Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Vision and Image Understanding 102 (2006) 22–41

Computer Vision and Image Understanding

www.elsevier.com/locate/cviu

Unsupervised scene analysis: A hidden Markov model approach

Manuele Bicego [1], Marco Cristani, Vittorio Murino *

Dipartimento di Informatica, University of Verona, Ca' Vignal 2, Strada Le Grazie 15, 37134 Verona, Italy

Received 23 November 2004; accepted 22 September 2005
Available online 8 November 2005

Abstract

This paper presents a new approach to scene analysis, which aims at extracting structured information from a video sequence using directly low-level data. The method models the sequence using a forest of Hidden Markov models (HMMs), which are able to extract two kinds of data, namely, *static* and *dynamic* information. The static information results in a segmentation that explains how the chromatic aspect of the static part of the scene evolves. The dynamic information results in the detection of the areas which are more affected by foreground activity. The former is obtained by a spatial clustering of HMMs, resulting in a spatio-temporal segmentation of the video sequence, which is robust to noise and clutter and does not consider the possible moving objects in the scene. The latter is estimated using an entropy-like measure defined on the stationary probability of the Markov chain associated to the HMMs, producing a partition of the scene in activity zones in a consistent and continuous way. The proposed approach constitutes a principled unified probabilistic framework for *low level* scene analysis and understanding, showing several key features with respect to the state of the art methods, as it extracts information at the lowest possible level (using only pixel gray-level temporal behavior), and is unsupervised in nature. The obtained results on real sequences, both indoor and outdoor, show the efficacy of the proposed approach.

Keywords: Scene analysis; Video processing; Hidden Markov models; Video segmentation; Scene understanding; Video surveillance

1. Introduction

Video analysis and understanding is undoubtedly an important research area, whose interest has grown in the last decade, promoting a set of interesting applications, each one characterized by different goals. For instance, video summarization [1,2] aims at subdividing a video in significant shots which characterize it overall. In video retrieval by content [3,4] the target is to retrieve videos from a database only on the basis of its content, trying to identify it using some specific features. Another class of approaches has increased its importance in the last few years, generally grouped under the name of "scene understanding," whose aim is to infer knowledge about a scene, easily interpretable by a human operator by analyzing a video sequence. An example is represented by video surveillance approaches [5,6], where the goal is to find "atypical" situations and behaviors in an outdoor/indoor environment. Furthermore, one can go beyond the detection of such situations, by carrying out higher-level analysis in several ways. In this class, methods classifying the activities occurring in a scene by analyzing the object trajectories have been proposed [7,8], other approaches studied the interactions between the objects in a scene [9–12]; modelling and synthesis of general complex behaviors have been also devised [13,48,49].

The key concept under all these approaches is the "learning," i.e., the capability of gaining such knowledge by training a particular model on the basis of the information extracted from a video sequence. In this way, the trained model can be subsequently used to generalize to other situations. All these approaches can be classified as *generative models* [14,15], in which the goal is to develop flexible models that can explain (*generate*) visual input as

* Corresponding author. Fax: +39 045 8027068.
E-mail addresses: bicego@uniss.it (M. Bicego), cristanm@sci.univr.it (M. Cristani), vittorio.murino@univr.it (V. Murino).

[1] Present address: DEIR, University of Sassari, via Torre Tonda 34, 07100 Sassari, Italy.

1077-3142/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2005.09.001

a combination of hidden variables and can adapt to new input types. A recent and complete review on generative models for scene analysis is presented in [14].

In this paper, we propose a computational framework aimed at processing low level data (the pixel gray levels) coming from a typical video surveillance sequence in which the camera is in a fixed location. The final aim is to provide knowledge usable to:

(1) draft a description about how the background evolves (e.g., periodic chromatic fluctuations possibly due to local/global illumination changes): this is accomplished via a segmented image in which each region corresponds to a compact patch of *background* pixels with similar gray level *and* with similar time-chromatic behavior[1];
(2) detect the degree of foreground activity in the scene: this is performed by an activity map, based on an entropy-driven measure;
(3) improve the preprocessing steps of the classical video surveillance flowchart, as background initialization [22,23], and other related tasks (listed in Section 4.4).

The first two points encourage us to define the proposed analysis as *low-level scene understanding*. This term highlights the fact that the analysis is performed on rough video data and the fact that, taken as independent process, our approach is able to easily extract human-interpretable knowledge about an observed scene.

This method presents several key characteristics that differentiate it from the state of the art, which will be reported and justified in the following.

Actually, many approaches in the literature [16,17,48] base their analysis on typical and well known operations, i.e., segmentation and tracking. Typically, these procedures are adequate when a priori knowledge is available (for example, the shape of an object to be identified, the number of objects to be tracked, the location of appearing/disappearing objects), but they are weak when this information is not provided. This problem occurs, for instance, when a camera is monitoring a crowded scene, in which multiple occlusions and clutter are present.

Our approach circumvents this problem by performing an analysis at the lowest level, i.e., by considering directly and only the temporal pixel-level behavior.

A somewhat similar idea is at the basis of the methods proposed in [18–21,54], in which the extraction of semantic information is carried out without segmentation or trajectory extraction, but performing low- (pixel) and mid-level (blob) analysis, after a background analysis modelling. Nevertheless, our approach is only similar in spirit to the above quoted work since, unlike those approaches, only low-level analysis at pixel level is performed and a different probabilistic method is used.

Another important characteristic of the approach regards the modelling tool used to analyze the video sequence. The sequence is modelled using a forest of Hidden Markov models (HMMs) [24], each one modelling the temporal evolution of a pixel. HMMs represent a widely employed generative model for probabilistic sequential data modelling [24], also used in the context of visual processing, using either the basic structure [48] or extensions (like transformed HMM [1], distance HMM [53], or dynamically multi-linked HMM [54]). An interested reader may refer to [51] for a complete review of the use of the HMMs in computer vision and other applications, and more in general to [14] for the use of the generative graphical models in the context of learning and understanding scene activity.

The popularity of HMMs derives from three appealing characteristics: the intrinsic capability to deal with sequential evolution, the effective and fast training algorithm (derived from the expectation maximization (EM) [27,28]), and the clear Bayesian semantic interpretation. The HMMs appear a suitable choice for our task: they represent a good combination between expressivity power and low computational complexity; in addition, we developed HMM clustering techniques that allow to infer similarity degrees among models, while exploiting inter-pixels analysis.

The main difference between our approach and those presented in the related literature based on HMMs is that in our case these models are used at the *lowest possible level* to directly model the evolution of each pixel gray level in a scene, rather than modelling high-level structured objects. In this sense, the modelling used by our approach is similar to that used in [29,30], where the temporal evolution of each pixel is modelled by a HMM. However, in these cases, the aim was not to infer knowledge from a sequence, but only to realize a robust background modelling module.

Moreover, our approach is inherently unsupervised in the sense that no learning step is necessary *before* processing the video data. The HMMs are actually trained using exactly the video sequence to analyze, and the inferences over the trained models provide the results of the analysis.

Finally, another characteristic feature of the proposed approach concerns its versatility, being able to contemporarily infer two types of data about a scene, namely *static* and *dynamic* information (Fig. 1). The terms static and dynamic indicates that we are extracting information about entities that are static or dynamic in a spatial sense.

The former gives some insight about the chromatic behavior of the static part of a scene, i.e., the background, whereas the latter aims at discovering the extent of foreground activity in the observed scene.

In this paper, we consider as "activity" a temporal pattern which cannot be classified as regular, or, in other words, which has not a predictable behavior. The static

[1] In the paper, the use of the words *chromo* or *chromatic* concerns the aspects related to the *gray-level* values of the image pixels as all the processed sequences are converted in gray-level values. Nevertheless, the extension to the RGB scale is straightforward, and does not raise particular issues to the proposed method.

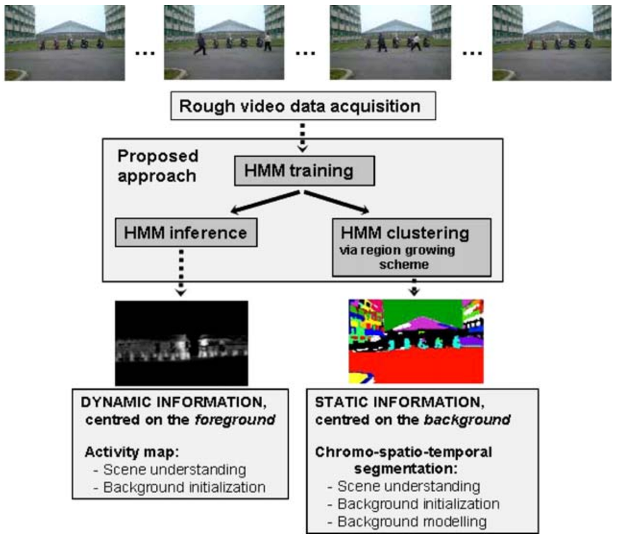

...
...
...
Rough video data acquisition
Proposed approach
HMM training
HMM inference
HMM clustering via region growing scheme
DYNAMIC INFORMATION, centred on the *foreground*
Activity map:
- Scene understanding
- Background initialization
STATIC INFORMATION, centred on the *background*
Chromo-spatio-temporal segmentation:
- Scene understanding
- Background initialization
- Background modelling

Fig. 1. Input and output of the proposed analysis: starting from the rough pixel data (top), the analysis performed can be represented using two images (bottom). On the left, the activity map, indicating the total amount of foreground activity in the scene (dynamic information). On the right, the chromo-spatio temporal segmentation, that describes how the chromatic aspect of the background evolve (static information): each region is characterized by a homogeneous gray-level *and* a similar temporal evolution. In the bottom of both the boxes, the related applications to which our analysis can be devoted, described in Section 4.4.

information is extracted by performing a chromo-spatio-temporal segmentation of the background, obtained by clustering the pixelwise HMMs. To this end, a new similarity measure between HMMs is proposed, able to remove non-stationary components of the sequence. Using this measure and a simple region-growing procedure, a segmentation of the scene is obtained in which the regions show a homogeneous gray level *and* a similar temporal evolution. In this case, the resulting segmentation is a spatial segmentation of the scene, obtained by using all available information: chromatic (different regions have different gray-level values), spatial (each region is connected in the image space), and temporal (each region varies its color homogeneously along time). Actually, our approach has two main advantages: first, the spatial knowledge, typically used to obtain standard segmentation, is augmented with temporal information. This is useful to discover, in a region with homogeneous color, additional subregions subjected to periodic chromatic fluctuations (caused for example by changes of illumination). Therefore, this segmentation could generally appear as over-segmented, but each region is however meaningful for the addressed task. The second advantage is that moving objects have not to be removed from the sequence (as in the single image segmentation), since this operation is automatically accomplished by the envisioned similarity measure.

The dynamic information is obtained by looking at the model parameters, and by inferring which pixels are mainly dealt with foreground "activity." To compute this activity measure, the stationary probability distribution of the Markov chain associated with the HMMs is used. By visualizing all these measures, we estimate the "activity map" of the scene, in which are recognizable the areas that are more engaged with activities, i.e., more affected by foreground motion.

The proposed approach has been tested using real experiments, showing that it represents a useful tool for scene analysis, which, starting from the lowest-level data representation, is able to support and increase the understanding of a monitored scene, as either an independent analysis module or embedded in a classical video-surveillance framework. Other perspective applications of this information are proposed at the end of the paper.

Summarizing, the main features of the proposed approach are: (1) scene analysis is carried out at the lowest possible level by directly processing the temporal behavior of the pixels' values, and without resorting to an explicit segmentation of moving objects; (2) this analysis is useful per se or embedded in a typical video surveillance structure as a preprocessing step (see Section 4.4); (3) the training phase represents the core of the process, from which follows the unsupervised character of the analysis; (4) the approach is based on a unified probabilistic framework, the hidden Markov modelling, which is able to simultaneously derive both static and dynamic information.

The rest of the paper is organized as follows. In Section 2, the basic principles of the Hidden Markov models are presented. The proposed strategy is then detailed in Section 3, and extensive experimental results and a comparative analysis are presented in Section 4, showing the superiority of the proposed approach. Finally, in Section 5, conclusions are drawn and future perspectives are envisaged.

2. Fundamentals

In this section, the fundamental instruments of the proposed approach are described. In particular, in Section 2.1 the definition of the Hidden Markov model approach is given, while in Section 2.2 the concept of stationary probability of a HMM, representing a key entity in the approach proposed in this paper, is introduced. Finally, Section 2.3 contains the description of the HMM-based clustering approach.

2.1. Hidden Markov models

A Hidden Markov model $\boldsymbol{\lambda}$ can be viewed as a Markov model whose states cannot be explicitly observed: each state has associated a probability distribution function, modelling the probability of emitting symbols from that state. The HMM methodology is not exhaustively described in this paper, and interested readers are referred to [24]. Briefly, a HMM is defined by the following entities:

- $S = \{S_1, S_2, \ldots, S_N\}$ the finite set of the possible hidden states;
- the transition matrix $\mathbf{A} = \{a_{ij}, 1 \leqslant j \leqslant N\}$ representing the probability of going from state S_i to state S_j

$$a_{ij} = P[S_i \to S_j], \quad 1 \leqslant i, \quad j \leqslant N$$

with $a_{ij} \geqslant 0$ and $\sum_{j=1}^{N} a_{ij} = 1$;

- the emission matrix $\mathbf{B} = \{b(o|S_j)\}$, indicating the probability of the emission of the symbol o when system state is S_j. In this paper continuous HMMs are employed, hence $b(o|S_j)$ is represented by a Gaussian distribution, i.e.

$$b(o|S_j) = \mathcal{N}(o|\mu_j, \Sigma_j), \tag{1}$$

where $\mathcal{N}(o|\mu, \Sigma)$ denotes a Gaussian density of mean μ and covariance Σ, evaluated at o;

- $\boldsymbol{\pi} = \{\pi_i\}$, the initial state probability distribution, representing probabilities of initial states, i.e.

$$\pi_i = P[q_1 = S_i], \quad 1 \leqslant i \leqslant N$$

with $\pi_i \geqslant 0$ and $\sum_{i=1}^{N} \pi_i = 1$.

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

The training of the model, given a set of sequences $\{\mathbf{O}_i\}$, is usually performed using the standard Baum–Welch re-estimation technique [24], able to determine the parameters $(\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ that maximize the probability $P(\{\mathbf{O}_i\}|\boldsymbol{\lambda})$. The evaluation step, i.e., the computation of the probability $P(\mathbf{O}|\boldsymbol{\lambda})$, given a model $\boldsymbol{\lambda}$ and a sequence $\mathbf{O}$ to be evaluated, is performed using the *forward–backward procedure* [24].

2.2. The stationary probability distribution

In this section, the stationary probability distribution of a HMM is defined, which represents the core of our approach.

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

$$\begin{aligned}\mathbf{p}_t &= [\mathbf{p}_t(1), \ldots, \mathbf{p}_t(j), \ldots, \mathbf{p}_t(N)] \\ &= \left[P(Q_t = S_1), \ldots, P\left(Q_t = S_j\right), \ldots, P(Q_t = S_N)\right],\end{aligned}$$

where $\mathbf{p}_t(i)$ represents the probability of being in state S_i at time t. Obviously, $\mathbf{p}_t$ 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 infinite 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 [31] 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. Finding $\mathbf{p}_\infty$ for a given $\mathbf{A}$ then amounts to solving the corresponding eigenvalue/eigenvector problem.

2.3. HMM-based clustering

HMMs have not been extensively employed for clustering sequences, so that only a few papers exploring this direction appeared in the literature. Even if some alternative approaches to HMM-based clustering have been proposed (e.g. [32,33]), the typically employed method is the so-called proximity-based approach [50], which uses the HMM modelling to compute distances between sequences, using another standard approach based on pairwise distance matrices (as hierarchical agglomerative) to obtain clustering [34–37]. The distance between sequences is typically based on the likelihood of the HMM, and could be obtained using several methods (for example, see [37,38,48]).

In more detail, given a set of R sequences $\{\mathbf{O}_1, \ldots, \mathbf{O}_R\}$, the standard approach to clustering trains one HMM $\boldsymbol{\lambda}_i$ for each sequence $\mathbf{O}_i$. Subsequently, a pairwise distance between sequences is defined using these models, in which the key entity is the likelihood L_{ij}, defined as $L_{ij} = P(\mathbf{O}_j|\boldsymbol{\lambda}_i)$. This probability is used to devise a distance (or a similarity) measure between sequences. The simplest example has been proposed in [50], and is defined as

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

A more complex one, inspired from the Kullback–Leibler measure [39] and proposed in [37], is defined as

$$D(i,j) = \frac{1}{2}\left\{\frac{L_{ij} - L_{jj}}{L_{jj}} + \frac{L_{ji} - L_{ii}}{L_{ii}}\right\}. \tag{3}$$

Once given these distances, any standard pairwise distance-based clustering algorithm could be used, such as those belonging to the hierarchical agglomerative family.

In Section 3.2, we will see how this standard method could be extended to deal with spatial segmentation, which represents a particular kind of clustering.

3. The proposed approach

In this section the proposed approach is presented. In Section 3.1, the probabilistic modelling of the sequence is introduced, while the following two sections describe how this representation is used to infer static (Section 3.2) and dynamic (Section 3.3) information about the scene.

3.1. The probabilistic modelling of video sequences

The proposed approach models the whole 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 a scene (since the camera is fixed). Given this set of sequences, we want to model them to capture their most important characteristics. In particular, we need a model able to determine: (1) the most stable gray-level components measured in the whole sequence; (2) the temporal chromatic variation of these components; (3) the sequentiality in which the components vary. An adequate computational framework showing these features is constituted by the Hidden Markov model (HMM) [24]. Using this model, all the above requirements can be accomplished. In particular, using HMMs with continuous Gaussian emission probability, the most important gray-level components are modelled by the means $\boldsymbol{\mu}_i$ of the Gaussian functions associated to the states, the variability of those components are encoded in the covariance matrices $\boldsymbol{\Sigma}_i$, and the sequentiality is encoded in the transition matrix $\mathbf{A}$. The HMM methodology has been preferred to other similar modelling techniques, such as Gaussian Mixture models (GMM [40]), due to its important characteristic of being able to deal with the temporal sequentiality of the data, which is crucial when analyzing video sequences. GMMs are indeed not able to capture the temporal variability, i.e., the model does not change if the frames of the video-sequence are randomly shuffled, as temporal information is not considered.

Summarizing, the sequence is modelled using a forest of HMMs, one for each pixel. For what concerns the model selection, the different approaches for determining the number of states of a HMM directly from data (e.g. [41–44]) are typically computationally demanding. Since the proposed approach trains one HMM per pixel, we have chosen to fix a priori the number of states to maintain the computational effort at a reasonable level. This choice is not critical and can be guided from opportune considerations about the complexity of the scene, especially in relation to the complexity of the background. Actually, three states are considered a reasonable choice, taking into account the possibility of a bimodal BG, and one component for the foreground activity [46].

Once fixed the number of states, the HMM training has been carried out using the standard Baum–Welch procedure, paying particular attention to the initialization. Since the Baum–Welch procedure, starting from some initial estimates, converges to the nearest local maximum of the likelihood function, which is typically highly multi-modal, the initialization issue is particularly crucial for the effectiveness of the training. In our approach, a Gaussian Mixture model (GMM) [40] clustering is used to initialize the emission matrix of the HMM before training. In particular, the initialization phase first considers the sequence of pixel gray levels as a set of scalar values (no matter in which order the coefficients appear); second, these values are grouped into three clusters by following a GMM clustering approach, i.e., fitting the data by using three Gaussian distributions, in which the Gaussian parameters are estimated by an EM-like [27,28] method. Finally, the mean and variance of each cluster are used to initialize the Gaussian of each state, with a direct correspondence between clusters and states.

The computational complexity of the training phase is $O(nI_{\max}N^2T)$, where n is the number of the pixels, $I_{\max}N^2T$ is due to the standard complexity of the Baum–Welch training phase for each pixel; $I_{\max}$ is the maximum number of iterations permitted during the learning step, N^2T is due to the forward and backward variables calculation, where N is the number of the states and T is the length of the sequence.

3.2. Static information: the spatio-temporal segmentation

The first kind of information extracted with the proposed approach is a *static* information, that provides knowledge about the structure of the scene. The probabilistic representation of the video sequence is used to obtain a "chromo-spatio-temporal segmentation" of the background. In other words, we want to segment the background of the video sequence in regions showing a homogeneous color and a similar temporal evolution, considering pixel-wise information. In this case, the result is a spatial segmentation, obtained by using all available information: chromatic (different regions have different gray-level values), spatial (each region is connected in the image space), and temporal (each region varies its color similarly along time). In this way, spatial knowledge, typically used to obtain spatial segmentation, is augmented with temporal information, allowing a more detailed and informative partitioning.

The proposed HMM representation implies 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 basic idea is to define a distance between locations on the basis of the distance between the trained Hidden Markov models: in this way the segmentation process is obtained using a spatial clustering of the HMMs. 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 important components. By using the measure proposed in Eqs. (2) or (3), we have that the Gaussian of each state contributes in the same way at the computation of the probability, because of the forward–backward procedure. For our goal, however, we need that the Gaussian of each state contributes differently to the probability computation, depending on the "importance" of the corresponding state.

To this end, we have regularized the HMMs' states S_i, for every HMMs, with respect to the related $\mathbf{p}_\infty(i)$, which is a quantitative index of the state importance. Actually, $\mathbf{p}_\infty$ indicates the "average" occupation of each state, after the Markov chain has achieved the stationary state [44], hence, it represents the degree of saliency associated to the states. This operation allows to normalize the behavior of the several HMMs so as to allow an effective and reliable comparison between them.

The normalization operation is carried out by operating on the Gaussian parameters of each state, in particular, each original model $\boldsymbol{\lambda}$ is transformed into a new model $\boldsymbol{\lambda}'$, where all components remain unchanged, except variances σ_i of state S_i, for each state $i = 1, \ldots, N$, for all HMMs, i.e.

$$\sigma_i' = \frac{\sigma_i}{\mathbf{p}_\infty(i)}. \tag{4}$$

The new distance, called D_{ES} (Enhanced Stationary), is then computed using Eq. (3) on the modified HMMs $\boldsymbol{\lambda}_k'$. The normalization of the state variances σ_i with respect to the related $\mathbf{p}_\infty(i)$, corresponds to associate the correct significance to the Gaussian $\mathcal{N}(\mu_i, \sigma_i^2)$, and has two beneficial effects: (1) Gaussians of unimportant states are undergraded, reducing their contribution to the probability computation, which results in eliminating moving objects from the video sequence, as they are considered as non-stationary components of the background model; (2) the possibility of match between Gaussians of important states of different models is increased. These concepts are exemplified in Fig. 2.

Assuming this kind of similarity measure between sequences, the segmentation process can be developed as an ordinary segmentation process of static images. We adopt a simple region growing algorithm: starting the process from some seed-points, we use a threshold θ to estimate when two adjacent sequences are similar using the distance D_{ES}. In our case, the threshold has been heuristically fixed after few experimental trials, and is not a particular critical parameter to set up. The complexity of the segmentation process is $\mathrm{O}(nN^2T)$, where N^2T is due to the calculation of the distance among models, and n is the total number of the pixels.

We will see in the experimental section that the modification of the metric in Eq. (3), with the integration of the chromatic-temporal information of the video-sequence, allows us to obtain a meaningful segmentation.

3.3. Dynamic information: the activity maps

The proposed method is also able to infer the degree of foreground activity in the scene; we characterize such information as *dynamic*, highlighting the main aspect of the foreground, i.e., of being (spatially) dynamic in the scene. Strictly speaking, we define a measure which is able to quantify for each pixel the related level of activity. By visualizing all these measures, we estimate the "activity map" of the scene, in which the areas more affected by foreground motion are recognizable.

A similar goal was achieved in [45], where the activity zones were found by clustering the object trajectories derived from the tracking. Nevertheless, in our case the analysis is performed without resorting to trajectories, but by the direct use of the pixel signals. Other approaches similar in spirit to our objectives are presented in [49], in which a motion energy image (MEI) is used to represent and index human gestures, in [55], where an enhancement of the MEI is proposed, namely the motion history image (MHI), and in [57], where spatio-temporal entropy image (STEI) were used to detect foreground activity. Some of these approaches are summarized in the experimental section, where they have been experimentally compared with our approach.

In our framework, we define a measure which is able to quantify for each pixel the level of activity related to that pixel, and this is carried out by analyzing the parameters of the associated HMM. The key idea is that the temporal evolution of the pixel gray level could be considered as

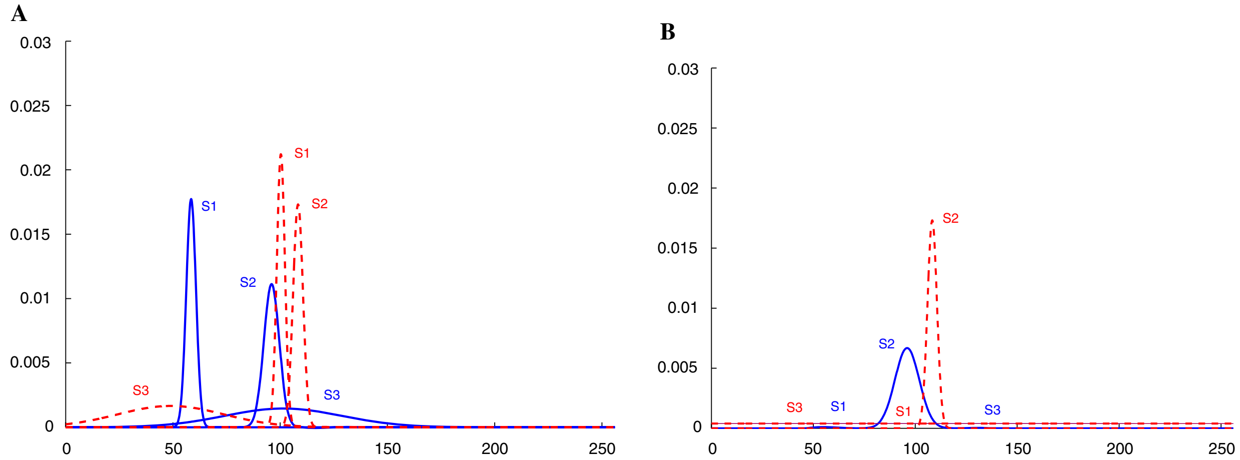


Fig. 2. Normalized variances: in (A) the original representation of the pdf of each HMM state is shown, for 2 HMM models, λ_1 (in red, dashed) and λ_2 (in blue solid). The variances of each state are $\sigma_1 = 2.2500$, $\sigma_2 = 3.5865$, $\sigma_3 = 27.4072$ for λ_1 and $\sigma_1 = 1.8804$, $\sigma_2 = 2.3042$, $\sigma_3 = 24.6667$ for λ_2. The associated $\mathbf{p}_\infty$ is $\langle 0.0228, 0.7752, 0.2020\rangle$ for S_1, S_2, S_3 of λ_1 and $\langle 0.0001, 0.9998, 0.0001\rangle$ for S_1, S_2, S_3 of λ_2, respectively. The result of the variance normalization is shown in (B): one could notice the high importance of states S_2 for both λ_1 and λ_2, and the low importance of the other state variances: in the similarity measure computation their contribution results therefore low. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.)

composed by different components, each one assigned to a particular state during the HMM training. Each component is then characterized by a degree of importance: some are more important, i.e., "explain" more data, others are less important since they result from disturbing sources (e.g., noise). Therefore, if we are able to measure the "importance" of a state, we could infer the importance of the components of the signal which represents the information that we will use to determine the activity zones. As explained in Section 3.2, the "state importance" can be measured using the stationary probability distribution of the Markov Chain associated with the HMM.

Given a HMM $\boldsymbol{\lambda}_{xy}$ with N states, trained on the sequence of the gray-level values assumed by the pixel (x, y), all the information we need is in the vector $\mathbf{p}_{\infty}^{xy}$. The activity measure $AL(x, y)$ should show some precise characteristics, i.e., it should discard the components relative to the background (unimodal or multimodal), and should clearly detect those relative to the foreground, giving a response proportional to the amount of foreground passed over the pixel. These requirements are accomplished by the following measure:

$$AL(x,y) = \sum_{i=1}^{N} \omega_i^{xy} \mathbf{p}_{\infty}^{xy}(i) \log_2 \frac{1}{\mathbf{p}_{\infty}^{xy}(i)} \tag{5}$$

with

$$w_i^{xy} = \log(1 + \sigma_i^{xy}), \tag{6}$$

where σ_i^{xy} is the variance of the Gaussian associated to the state S_i of the HMM $\boldsymbol{\lambda}_{xy}$. The term 1 added to the variance ensures that the weights are all positive.

We use the logarithm to ensure a smoother increasing behavior of the weights. This formula is a sort of "weighted entropy," and is the result of two ideas: the use of the entropy, and the weighting of the components in the entropy computation. The measure of entropy has been chosen since it is able to quantify the uncertainty linked to the model of the pixel gray-level evolution. The idea of weighting has been introduced to deal with multi-modal background (for example a moving foliage), which produces an erroneous high entropy: the idea is to assign lower weight to the terms of the computation that are most related to the background. In this case, we are in fact more interested in the entropy of the states that most probably do not correspond to background, since they represent the activity. The weight is linked to the variance of the Gaussian of the state, so that the lower the variance, the higher the probability that the state corresponds to a background component. By computing this quantity for all the frame pixels, we could finally obtain an activity map of the observed scene. The computational effort required to calculate the activity map is $O(Nn)$.

An immediate consideration that could arise is why we do not directly use the entropy of the gray-level evolution of the pixel, rather that the pseudo entropy of the *model* of the gray-level evolution. The reasons are essentially two: first, the use of a HMM permits to recover from noise that is present in the video sequence, which cannot be accomplished by the raw entropy computation. Second, and more important, the use of HMMs permits to deal also with multimodal background: the entropy of the raw signal results large in case of multimodal background, whereas with our approach this does not occur since the background states are in some way disregarded from the measure computation. This behavior is confirmed by results presented in the experimental section. Moreover, it is important to note that the dynamic information is only one of the by-products of the proposed approach: using the same probabilistic modelling we are also able to infer static information.

4. Experimental trials and comparative analysis

In this section, some comparative experimental evaluations of the proposed approaches are presented.[2] In particular, in Section 4.1 some results regarding the video-segmentation are presented, while Section 4.2 contains results from the activity maps detection process. In the above two sections, some experiments are related to the same sequence, and others are related to different sequences to highlight the particular features of each part of the methodology. Global experiments are then presented in Section 4.3, in which the strengths and the limitations of the whole proposed approach are discussed. Finally, Section 4.4 contains some suggestions about the possible use of the information extracted from the video sequence.

4.1. Static information: the spatio-temporal segmentation

The approach proposed in Section 3.2 is tested using two real sequences: the first one regards a person walking in a corridor in which several doors are present. Some frames of the sequence are presented in Fig. 3. Looking at the figure (video sequence), you can notice that some doors are 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, that it would be reasonable to separate. These different spatial chromatic zones are highlighted in Fig. 4: one is on the left part of the corridor, and the other on the right part. This example shows all the potentialities of the proposed approach: the sequential information employed by our approach is essential to recover all the different semantic regions of the scene. As an example, let us consider only the median (or the mean) of the sequence, i.e., the image formed by the median (mean) values of each pixel signal, displayed in Fig. 5. From these images it is not possible to detect the two semantically

[2] All the video sequences presented in this section are available at the web site profs.sci.univr.it/~bicego/VideoSequences/Scene Understanding.htm.

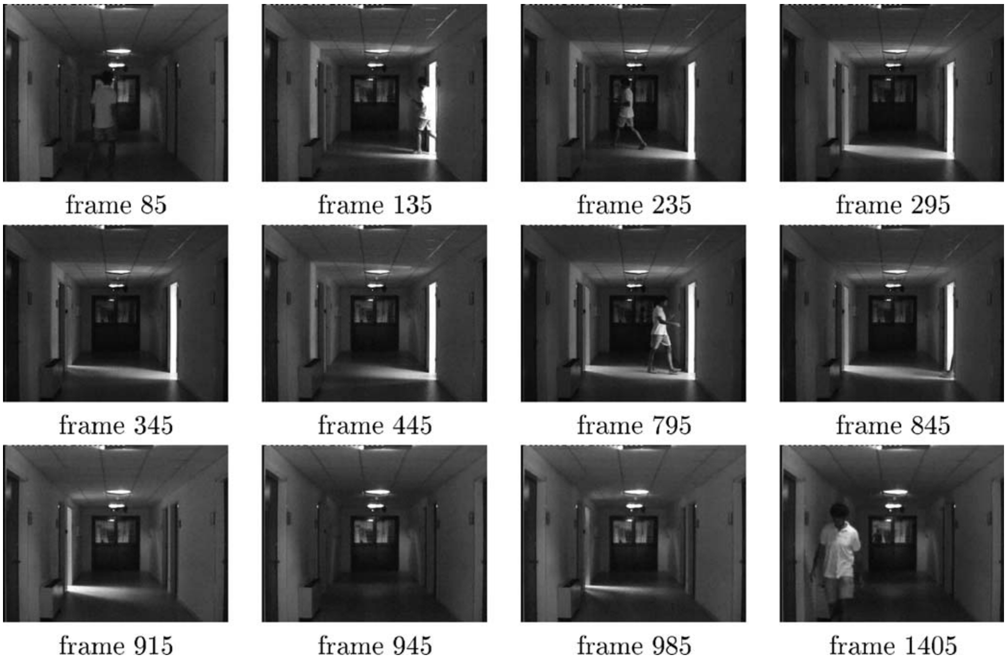


Fig. 3. Frames of the first indoor sequence.

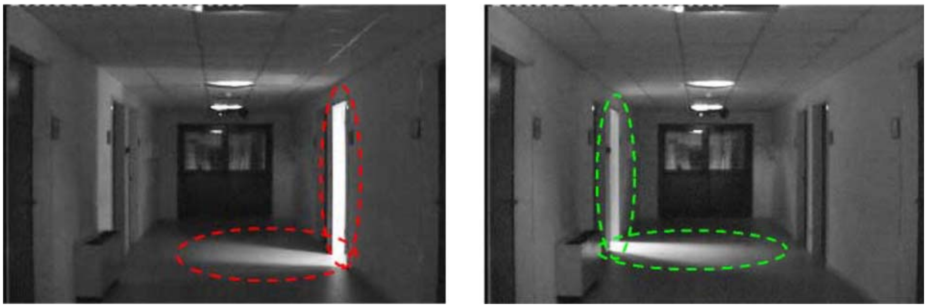


Fig. 4. Different spatial chromatic zones.

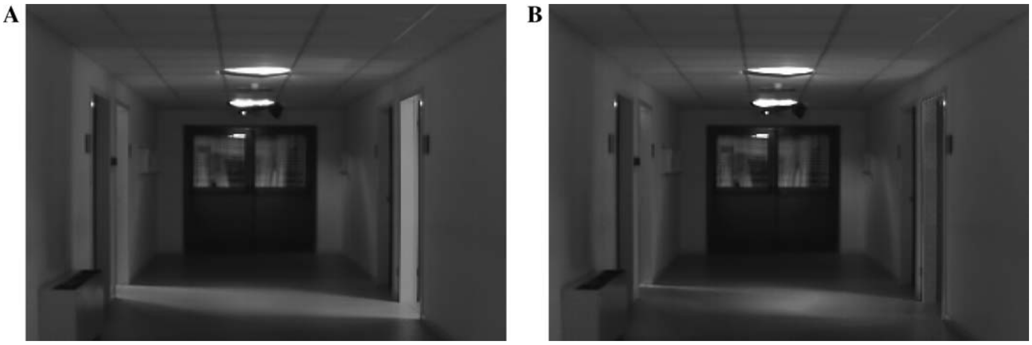


Fig. 5. (A) Average frame; (B) median frame.

different zones of the background. Actually, any spatial segmentation technique applied to these images would segment the zone between the two doors as belonging to the same region. In Fig. 6, the segmentation resulting from our approach is displayed. One can easily notice that our approach clearly separates the two zones, labelled as different regions of the scene. To assess the gain obtained with the Enhanced Stationary similarity measure D_{ES}, the segmentation of the corridor sequence based on the measure of the Eq. (3) is depicted in Fig. 7. It is evident that the noise affecting the video sequence and the presence of foreground produce a very noisy and heavy over-segmentation, whereas our approach is able to manage foreground objects and noise.

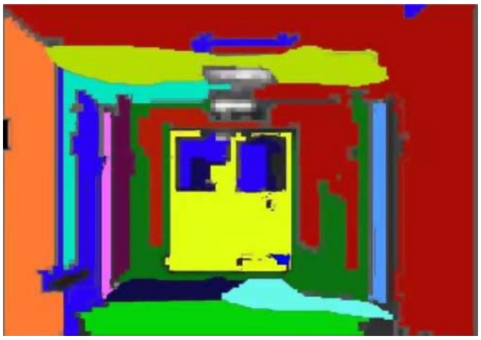


Fig. 6. Static information: spatio-temporal segmentation of the first indoor sequence.

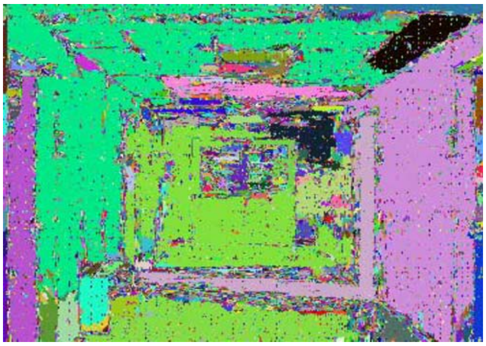


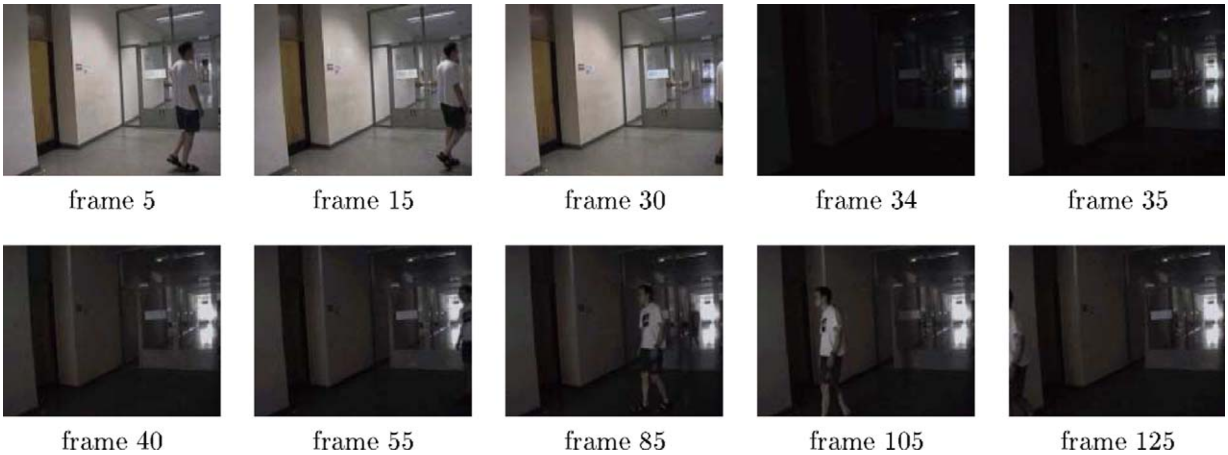
Fig. 7. Segmentation of the first indoor sequence using the similarity measure without the HMM states' normalization.

The second sequence used for testing is obtained from [29] and regards the monitoring of an indoor environment with one moving object. The sequence is formed by 135 frames (320 × 240 pixels) acquired at 20 frame/s. Some of the frames of the sequence are presented in Fig. 8, showing a sudden not uniformly distributed change of the illumination. Such non-uniform luminosity change could drastically affect the comprehension of the sequence, and only a method that uses spatio-temporal information can be able to correctly identify the semantically separated regions. To slow down the computational effort, we partitioned the field of view in a grid with circular Gaussian filters of 5×5 pixels, and at each time step each filter provides one single weighted value (this improvement has drastically reduced the computation time). The result of the segmentation, after the HMM training is reported in Fig. 9: the segmentation is highly informative in that the foreground does not appear in the resulting segmentation, and the change of illumination does not influence the spatial chromatic structure of the scene. Actually, areas of different chromaticity (the floor, portions of the wall) remain separated despite the light reduction narrows down the chromatic difference among them.

4.2. Dynamic information: the activity maps

The method for the extraction of dynamic information, described in Section 3.3, is first tested using three video sequences, to highlight the specific features of this part. Some further comparative evaluations could also be found in the following section, in which complete examples are proposed. As comparative techniques, we considered methods present in the literature (see [49,57]) and simple modified versions of them. Summarizing, all the approaches employed in this section are named as follows:

- Motion energy image (MEI) [49]: the MEI is the sum of the squared differences between each frame and one chosen as reference (the first of the sequence); in particular, to each difference image is applied a threshold T_{MEI} to disregard little values due to noise. The best results of this approach have been obtained using $T_{\mathrm{MEI}} = 4$.
- Modified motion energy image (MMEI): the same approach as above, but the differences are calculated between consecutive frames. This measure will weight much more sudden foreground activities.
- Median over reference squared difference (MedReF): the median operator is applied over the volume of the squared differences with respect to the first frame.



frame 5 frame 15 frame 30 frame 34 frame 35

frame 40 frame 55 frame 85 frame 105 frame 125

Fig. 8. Frames of the second indoor sequence.

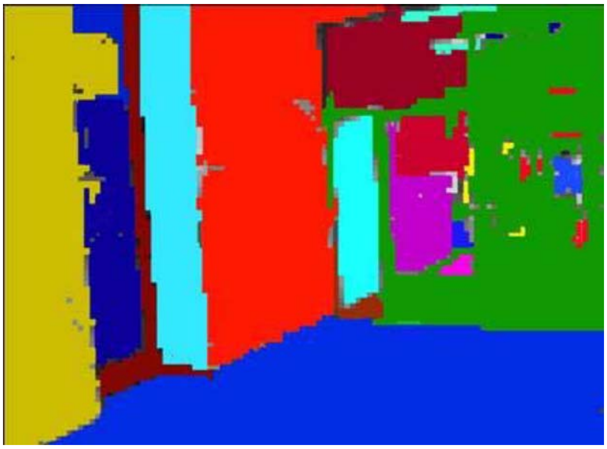


Fig. 9. Static information: spatio-temporal segmentation of the second indoor sequence.

- Median over consecutive squared difference (MedCDif): the median operator is applied over the volume of the consecutive squared differences.
- Simple entropy: for each pixel, we calculate the associated signal entropy in a range of 255 gray-level values. This measure is quite similar to that proposed in [57]: in such approach the entropy is calculated over a time interval of five frames, and over a square spatial window of 3 × 3 pixels.
- The proposed approach.

The first test sequence is composed of 390 frames, acquired at a rate of 15 frames/s. The sequence regards an indoor scene, where a man is entering from the left, walking to a desk, and making a phone call. After the phone call, he leaves the scene going out to the right. Some frames of the sequence are shown in Fig. 10.

The activity zones resulting from the application of the proposed approach and the comparative methods are displayed in Fig. 11, in which higher gray-level values correspond to larger activity. All the output values of the different methods are scaled in the pictures in the interval [0,255]. The results show that the methods based on the differences with respect to an initial frame (Figs. 11A and C) are "complementary" with respect to the ones based on consecutive differences (Figs. 11B and D). In the former case, the person near the phone represents the biggest amount of activity, while in the latter the slow motion of the person makes the vibrating phone wire as the strongest activity. The simple entropy method (Fig. 11E) includes both the person and the wire as energetic objects in the scene, and it is also visible in the center of the scene a mild energy zone, due to the approaching phase of the person to the phone. Another drawback of this method is that also the background signals (due to reflecting effects in the scene and in the decoding of the movie) are taken into account in the calculus of the activity map. Therefore, high energy patterns are detected in correspondence of the bookshelf, over the chair and under the phone; moreover, a general energy amount is detected over all the scene, due to the compression coding of the sequence. Our method (Fig. 11F) avoids all the noise due to the background, highlighting a more precise description of the activity present in the scene. The resulting image is very informative: one could see that the walking zone (i.e., the zone to the left of the desk) is quite active, while the zone near the phone is very active. The zone in the top of the image, where no foreground objects pass, is darker, i.e. no activity is present, and only some noisy behavior is visible.

Another interesting example is proposed in Fig. 12, where some frames of the video sequence are shown. The camera is monitoring an outdoor scene, where there is a starting fire (please note the smoke in the middle). This is a clear example in which object-centered trajectories cannot be extracted, since the moving object has neither a clear shape nor a well-defined contour. The sequence is 450 frames long, which represents 30 s of observation. The activity zones, extracted from this video sequence using all the methods, are shown in Fig. 13.

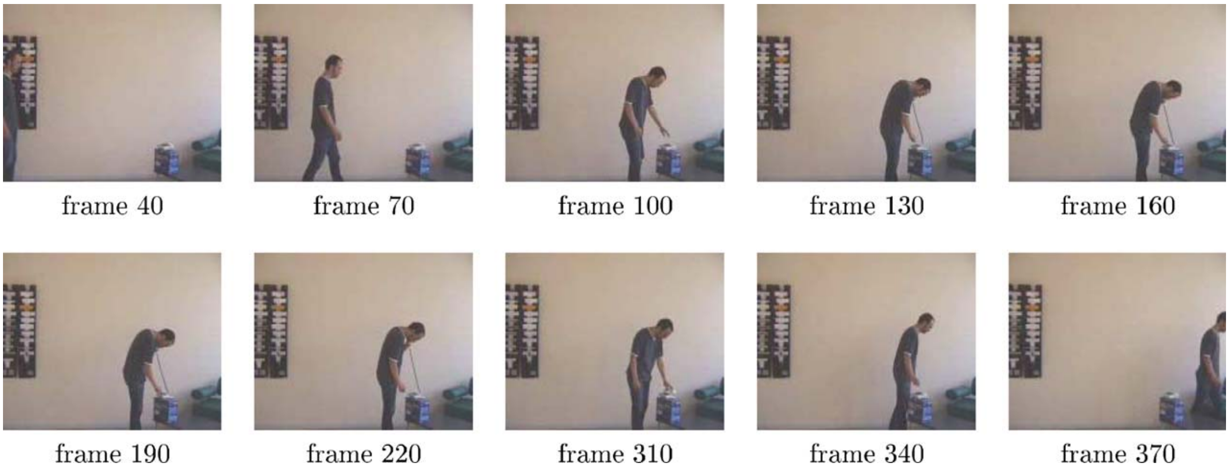


Fig. 10. Some frames of the original sequence.

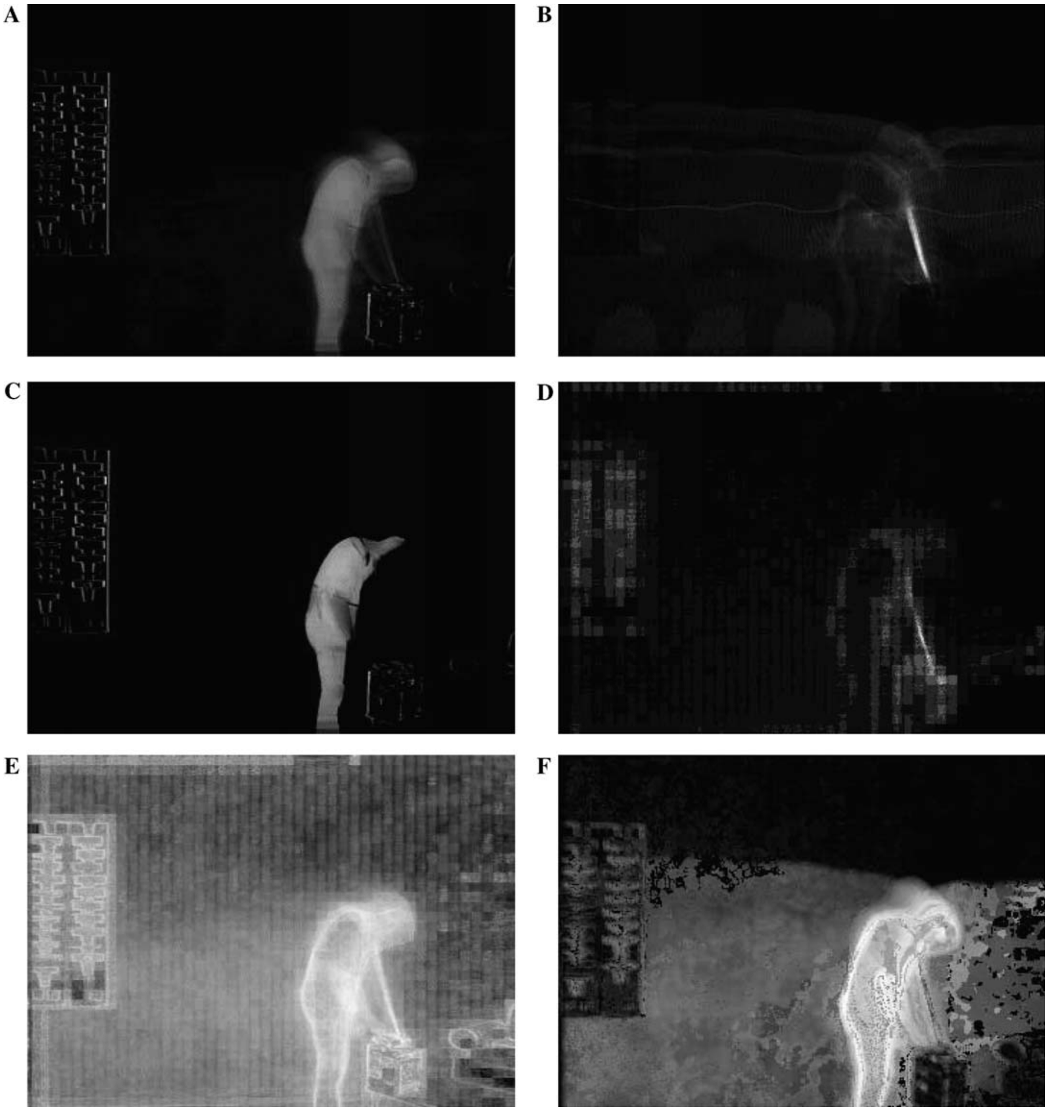


Fig. 11. Activity zones resulting from: (A) MEI over reference frame; (B) MEI over consecutive difference (MMEI); (C) median over reference squared difference (MedReF); (D) median over consecutive squared difference (MedCDif); (E) simple entropy; (F) the proposed approach. The whiter the pixels the higher the activity.

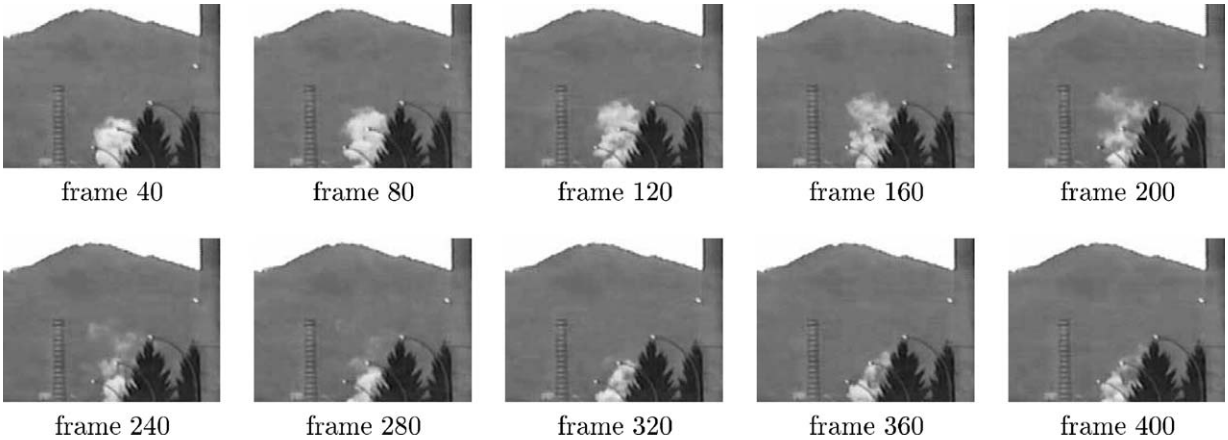


Fig. 12. Some frames of the sequence of the fire.

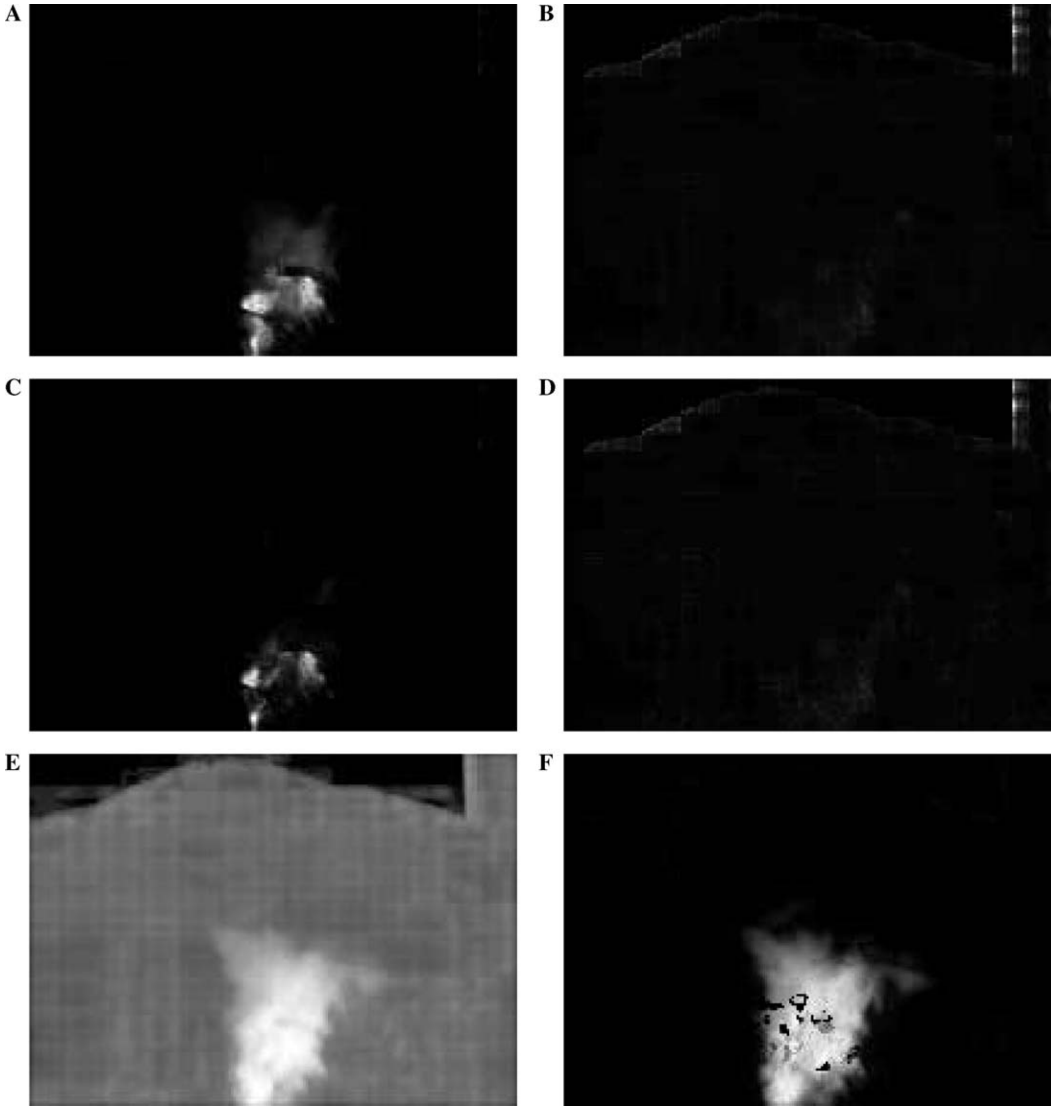


Fig. 13. Activity zones resulting from: (A) MEI over reference frame; (B) MEI over consecutive difference; (C) median over reference squared difference (MedReF); (D) median over consecutive squared difference (MedCDif); (E) simple entropy; (F) the proposed approach. The whiter the pixels the higher the activity.

In this case, all the methods based on consecutive differences (Figs. 13B and D) fail due to the slow motion of the smoke. The methods based on the difference with respect to a reference frame (Figs. 13A and C) perform better, even if a clear pattern is not identifiable. In general, all these methods are able to absorb the background noise. The entropy of the sequence, shown in Fig. 13E, highlights also the background noise, resulting in a overall high energy scene. Using our approach, depicted in Fig. 13F, it is possible to clearly identify the smoke zone, indicating that there is a certain activity. Further, it is important to note that the fire has been detected analyzing only 30 s of the scene. The holes present in the image can derive from the lamp-posts which are located ahead of the smoke, as can be noticed by looking at Fig. 12. Comparing the two last images we can also notice that they carry similar information: in both cases the smoke area is clearly identified. This is obvious, since the same guiding principle is used: in our case, it is the entropy of the *model* of the signal, whereas in the second it is the entropy of the signal itself. However, the image resulting from our approach clearly separates activity from inactivity (all the remaining part of the scene is dark), while using the "simple entropy" approach the activity in the mountains zone is larger than that of the sky, and this represents an erroneous interpretation.

Another interesting aspect has to be pointed out. In the MEI and MedReF approaches, the reference frame has to be carefully chosen: essentially, being the reference frame fixed over time, the evolution of the light and the weather (the background) is not modelled. Our method, as shown in the following example, is able to deal with such kinds of situations. We employed the video sequence presented in Fig. 8, applying the different approaches: results are presented in Fig. 14.

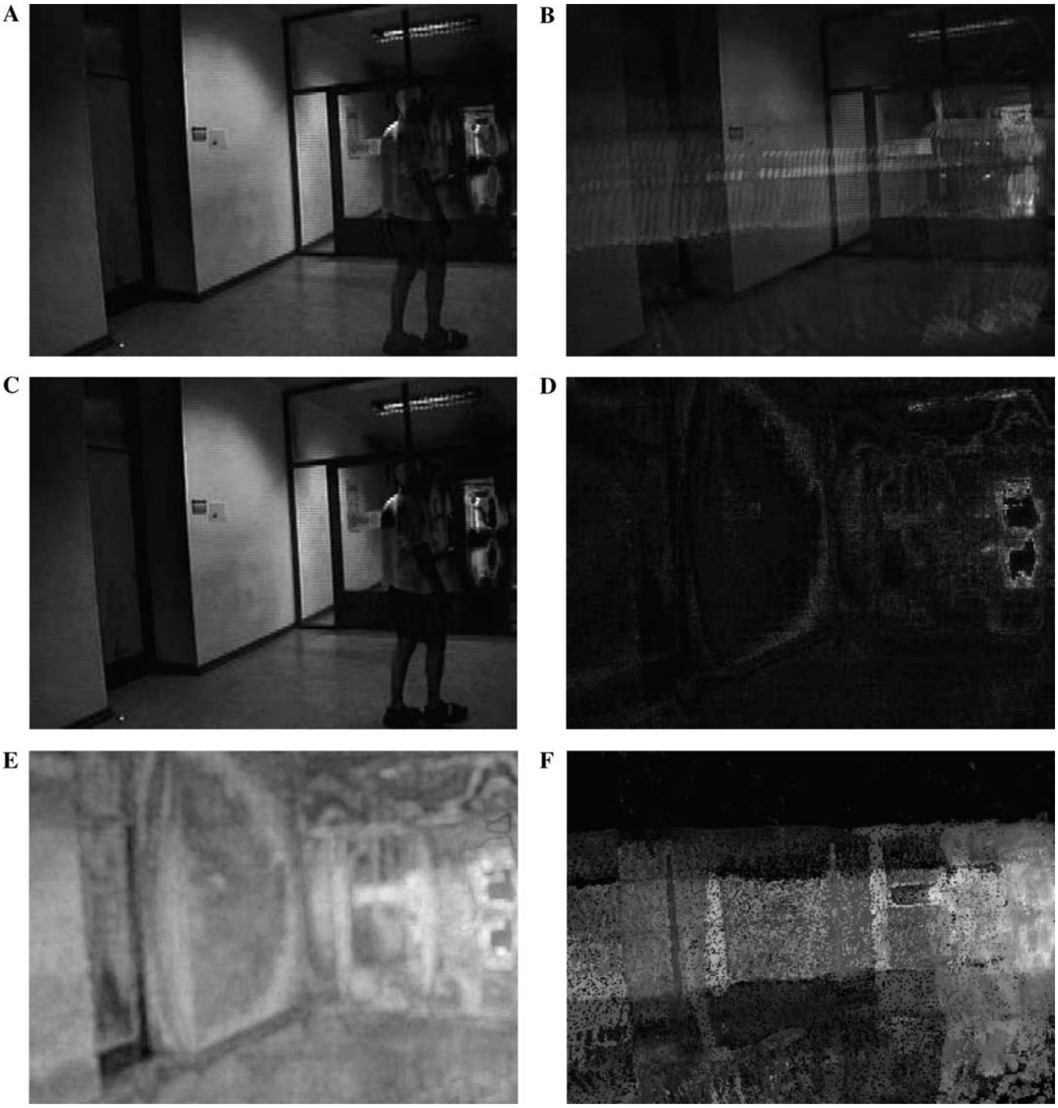


Fig. 14. Activity zones resulting from: (A) MEI over reference frame; (B) MEI over consecutive difference (MMEI); (C) median over reference squared difference (MedReF); (D) median over consecutive squared difference (MedCDif); (E) simple entropy; (F) the proposed approach. The whiter the pixels the higher the activity.

The change of illumination in the sequence produces erroneous activity maps in the methods based on differences over a reference frame (Figs. 14A and C). As stated before, all these methods work well in the situations in which the background is highly static, as in the case of well constrained indoor environments, or environments considered over short periods of time. In situations in which the chromatic aspect of the background is changing over time, all these methods are not applicable. In the methods based on consecutive differences (Figs. 14B and D), the change of illumination is better absorbed: it is 5 frames long, therefore, each consecutive difference image has smaller pixel (absolute) values than the one built between the current frame and the reference one. Nevertheless, that quantity is bigger with respect to the values of the consecutive differences due to the moving person in the scene: the overall result is that the change of illumination visually predominates on the moving object. Looking at the Fig. 14E, we can notice that the "simple entropy" method completely fails in that the illumination change occurring in the middle of the sequence makes not possible to recover any meaningful information. Actually, one can notice that the resulting image does not provide any expressive interpretation being quite uniform. On the other side, our method is able to recover useful information about the movement of the person in the hallway. In particular, looking at Fig. 17F, we could infer three correct information: (1) the top part of the scene is not active, which is correct; (2) there is something moving in the bottom, going through all the scene, and this is also correct; (3) the right part of the scene is more active than the left part: this is still correct, since the man starts walking (Fig. 8) in the middle part of the scene and come back in from the right.

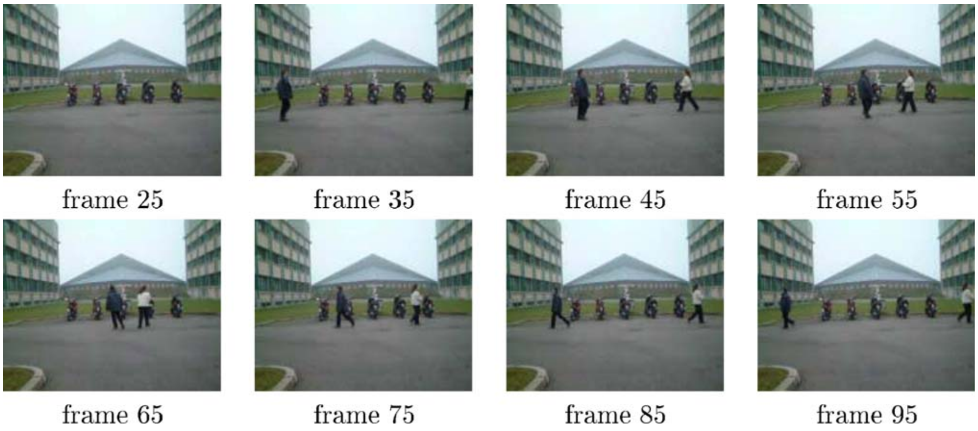


Fig. 15. Some frames from the outdoor sequence.

4.3. Complete examples

In this section, we present three additional complete applications of the proposed approach: given a video sequence, we build the probabilistic representation, and we extract both static and dynamic information. The first testing sequence regards an outdoor environment where two persons are closing and come back. A few frames of the sequence are presented in Fig. 15.

After building the probabilistic approach, we extract static and dynamic information.

Looking at the static part of the analysis (Fig. 16), one could notice that the segmentation is clear, expressive, and quite accurate: zones with similar gray level and similar chromatic behavior (the road, the sky, and the motorbikes) are represented as single regions. Other zones characterized by a different chromatic behavior (the two buildings and part of the pyramid) are oversegmented. It is worthwhile to notice that this segmentation is obtained by processing the whole sequence, without any need to remove the moving objects, in that they are naturally removed by the procedure used to compute the enhanced stationary distance D_{ES}.

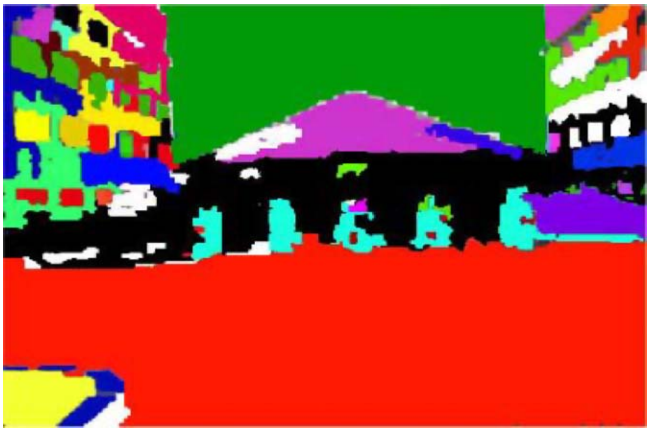


Fig. 16. Information extracted from the outdoor sequence using the proposed approach: static information (spatio-temporal segmentation).

The comparative results related to the dynamic part express the same considerations made for the examples relative to the Section 4.2 (see Fig. 17). In general, all the noisy background situations have great impact over the final energy image: the more noisy the background, the less important the role of the foreground on the final map. In general, the entropy-based method over relatively short sequences (100–500 frames) is highly prone to over estimated energy errors. Moreover, when the foreground appears briefly in the scene, the median-based methods (Figs. 17C and D) tend to prune away the correspondent activity, and the simple entropy method (Fig. 17E) strongly highlights the light noise activity, in this case due to the video compression. Looking at the proposed approach (Fig. 17F), one could notice that the image is quite informative: the part of the scene where people are walking is clearly expressed, as well as the non-active part. Moreover, it is interesting to notice that it is possible to precisely infer also some further details, as the positions where the legs of the people are standing more time, which represents a larger level of detail. This detail is also represented by the MEI and MMEI approaches (Figs. 17A and B, respectively), although with less strength.

The second and the third sequences of this section should be considered the most hard ones, in which our current approach shows its limits, regarding in particular the static analysis. These limits will draw the directions of our research. The former sequence represents an outdoor environment,[3] in which a traffic situation over a square is monitored via a fixed camera (Fig. 18). The sequence is 1710 frames long, acquired at 30 fps. The chromo-spatio-temporal segmentation, in this case, is highly over-segmented. This is due to the intrinsic irregularity with which the static zone evolve, and to the difficulty to clearly distinguish what is the background and what the foreground (some blocked cars could be detected as background). One of the possible solutions is to restrict the static analysis to the zones where the activity map gives low values.

[3] Downloaded from ftp://ftp.ira.uka.de/pub/vid-text/image_sequences/kwbB/sequence.mpg.

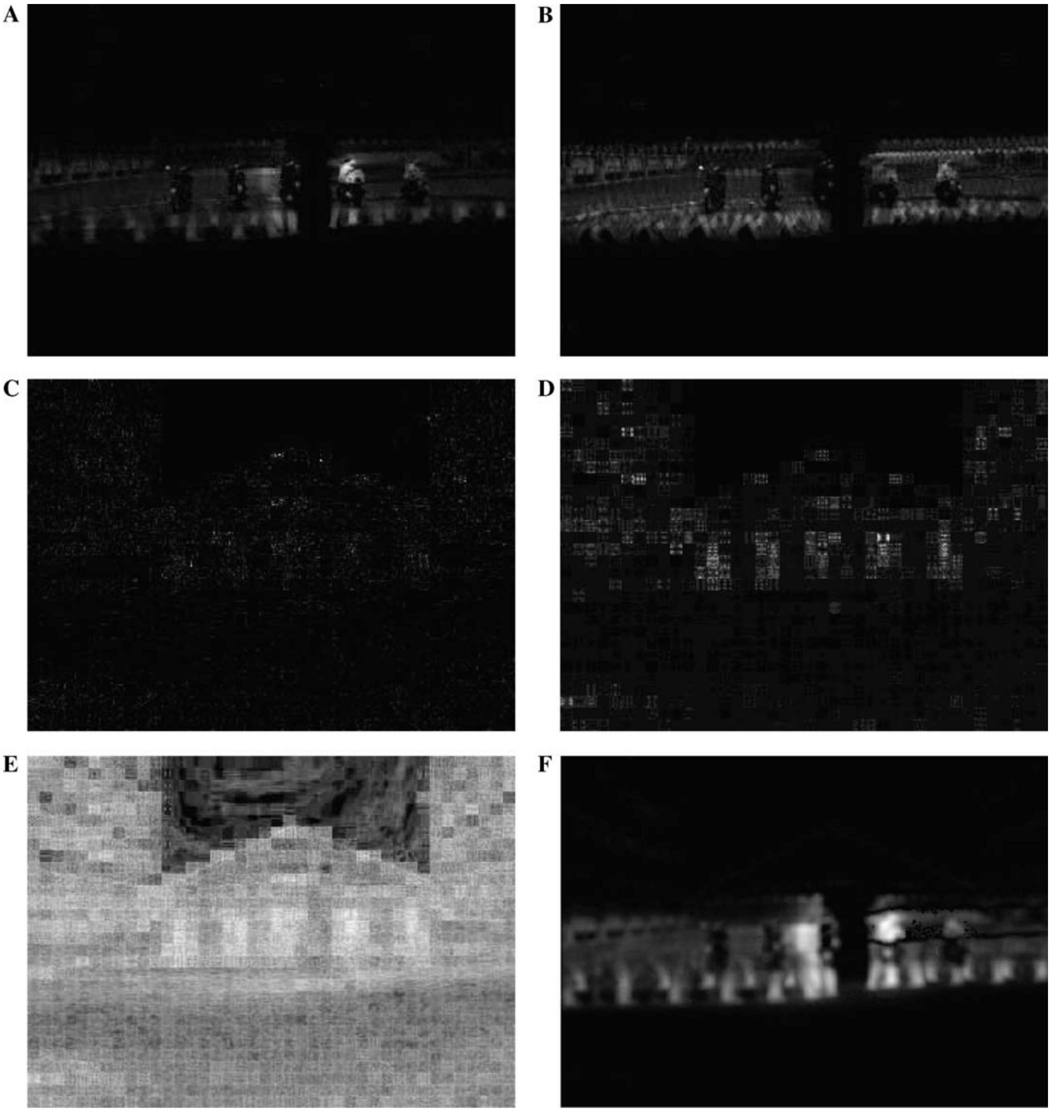


Fig. 17. Activity zones resulting from: (A) MEI over reference frame; (B) MEI over consecutive difference (MMEI); (C) median over reference squared difference (MedReF); (D) median over consecutive squared difference (MedCDif); (E) simple entropy; (F) the proposed approach. The whiter the pixels the higher the activity.

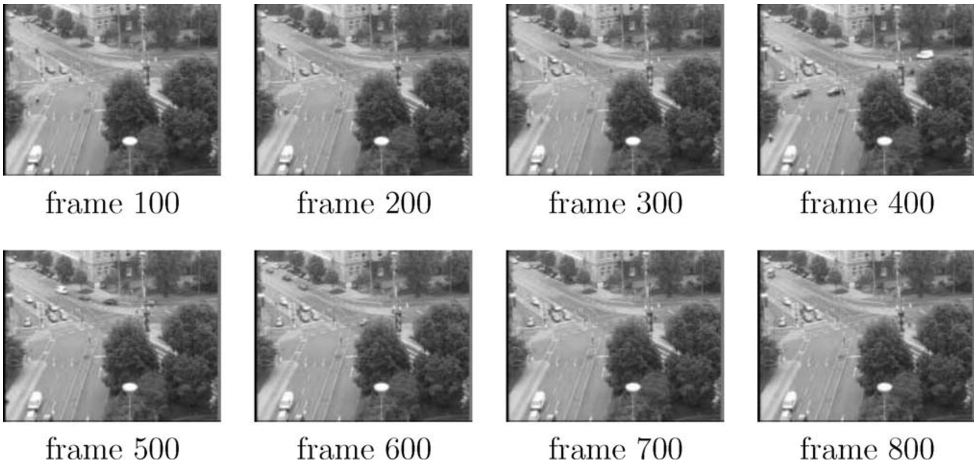
The results of the dynamic part are shown below (Fig. 19).

As one can notice, the only useful results are the ones relatives to the entropy-based approaches (E and F) (MMEI is also good, but not so informative; actually, the activity due to the people in the upper right part of the scene and in the middle of the crossing street is not shown). It is interesting to note the cylindric high energy zone detected in the middle-right part of the scene. That part represents a rotating billboard, detected as high energetic foreground pattern. This represents a wrong estimation, due to the low number of states with which the HMMs have been trained. Actually, the behavior of the area is 5-modal, with the modes fixed over time. Intuitively, this area should be interpreted as background, with low energy in the activity map, and this situation could be recovered using HMM with 5 number of states (Fig.20).

The last sequence, 8 min long acquired at 25 fps, represents an indoor environment of a mall[4] (some frames are depicted in Fig. 21). One of the purpose of this experiment is to assess the performance of our method over different sequence lengths. The test is divided in two stages: the first in which an initial short part of the sequence is evaluated (48 s long, 1/10 of the original one); in the second stage the whole sequence is analyzed.

This is the most difficult indoor sequence analyzed, due to strong noisy effects degrading the data quality, like reflections over the floor and over the lateral windows. In both processing cases, the outcome of the static analysis results over-segmented, as expected. Since the results on

[4] Downloaded from http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1.



frame 100 frame 200 frame 300 frame 400

frame 500 frame 600 frame 700 frame 800

Fig. 18. Some frames from the traffic sequence.

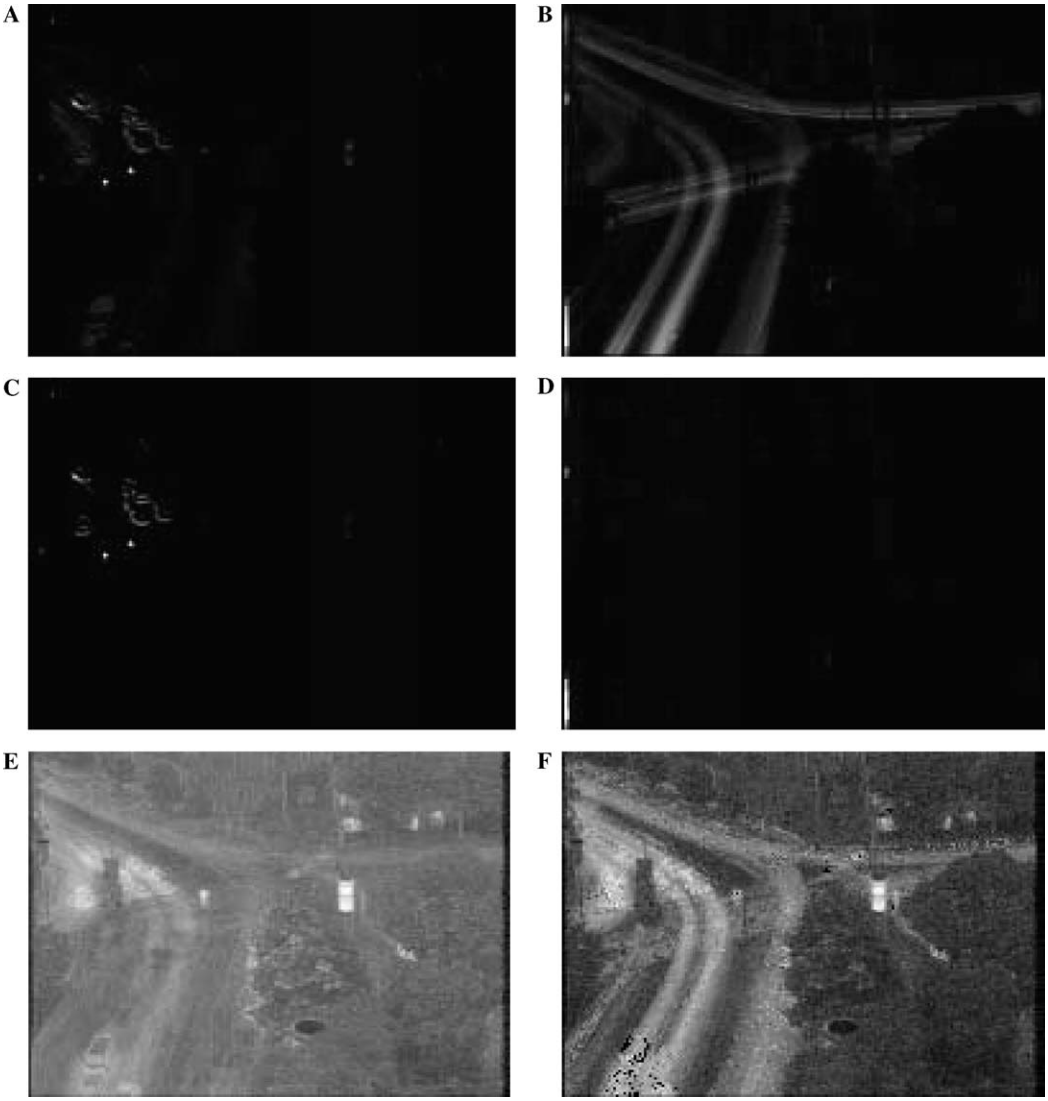


Fig. 19. Activity zones resulting from: (A) MEI over reference frame; (B) MEI over consecutive difference (MMEI); (C) median over reference squared difference (MedReF); (D) median over consecutive squared difference (MedCDif); (E) simple entropy; (F) the proposed approach. The whiter the pixels the higher the activity.

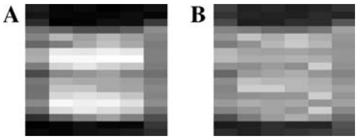


Fig. 20. Traffic sequence: the detail of the billboard in Fig. 19, whose pixels are trained with HMMs having 3 (A) or 5 states (B). It is possible to evaluate the decreasing of activity detected in (B) due to the correct estimation of the five modalities which characterize the billboard behavior.

the short sequence are not significative, only the results related to the longer sequence are shown in Fig. 22.

Looking at this figure, it is possible to reason about what are the biggest areas whose chromatic behavior is similar in time, so as to detect the most "stable" scene areas. In particular, it is possible to detect stable zones in proximity of the lateral columns, and in some parts of the floor, while the area corresponding to the left wall, with several glass windows, is in general over-segmented.

For what concerns the dynamic analysis, the techniques based on differences between frames give poor results in both tests, hence only the method based on the entropy and the proposed approach are presented.

As shown in the previous results, the proposed approach works better compared to the entropy method, individuating the zones in which the foreground activity is located, disregarding the noise. In particular, in Fig. 23F1 is possible to clearly detect the left drift of energy, that models the fact that the people enters frequently in the mall. In Fig. 23E1 this aspect is not so clearly highlighted. Moreover, augmenting the sequence length, the effect of the entropy takes strongly into account the noise due to the reflection on the floor: this results in an activity map that "forgets" the amount of activity present in the bottom part of the hallway. Conversely, our approach is able to represent all the foreground activity as shown in Fig. 23F2.

Summarizing, the experiments have assessed that the proposed method, concerning the static information extraction, provides a novel kind of analysis able to explain the chromatic evolution of the static part of the sequence, individuating regions with similar temporal-chromatic profile. The main drawback results in the generation of an over segmentation, which occurs in noisy cases. For what concerns the dynamic analysis of the foreground, the proposed approach outperforms in general all the tested comparative methods, showing a certain degree of robustness over all the input (long/short sequences with low/high noise levels).

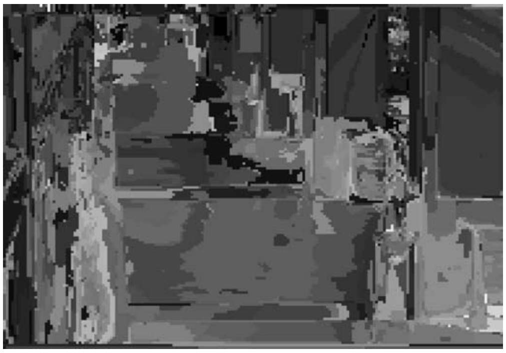


Fig. 22. Information extracted from the whole outdoor sequence using the proposed approach: static information.

4.4. Possible applications

In this section, the possible uses of the information extracted with the proposed approach are investigated. A first example can be found in [22], where this information has been used in order to initialize an integrated pixel- and region-based approach to background modelling, proposed in [23]. This background model uses information derived from a spatial segmentation of the scene in order to modulate the response of a standard pixel-level background modelling scheme [46], increasing the robustness against local non uniform illumination changes. The information extracted with our approach is used to initialize this model: in particular, the initialization of the pixel level part of the model straightforwardly derives from the

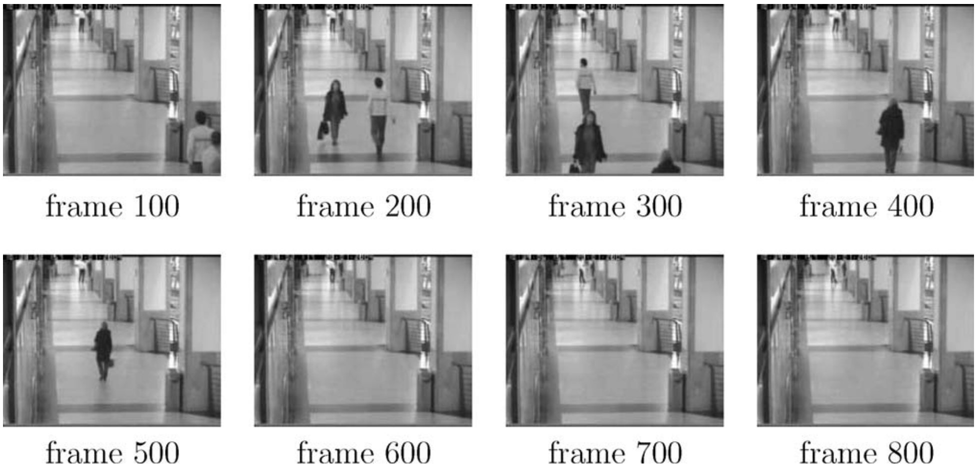


Fig. 21. Some frames from the traffic sequence.

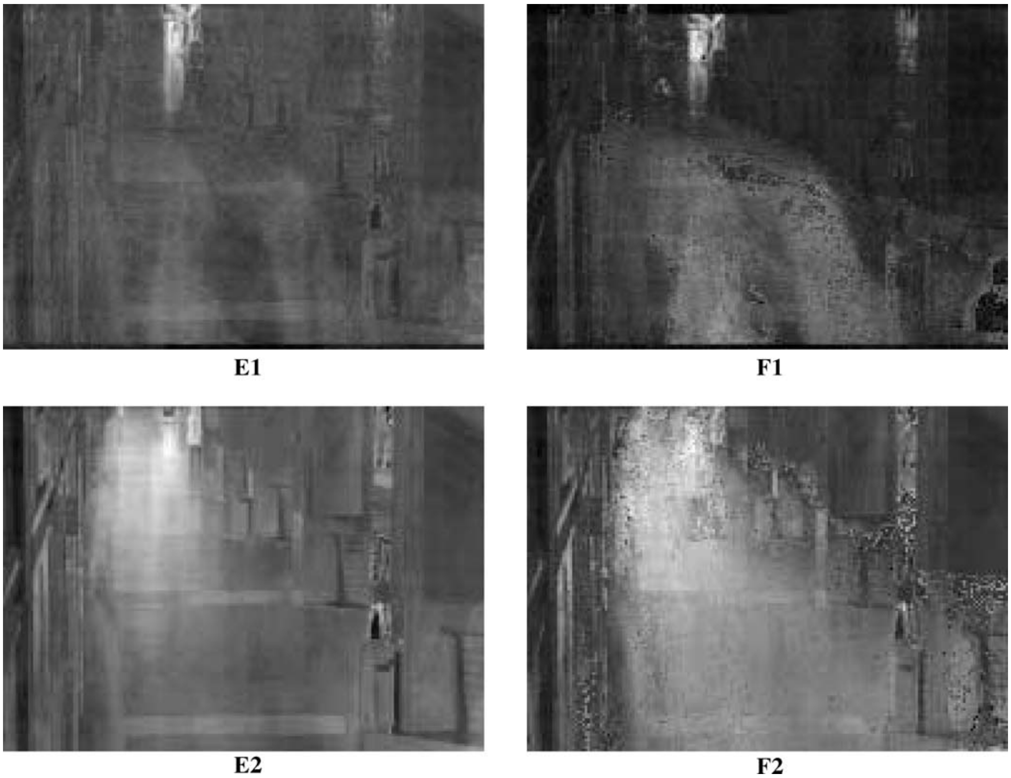


Fig. 23. Activity zones resulting from: (E1) simple entropy and (F1) the proposed approach calculated over the 48 s sequence; (E2) simple entropy and (F2) the proposed approach calculated over the entire 8 min sequence; the whiter the pixels the higher the activity.

probabilistic modelling of the video sequence, while the initialization of the region level part is the spatio temporal segmentation described in Section 3.2.

There are two other ways of employing the information extracted from the proposed approach, which are currently in progress. The first is to use the activity map to decide the level of detail of a variable resolution background modelling scheme: the idea is that in those zones where no activities typically occur, a very accurate background analysis is not necessary, and a coarse analysis could be sufficient. The second application can be to use the activity maps to infer the zones of appearance of the foreground with high probability, the so-called source detection problem [47]. The idea is that it is not useful to accurately monitor the zones of the scene where typically no foreground objects are likely to occur.

One constraint of the described approach regards the requirement of a fixed camera. In principle, this condition can be relaxed by performing a pre-registration of the image pixels using an estimate of dominant motion of the scene so that temporal gray-level profiles can be reliably evaluated. Further, such registration could not be critical if small local areas are considered instead of single pixels like in one of the experiments above.

5. Conclusions

In this paper, a new method for scene analysis from video sequences has been proposed, using only very low-level data, that is just pixel behavior. The proposed approach models the sequence using a forest of Hidden Markov models, each one devoted to modelling the temporal evolution of the gray level of each pixel. Given this representation, two kinds of analysis have been developed: the first one clusters the HMMs to obtain a spatio-temporal segmentation of the background, and the second one defines an entropy based measure computed on the stationary probability distribution of each HMM to infer the activity zones of the scene. The proposed approach has several key features with respect to the methods in the state of the art: it extracts information from the lowest possible level (the pixel level), it is unsupervised in nature, it uses HMM at a very basic level, and it employs the same principled probabilistic modelling to infer both static and dynamic information. The results obtained from real experiments have shown the effectiveness of the proposed approach, also with respect to state of the art methods, able to get a better insight of the scene and on the interpretation of the activities occurring therein.

References

[1] N. Jojic, N. Petrovic, B. Frey, T. Huang, Transformed hidden Markov models: Estimating mixture models of images and inferring spatial transformations in video sequences, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2000, pp. 26–33.

[2] M. Naphade, T. Huang, A probabilistic framework for semantic indexing and retrieval in video, in: Proc. IEEE Internat. Conf. on Multimedia and Expo(I), 2000, pp. 475–478.

[3] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, Query by image and video content: the qbic system, IEEE Computer 28 (9) (1995) 23–32.

[4] S. Maillet, Content-based video retrieval: An overview, available at http://viper.unige.ch/ marchand/CBVR/overview.html (2000).

[5] PAMI, Special issue on video surveillance, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8), 2000.

[6] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, A system for video surveillance and monitoring, Tech. Rep. CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, 2000.

[7] N. Johnson, D. Hogg, Learning the distribution of object trajectories for event recognition, Image Vision Comput. 14 (1996) 609–615.

[8] D. Gavrila, The visual analysis of human movement: a survey, Comput. Vision Image Understand. 73 (1) (1999) 82–98.

[9] M. Brand, N. Oliver, S. Pentland, Coupled hidden markov models for complex action recognition, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 1997, pp. 994–999.

[10] T. Jebara, A. Pentland, Action reaction learning: Automatic visual analysis and synthesis of interactive behavior, in: Proc. Internat. Conf. on Computer Vision Systems, 1999.

[11] R. Morris, D. Hogg, Statistical models of object interaction, Internat. J. Comput. Vision 37 (2000) 209–215.

[12] N. Oliver, B. Rosario, A. Pentland, Graphical models for recognising human interactions, in: Advances in Neural Information Processing Systems, 1998.

[13] A. Galata, N. Jonhson, D. Hogg, Learning variable-length markov models of behavior, Comput. Vision Image Understand. 81 (2001) 398–413.

[14] H. Buxton, Learning and understanding dynamic scene activity: a review, Image Vision Comput. 21 (2003) 125–136.

[15] B. Frey, N. Jojic, Transformation-invariant clustering using the EM algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 25 (1) (2003) 1–17.

[16] I. Haritaoglu, D. Harwood, L. Davis, W^4: real-time surveillance of people and their activities, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 809–830.

[17] T. Wada, T. Matsuyama, Multiobject behavior recognition by event driven selective method, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 873–887.

[18] S. Gong, J. Ng, J. Sherrah, On the semantics of visual behaviour, structured events and trajectories of human action, Image Vision Comput. 20 (12) (2002) 873–888.

[19] J. Ng, S. Gong, Learning pixel-wise signal energy for understanding semantics, in: Proc. of the British Machine Vision Conference, BMVA Press, 2001, pp. 695–704.

[20] J. Sherrah, S. Gong, Continuous global evidence-based bayesian modality fusion for simultaneous tracking of multiple objects, in: Proc. Internat. Conf. on Computer Vision, 2001, pp. 42–29.

[21] T. Xiang, S. Gong, D. Parkinson, Autonomous visual event detection and classification without explicit object-centred segmentation and tracking, in: Proc. of the British Machine Vision Conference, BMVA Press, 2002, pp. 233–242.

[22] M. Cristani, M. Bicego, V. Murino, Multi-level background initialization using hidden markov models, in: Proc. ACM SIGMM Workshop on Video Surveillance, 2003, pp. 11–19.

[23] M. Cristani, M. Bicego, V. Murino, Integrated region- and pixel-based approach to background modelling, in: Proc. IEEE Workshop on Motion and Video Computing, 2002, pp. 3–8.

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

[27] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Statist. Soc. B 39 (1977) 1–38.

[28] C. Wu, On the convergence properties of the EM algorithm, Ann. Stat. 11 (1) (1983) 95–103.

[29] B. Stenger, V. Ramesh, N. Paragios, F.Coetzee, J.M. Buhmann, Topology free hidden Markov models: application to background modeling, in: Proc. Internat. Conf. Computer Vision, vol. 1, 2001, pp. 294–301.

[30] J. Rittscher, J. Kato, S. Joga, A. Blake, A probabilistic background model for tracking, in: Proc. Eur. Conf. on Computer Vision, 2000, pp. 336–350.

[31] P. Brémaud, Markov Chains, Springer, Berlin, 1999.

[32] M. Bicego, V. Murino, M. Figueiredo, Similarity-based clustering of sequences using hidden Markov models, in: P. Perner, A. Rosenfeld (Eds.), Machine Learning and Data Mining in Pattern Recognition, vol. LNAI 2734, Springer, Berlin, 2003, pp. 86–95.

[33] M. Law, J. Kwok, Rival penalized competitive learning for model-based sequence, in: Proc. Internat. Conf. on Pattern Recognition, vol. 2, 2000, pp. 195–198.

[34] P. Smyth, Clustering sequences with hidden Markov models, in: M. Mozer, M. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems, vol. 9, MIT Press, Cambridge, 1997, p. 648.

[35] C. Li, G. Biswas, A bayesian approach to temporal data clustering using hidden Markov models, in: Proc. Internat. Conf. on Machine Learning, 2000, pp. 543–550.

[36] C. Li, G. Biswas, Applying the Hidden Markov Model methodology for unsupervised learning of temporal data, Int. J. Knowl.-based Intell. Eng. Syst. 6 (3) (2002) 152–160.

[37] A. Panuccio, M. Bicego, V. Murino, A Hidden Markov Model-based approach to sequential data clustering, in: T. Caelli, A. Amin, R. Duin, M. Kamel, D. de Ridder (Eds.), Structural, Syntactic and Statistical Pattern Recognition, Lecture Notes in Computer Series, vol. 2396, Springer, 2002, pp. 734–742.

[38] C. Bahlmann, H. Burkhardt, Measuring HMM similarity with the Bayes probability of error and its application to online handwriting recognition, in: Proc. Internat. Conf. on Document Analysis and Recognition, 2001, pp. 406–411.

[39] S. Kullback, R. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1951) 79–86.

[40] G. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.

[41] A. Stolcke, S. Omohundro, Hidden Markov Model induction by Bayesian model merging, in: S. Hanson, J. Cowan, C. Giles (Eds.), Advances in Neural Information Processing Systems, vol. 5, Morgan Kaufmann, San Mateo, CA, 1993, pp. 11–18.

[42] M. Brand, An entropic estimator for structure discovery, in: M. Kearns, S. Solla, D. Cohn (Eds.), Advances in Neural Information Processing Systems, vol. 11, MIT Press, Cambridge, 1999, pp. 723–729.

[43] M. Bicego, A. Dovier, V. Murino, Designing the minimal structure of Hidden Markov Models by bisimulation, in: M. Figueiredo, J. Zerubia, A. Jain (Eds.), Energy Minimization Methods in Computer Vision and Pattern Recognition, Lecture Notes in Computer Series, vol. 2134, Springer, 2001, pp. 75–90.

[44] M. Bicego, V. Murino, M. Figueiredo, A sequential pruning strategy for the selection of the number of states in Hidden Markov Models, Patt. Recogn. Lett. 24 (9–10) (2003) 1395–1407.

[45] D. Demirdjian, K. Tollmar, K. Koile, N. Checka, T. Darrell, Activity maps for location-aware computing, in: Proc. IEEE Workshop on Applications of Computer Vision, 2002, pp. 70–75.

[46] C. Stauffer, W. Grimson, Adaptive background mixture models for real-time tracking, in: Proc. Internat. Conf. on Computer Vision and Pattern Recognition, vol. 2, 1999, pp. 246–252.

[47] C. Stauffer, Estimating tracking sources and sinks, in: Proc. IEEE Workshop on Event Mining, 2003, p. 34.

[48] J. Alon, S. Sclaroff, G. Kollios, V. Pavlovic, Discovering clusters in motion time-series data, in: Proc. Internat. Conf. on Computer Vision and Pattern Recognition, 2003, pp. 375–381.

[49] A. Bobick, J. Davis, An appearance-based representation of action, in: Proc. Internat. Conf. on Pattern Recognition (ICPR '96), vol. 1, 1996, pp. 307–310.

[50] B. Juang, L. Rabiner, A probabilistic distance measure for Hidden Markov Models, AT&T Tech. J. 64 (2) (1985) 391–408.

[51] O. Cappé, Ten years of HMMs. Available from: http://www.tsi.enst.fr/cappe/docs/hmmbib.html, 2001.

[53] S. Zhong, J. Ghosh, HMMs and coupled HMMs for multi-channel EEG classification, in: Proc. IEEE Internat. Joint Conf. on Neural Networks, vol. 2, 2002, pp. 1154–1159.

[54] S. Gong, T. Xiang, Recognition of group activities using a dynamic probabilistic network, in: IEEE Proc. Internat. Conf. on Computer Vision, 2003, pp. 742–749.

[55] J.W. Davis, A.F. Bobick, The representation and recognition of human movement using temporal templates, in: IEEE Proc. Internat. Conf. on Computer Vision and Pattern Recognition, 1997, pp. 928–934.

[57] G. Jing, C.E. Siong, D. Rajan, Foreground motion detection by difference-based spatial temporal entropy image, in: IEEE Tencon 2004, 2004.