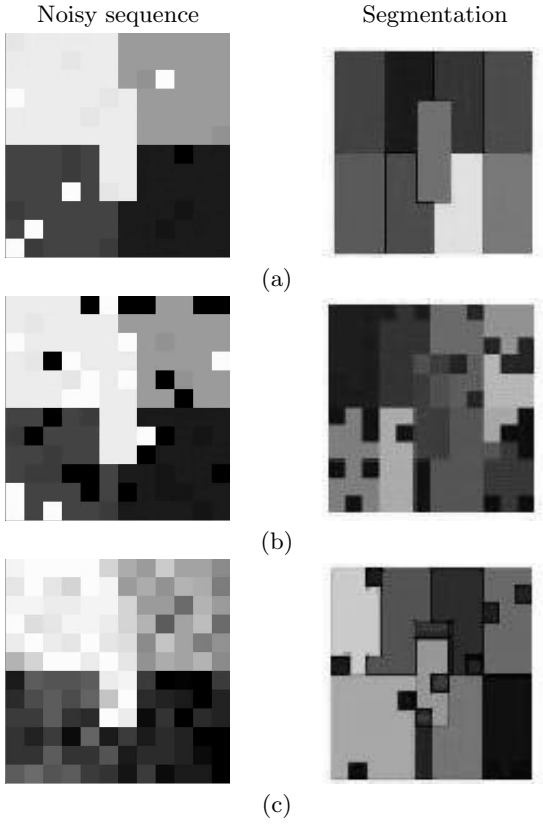


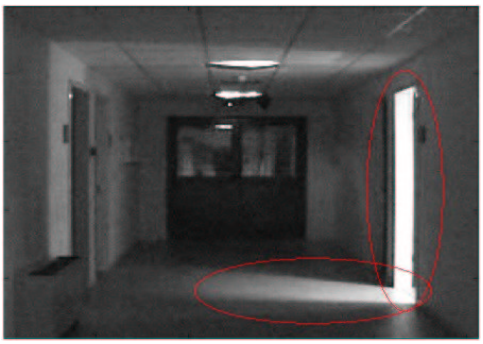
Figure 6: Synthetic experiment with added noise. In the left column, a frame from the noisy sequence, and, in the right column, the resulting segmentation, for different noise type and level. (a) Salt & Pepper, density 5%, (b) Salt & Pepper, density 25%, and (c) Gaussian, variance 0.01.

some extent, our approach shares some intuitions of the so-called JSEG algorithm [6], in which a spatial segmentation is performed in the initial frame, and is propagated in the subsequent frames using temporal constraints. The difference is that JSEG uses only the frame by frame temporal information in order to obtain a series of spatial segmentated images (one for each frame), whereas our approach considers the sequence as a whole in order to get a single summarizing spatio-temporal scene representation.

To assess the robustness of our approach to noisy sequences, we add two types of synthetic noise: a Salt & Pepper noise, of density 5% and 25%, and a white Gaussian noise, of variance 0.01. An example of a noisy frame and the corresponding sequence segmentation are presented in Fig. 6 for all noisy situations. As one can notice, our approach is quite robust to recover from both types of noise: even if the sequence is quite corrupted, the different semantic regions are identified rather well.

The proposed approach has been tested also with some

real sequences. The first example regards the sequence presented in the previous subsection, and is aimed at explaining how the temporal information used in the proposed segmentation could be useful in order to solve the problem of a sudden change of illumination at region level. Looking at the Fig. 1, you can notice that during time some doors were opened and closed several times, each one with a random different frequency. The action of opening-closing a door determines a local variation of the illumination, *i.e.*, there are two particular regions of the corridor in which the illumination changes with different frequencies. These different spatial chromatic zones are highlighted in Fig. 7: one is on the left part of the corridor, and the other on the right part. Considering only the median (or the mean) of the

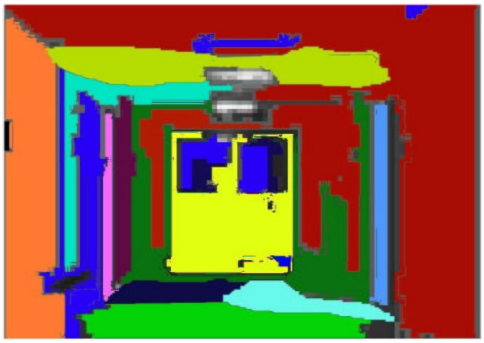


(a)

(b)

Figure 7: Different spatial chromatic zones.

sequence, displayed in Fig. 2, it is not possible to detect the two semantically different zones of the background. Actually, any spatial segmentation technique applied to the image segments the zone between the two doors as belonging to the same region. In Fig. 8, a comparative result between the segmentation resulting from our approach and an ordinary region growing based segmentation on the median image is shown. One can easily notice that our approach clearly separates the two zones, labelled as different regions of the scene. This is important since the integrated pixel- and region-based approach to background modelling uses the region information derived from the segmentation of the background as the modulating information. In order to assess the gain obtained with the Enhanced Stationary similarity measure, the segmentation of the corridor sequence based on the measure of the eq.(2) is depicted in Fig. 9. It is evident that the noise of the sequence and the presence of foreground produce a very noisy over-segmentation. Actu-



(a)

(b)

Figure 8: Segmentation of the corridor sequence (a) using the proposed approach, and (b) using an ordinary method of region growing based segmentation.

ally, only if the segmentation is able to correctly detect all the semantically different regions of the scene, the method can correctly work. It has been shown in [5] that, using this segmentation, a video surveillance system is able to correctly track objects also in presence of sudden non-uniform changes of illumination.

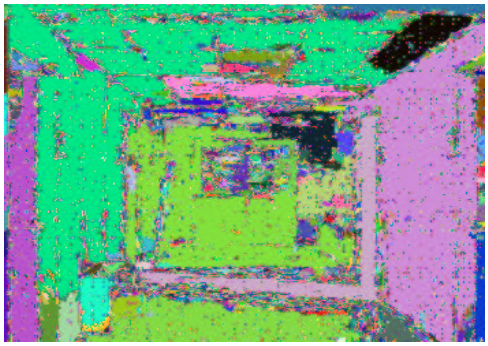


Figure 9: Segmentation of the corridor sequence using the HMM similarity measure without the flattening of the non stationary states.

We perform a further test on another sequence, consisting in two moving objects in a outdoor scene. A few frames of the sequence are presented in Fig. 10, and the resulting segmentation is proposed in Fig. 11(a). The resulting segmentation is clear, expressive, and quite accurate, even more

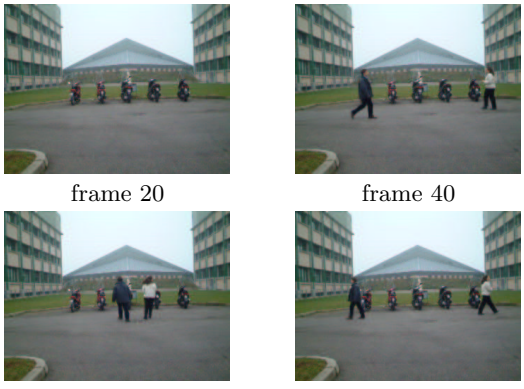


Figure 10: Few frames from the outdoor sequence.

valuable if one notices that it is obtained by processing the whole sequence, without any need to remove the moving objects, as they are naturally taken out by the procedure used to compute the distance. This result was compared to that obtained by an ordinary region growing approach performed on the first frame, presented in Fig. 11(b). The result of the

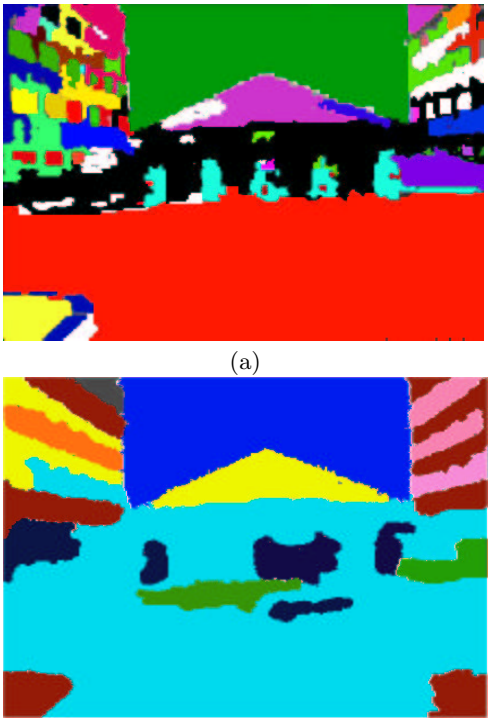


Figure 11: Segmentation of the outdoor sequence:(a) proposed approach, (b) standard region growing method.

proposed approach appears more accurate, in particular in the ground in front of the scooters and in the segmentation of the windows of the two lateral buildings.

5. CONCLUSIONS AND FUTURE WORK

In this paper, a novel algorithm for background initialization is proposed, able to characterize the chromatic and spatial behavior of a scene at pixel- and region-levels using an arbitrary uncontrolled training sequence. The process realizes a probabilistic modelling of the video sequence using a battery of Hidden Markov Models (HMM), modelling the gray intensity values assumed by each pixel as a set of independent process. From this probabilistic representation it is possible to initialize the background modelling scheme at both pixel- and region-levels. The former is obtained by observing the stationary probability distribution of each HMM, in order to infer the most stable gray intensity values. The latter is obtained by clustering the HMMs, using a novel similarity measure between HMMs. This measure, together with a region growing process, couples neighboring pixels that exhibit a similar chromatic-temporal behavior, providing a segmented image. In this image each cluster indicates a spatial region with a homogeneous gray level that changes its intensity similarly along time. The proposed measure is also able to consider only stationary components of the scene, removing possible moving objects.

The main drawback of this method is the strong computational effort, but an off-line computation for an initialization algorithm is indeed allowed. Nevertheless, a parallel computational architecture may solve this problem, permitting a very quickly and useful batch mode scheme.

Finally, the use of an on-line HMM training [22] together with a parallel architecture would provide the first effective background modelling algorithm based on region information only, able to identify the foreground using regions as elementary units. This is our main interest for the progress of this work.

6. REFERENCES

[1] C. Bahlmann and H. Burkhardt. Measuring hmm similarity with the bayes probability of error and its application to online handwriting recognition. In *Proc. Int. Conf. Document Analysis and Recognition*, pages 406–411, 2001.

[2] M. Bicego, V. Murino, and M. Figueiredo. A sequential pruning strategy for the selection of the number of states in Hidden Markov Models. *Pattern Recognition Letters*, 24(9–10):1395–1407, 2003.

[3] M. Bicego, V. Murino, and M. Figueiredo. Similarity-based clustering of sequences using hidden Markov models. In P. Perner and A. Rosenfeld, editors, *Machine Learning and Data Mining in Pattern Recognition*, volume LNAI 2734, pages 86–95. Springer, 2003.

[4] P. Brémaud. *Markov Chains*. Springer-Verlag, 1999.

[5] M. Cristani, M. Bicego, and V. Murino. Integrated region- and pixel-based approach to background modelling. In *Proc. of IEEE Workshop on Motion and Video Computing*, pages 3–8, 2002.

[6] Y. Deng and B. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(8):800–810, 2001.

[7] N. Friedman and S. Russell. Image segmentation in video sequences: A probabilistic approach. In *Uncertainty in Artificial Intelligence: Proceedings of*

the Thirteenth Conference (UAI-1997), pages 175–181, San Francisco, CA, 1997. Morgan Kaufmann Publishers.

[8] X. Gao, T. Boult, F. Coetzee, and V. Ramesh. Error analysis of background adaption. In *Proc. of IEEE Conf. on Computer Vision Pattern Recognition*, volume I, pages 503–510, 2000.

[9] B. Gloyer, H. Aghajan, K.-Y. Siu, and T. Kailath. Video-based freeway monitoring system using recursive vehicle tracking. In C. M. Bishop and B. J. Frey, editors, *Proc. SPIE - The International Society for Optical Engineering*, volume 2421, pages 747–757, 1995.

[10] B. Gloyer, H. K. Aghajan, K. Y. Siu, and T. Kailath. Video-based freeway monitoring system using recursive vehicle tracking. In *IS&T-SPIE Symposium on Electronic Imaging: Image and Video Processing*, 1995.

[11] D. Gutchessy, M. Trajkovicz, E. Cohen-Solalz, D. Lyonsz, and A. K. Jain. A background model initialization algorithm for video surveillance. In *Proc. of IEEE Conf. on Computer Vision*, 2001.

[12] I. Haritaoglu, D. Harwood, and L. Davis. W^4: real-time surveillance of people and their activities. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):809–830, 2000.

[13] M. Harville. A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models. In *European Conf. Computer Vision*, volume 3, pages 543–560, 2002.

[14] O. Javed, K. Shafique, and M. Shah. A hierarchical approach to robust background subtraction using color and gradient information. In *Proc. of IEEE Workshop on Motion and Video Computing*, pages 22–27, 2002.

[15] M. Law and J. Kwok. Rival penalized competitive learning for model-based sequence. In *Proc. Int. Conf. Pattern Recognition*, volume 2, pages 195–198, 2000.

[16] C. Li and G. Biswas. A bayesian approach to temporal data clustering using hidden Markov models. In *Proc. Int. Conf. on Machine Learning*, pages 543–550, 2000.

[17] C. Li and G. Biswas. Applying the Hidden Markov Model methodology for unsupervised learning of temporal data. *Int. Journal of Knowledge-based Intelligent Engineering Systems*, 6(3):152–160, 2002.

[18] W. Long and Y. Yang. Stationary background generation: An alternative to the difference of two images. *Pattern Recognition*, 23:1351–1359, 1990.

[19] N. Ohta. A statistical approach to background subtraction for surveillance systems. In *Int. Conf. Computer Vision*, volume 2, pages 481–486, 2001.

[20] A. Panuccio, M. Bicego, and V. Murino. A Hidden Markov Model-based approach to sequential data clustering. In T. Caelli, A. Amin, R. Duin, M. Kamel, and D. de Ridder, editors, *Structural, Syntactic and Statistical Pattern Recognition*, LNCS 2396, pages 734–742. Springer, 2002.

[21] M. Petkovic and W. Jonker. Content-based video retrieval by integrating spatio-temporal and stochastic recognition of events. In *Proc. of IEEE Workshop on Detection and Recognition of Events in Video*, pages 75–82, 2001.

[22] N. Petrovic, N. Jojic, B. J. Frey, and T. S. Huang. Real-time on-line learning of transformed hidden markov models from video. In C. M. Bishop and B. J. Frey, editors, *Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[23] L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. of IEEE*, 77(2):257–286, 1989.

[24] P. Smyth. Clustering sequences with hidden Markov models. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing*, volume 9. MIT Press, 1997.

[25] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Int. Conf. Computer Vision and Pattern Recognition*, volume 2, 1999.

[26] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[27] B. Stenger, V. Ramesh, N. Paragios, F.Coetzee, and J. M. Buhmann. Topology free hidden Markov models: Application to background modeling. In *Int. Conf. Computer Vision*, volume 1, pages 294–301, 2001.

[28] H. Tao, H. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):75–89, 2002.

[29] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Int. Conf. Computer Vision*, pages 255–261, 1999.

[30] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[31] Y. Yacoob and M. Black. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding: CVIU*, 73(2):232–247, 1999.