Trajectory Classification Using Switched Dynamical Hidden Markov Models

Jacinto C. Nascimento, *Member, IEEE*, Mário A. T. Figueiredo, *Fellow, IEEE*, and Jorge S. Marques

***Abstract*—This paper proposes an approach for recognizing human activities (more specifically, pedestrian trajectories) in video sequences, in a surveillance context. A system for automatic processing of video information for surveillance purposes should be capable of detecting, recognizing, and collecting statistics of human activity, reducing human intervention as much as possible. In the method described in this paper, human trajectories are modeled as a concatenation of segments produced by a set of low level dynamical models. These low level models are estimated in an unsupervised fashion, based on a finite mixture formulation, using the *expectation-maximization* (EM) algorithm; the number of models is automatically obtained using a *minimum message length* (MML) criterion. This leads to a parsimonious set of models tuned to the complexity of the scene. We describe the switching among the low-level dynamic models by a hidden Markov chain; thus, the complete model is termed a *switched dynamical hidden Markov model* (SD-HMM). The performance of the proposed method is illustrated with real data from two different scenarios: a shopping center and a university campus. A set of human activities in both scenarios is successfully recognized by the proposed system. These experiments show the ability of our approach to properly describe trajectories with sudden changes.**

***Index Terms*—Expectation-maximization, hidden Markov models(HMMs), human activities, minimum message length, mixture models, unsupervised learning, visual surveillance.**

I. Introduction

HUMAN activity recognition (HAR) aims at understanding what people are doing from their position, shape, or movement, observed in video sequences. HAR is a key technical component in boosting the interaction between people and environments equipped with sensors and computational resources. *Smart rooms* are examples of environments equipped with cameras and microphones, which are used to infer what people are doing in order to interact with them [21], [40], [59]. Another widespread example of environments equipped with sensors is video surveillance of large infrastructures, such as parks, airports, shopping malls. In the surveillance case, one standard goal is to characterize trajectories and behaviors, in order to detect abnormal situations, such as running or fighting [23].

An automatic surveillance system should have the ability to learn typical behaviors from video data, without involving specific knowledge about the actions performed by humans in the monitored environment. This objective has been addressed in several works, such as [33] and [59]. Typical approaches comprise three steps [3], [8], [9], [13], [20], [22], [25], [26], [31], [33], [41], [47], [51], [59]: (a) the objects of interest (people) are segmented and tracked; (b) a sequence of features (position, motion, shape) is extracted from the tracked objects; (c) these features are used to classify the observed behavior into one of several classes of interest. This classification can be based on simple rules [3], [13], [20], or using learning methods exploiting the statistical regularities of the data [60]. Since human motion may be complex, it is often more efficiently modeled using a concatenation of simple models. This idea underlies the use of hierarchical models, in which the lower level describes simple events, while the higher levels describe complex actions [21].

This paper proposes a two-layer hierarchical model for human activity recognition. The lower layer consists of a bank of dynamical models, each of which is tailored to describe a specific motion regime. The higher layer models the switching among the lower layer dynamical systems. For example, the trajectory of a person "entering a shop" in a shopping mall can be decomposed into a set of segments, with each segment described by a different motion regime (or dynamical model). Both the low level models and the high level switching are learned from training data in a fully automatic way. The low level models are common to all the activities and their parameters (as well as their number) are learned in an unsupervised way.

Underlying our approach is the observation that people tend not to move randomly through their environments. Instead, they usually engage in motion patterns, related to typical activities or specific locations that they might be interested in approaching. This observation suggests that it is possible to obtain the most significant low level dynamical models within the observed trajectories, a task we address using unsupervised learning of Gaussian mixture models (GMM) with embedded model selection [16].

Building on the low-level models, the activities to be recognized are represented by modeling the switching patterns among these models as Markov processes. We thus have a complete generative model of the trajectories in each activity, which is used for recognition.

Manuscript received July 25, 2008; revised December 03, 2009. First published December 31, 2009; current version published April 16, 2010. This work was supported by the FCT (ISR/IST plurianual funding) through the PIDDAC Program funds and by project PTDC/EEA-CRO/098550/2008. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sabine Susstrunk.

J. C. Nascimento and J. S. Marques are with the Instituto de Sistemas e Robótica, Instituto Superior Técnico, 1049-001 Lisboa, Portugal (e-mail: jan@isr.ist.utl.pt; jsm@isr.ist.utl.pt).

M. A. T. Figueiredo is with the Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001 Lisboa, Portugal (e-mail: mario.figueiredo@lx.it.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2009.2039664

The remainder of the paper is organized as follows. Section II presents an overview of previous work. In Section III, we describe the model of the pedestrian's activities. Section IV details the estimation of the low level dynamic models, while the top level activity model is explained in the Section V. Section VI describes the classification of the activities. Section VII reports experimental results with real data. Appendix A describes the tracking procedure used to extract the trajectories from the video data.

II. Previous Work

Activity recognition systems can be grouped in two broad classes: systems targeted to the short range analysis (e.g., gesture recognition [22], [25], [47], [59], smart rooms [59]), and systems tailored to long range analysis (e.g., trajectory recognition and prediction [26], surveillance of large infrastructures [8], [13], [31], [33], [41], [51]). The work described in this paper falls in the second class.

As explained above, the recognition of human activities is usually done using a three-step bottom-up approach: segmentation and tracking, feature extraction, and classification. Several approaches to these three steps will be briefly discussed in the next paragraphs. Detailed surveys on HAR can be found in [1], [10], [19], [23], and [36].

A. Segmentation and Tracking

Most systems use static cameras and assume that the background is static as well. In this case, moving object detection (segmentation) can be performed by detecting changes in the image, i.e., by separating active regions (foreground) from the static background. In the simplest cases, the current image is subtracted from the static background image and the differences are compared with a threshold. However, background subtraction is extremely sensitive to scene changes, as well as to lighting changes and other extraneous events, which has stimulated several improvements to this approach. For example, the uncertainty associated to each background pixel may be taken into account, by modeling it as a random variable with a Gaussian distribution [59], or a mixture of Gaussians distributions [41], [51]. Other systems model background fluctuations by using more than one background image [8], [20].

The detected objects are then tracked in order to obtain their trajectories in the video sequence. This can be done by region matching, using feature tracking [14], by template or object matching [7], [13], [20], [33], or by Kalman filtering [30], [40], [59]. Region matching in consecutive frames is usually a simple operation if the person is isolated and unoccluded, but becomes more difficult under occlusions and with groups merging and/or splitting; these difficulties have been tackled in [27], [61]. Some feature tracking methods try to overcome the occlusion problem by tracking local features (provided that some small regions of the object remain unoccluded); this can be done using *scale-invariant feature transforms* (SIFT) [32] or the mean shift algorithm [14]. An additional operation has to be done in this case in order to associate the detected features with each object of interest. This is not a trivial task since there is object overlap and the number of detected features changes during the observation period [48]. Both techniques (region matching and feature tracking) can be used in a complementary way to improve the performance of the tracker [48]. Template matching is also used in some tracking methods [20]; however, it is often difficult to cope with changes in the visual appearance of the object to be tracked. Adaptive templates can partially overcome these difficulties but template updating is not a robust operation. In long range settings, the active regions detected in the image are often represented by their centroids and tracked using Kalman filtering [40]. In the presence of outliers produced by clutter, more sophisticated approaches are needed, such as the *multiple hypothesis tracker* [41].

B. Features

Human activity recognition is usually based on features, the choice of which depends on the particular application in hand and on the geometry of the camera and scene. In long range problems (e.g., outdoor surveillance), the most commonly used feature is the location (of the centroid) of the person in the scene [21], [40], [41], [51], since it is a stable and robust feature. Other features used include shape features, such as the silhouette, the star skeleton [13], and temporal templates [7].

In short range problems, additional features are often used, such as the position of the feet, hands and head [20], [59], 2-D or 3-D estimates of the human body segments [44], facial features [12], periodicity features [11].

In this paper, since our goal is long range surveillance, we focus on trajectories.

C. Activity Recognition

A formal definition of what is an activity depends strongly on the application domain considered and on the specific problem being addressed. In a surveillance application, examples of activities can be "walking", "running", "entering", "leaving", "fighting", or other activities which the users/owners of the system find of interest. At this point, we will not attempt a formal definition of activity classes other than say that the goal of activity classification is to classify the observed persons' behavior (as observed by the video camera) into one element of this set of classes.

Activity classification is usually based on generative probabilistic dynamical models of trajectories, with hidden Markov models (HMM, [42]), and variants thereof, being the most widely adopted tool. Being a generative model, an HMM can be used to produce synthetic sequences with the same probabilistic properties as the observations. Discriminative methods have also been used in activity recognition since they do not attempt to model the data, but only the decision; see [53], where a comparison between *maximum entropy Markov models* and *conditional random fields* (CRF) is carried out. A problem which remains unsolved to a large extent is the representation of interacting people and group activities. Two steps towards this direction can be found in [40], using *coupled HMMs*, and in [21], based on *Bayesian networks*. Other machine learning methods have also been used with some success, namely, neural networks [26], [47], *dynamic Bayesian networks* [60] and the so-called *abstract* HMM [30], [54], [55].

In [55], the scene is split into a set of square cells and the movement of pedestrians between neighboring cells is

described by Markov models (primitive behaviors). Complex behaviors are defined by the concatenation of primitive behaviors. This system is based on a hierarchical generative model as ours. However, there are several differences. First, space is discretized into a small number of cells appropriate for indoor scenes while we use continuous trajectories which are more appropriate for outdoor/far-field settings. Second, the system in [55] is event driven (switching can only occur when the person meets a landmark) while this restriction is not imposed in ours.

III. Human Activity Model

A. Class-Conditional Generative Model

The inputs to our human activity classifier are trajectories of the centers of mass of the people tracked in the video sequence. Formally, each trajectory is a length-n sequence of positions, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with $\mathbf{x}_t \in \mathbb{R}^2$. Consider that the activity class is represented by variable $a \in \{1, \ldots, A\}$; our classifier is based on class-conditional generative models for the trajectories, $p(\mathbf{x}|a)$, which are next described.

Since the trajectories are nonstationary, they are not easily described by a single dynamical model. Instead, we use a set of M dynamical models, each of which associated to a different motion direction and speed. Thus, we assume that $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is produced by a switched dynamical system

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{T}_{z_t} + \mathbf{Q}_{z_t}^{1/2}\mathbf{w}_t \tag{1}$$

where $z_t \in \{1, \ldots, M\}$ is the label of the low level model at time t, and $(\mathbf{w}_1, \ldots, \mathbf{w}_n)$ are independent samples of a zero-mean Gaussian random vector with identity covariance; the parameters of this system are $\{\mathbf{T}_1, \ldots, \mathbf{T}_M\}$, the mean displacement vectors of each model, and $\{\mathbf{Q}_1, \ldots, \mathbf{Q}_M\}$, the covariance matrices of the random displacements under each model. Naturally, model (1) specifies the conditional probability density of a trajectory $\mathbf{x}$, given a sequence of model labels $\mathbf{z} \in \{1, \ldots, M\}^n$, i.e.,

$$p(\mathbf{x}|\mathbf{z}) = p(\mathbf{x}_1)\prod_{t=2}^{N} \mathrm{N}\left(\mathbf{x}_t - \mathbf{x}_{t-1}|\mathbf{T}_{z_t}, \mathbf{Q}_{z_t}\right) \tag{2}$$

where $p(\mathbf{x}_1)$ is the probability density of the initial position and $\mathcal{N}(\mathbf{v}|\boldsymbol{\mu}, \mathbf{C})$ denotes a Gaussian probability density function of mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$, computed at $\mathbf{v}$. In this paper, we assume that the M low level dynamical models characterized by $\{\mathbf{T}_1, \ldots, \mathbf{T}_M\}$ and $\{\mathbf{Q}_1, \ldots, \mathbf{Q}_M\}$ are the same for all the classes (activities), this being the reason why $p(\mathbf{x}|\mathbf{z})$ is not conditioned on a, i.e., $p(\mathbf{x}|\mathbf{z}, a) = p(\mathbf{x}|\mathbf{z})$.

At the higher level, we assume that the label sequence $\mathbf{z} = (z_1, \ldots, z_n) \in \{1, \ldots, M\}^n$ of a trajectory of class a is a sample of a Markov chain with transition matrix $\mathbf{B}_a$ (of dimension $M \times M$), which is characteristic of each activity class. Formally

$$P(\mathbf{z}|a) = P(\mathbf{z}|\mathbf{B}_a) = P(z_1|a)\prod_{t=2}^{n} B^{(a)}_{z_{t-1}, z_t} \tag{3}$$

where $B^{(a)}_{ij}$ denotes the (i, j)th element of matrix $\mathbf{B}_a$.

The sequence $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is of course obtained from the video sequence, but $\mathbf{z} = (z_1, \ldots, z_n)$ is an unobserved (hidden) sequence. The class-conditional generative model is thus finally obtained by marginalizing with respect to the missing label sequence

$$p(\mathbf{x}|a) = \sum_{\mathbf{z}\in\{1,\ldots,M\}^n} p(\mathbf{x}, \mathbf{z}|a) = \sum_{\mathbf{z}\in\{1,\ldots,M\}^n} p(\mathbf{x}|\mathbf{z})P(\mathbf{z}|a). \tag{4}$$

The next sections describe how the parameters of the low level models, $\boldsymbol{\theta} = \{\mathbf{T}_1, \ldots, \mathbf{T}_M, \mathbf{Q}_1, \ldots, \mathbf{Q}_M\}$, and the transition matrices of the high level models, $\{\mathbf{B}_1, \ldots, \mathbf{B}_A\}$, are estimated from training data.

IV. Estimation of Low Level Model

The maximum likelihood estimate of all the parameters of our model can be obtained from a training set of labeled training trajectories, $\mathcal{T} = \{(\mathbf{x}^{(1)}, a_1), \ldots, (\mathbf{x}^{(N)}, a_N)\}$. Each pair $(\mathbf{x}^{(j)}, a_j)$ denotes that training trajectory $\mathbf{x}^{(j)} = (\mathbf{x}_1^{(j)}, \ldots, \mathbf{x}_{n_j}^{(j)})$ was generated by an activity of class $a_j \in \{1, \ldots, A\}$, and n_j is the length of the training trajectory $\mathbf{x}^{(j)}$.

Assuming that the training trajectories are independent, the log-likelihood function for the model parameters is given by

$$\begin{aligned}
&\log p\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}|a_1, \ldots, a_N, \boldsymbol{\theta}, \mathbf{B}_1, \ldots, \mathbf{B}_A\right) \\
&\quad = \sum_{j=1}^{N} \log p\left(\mathbf{x}^{(j)}|\boldsymbol{\theta}, a_j\right) \qquad (5) \\
&\quad = \sum_{j=1}^{N} \log \sum_{\mathbf{z}^{(j)}\in\{1,\ldots,M\}^{n_j}} p\left(\mathbf{x}^{(j)}|\mathbf{z}^{(j)}, \boldsymbol{\theta}\right) \\
&\qquad \times P\left(\mathbf{z}^{(j)}|\mathbf{B}_{a_j}\right). \qquad (6)
\end{aligned}$$

where $p(\mathbf{x}^{(j)}|\mathbf{z}^{(j)}, \boldsymbol{\theta})$ is given by (2) and $P(\mathbf{z}^{(j)}|\mathbf{B}_{a_j})$ is given by (3). Maximizing (6) with respect to $\boldsymbol{\theta}$ and $\{\mathbf{B}_1, \ldots, \mathbf{B}_A\}$ can be done using a slightly modified version of the Baum-Welch algorithm (BMA) [42]. Essentially, this maximization corresponds to estimating the parameters of a set of A hidden Markov models which share a common set of M Gaussian emission densities parameterized by $\boldsymbol{\theta} = \{\mathbf{T}_1, \ldots, \mathbf{T}_M, \mathbf{Q}_1, \ldots, \mathbf{Q}_M\}$, but with different transition matrices $\{\mathbf{B}_1, \ldots, \mathbf{B}_A\}$.

In this paper, we follow an alternative route motivated by the following considerations. Since the parameters of the low-level models, $\boldsymbol{\theta} = \{\mathbf{T}_1, \ldots, \mathbf{T}_M, \mathbf{Q}_1, \ldots, \mathbf{Q}_M\}$, are the same for all the activity classes, it makes sense to estimate them in an unsupervised way directly from a training set $\boldsymbol{\mathcal{X}} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$, containing unlabeled trajectories of all the activities. Under model (2), the observed increments in the jth training trajectory $\{\mathbf{d}_t^{(j)} = \mathbf{x}_t^{(j)} - \mathbf{x}_{t-1}^{(j)}, t = 2, \ldots, n_j\}$, are conditionally independent, given the (unobserved) low-level model labels, $\{\mathbf{z}_t, t = 2, \ldots, n\}$; thus, in the presence of these labels, estimating $\boldsymbol{\theta}$ would be trivial. Our proposal is to treat these increments as independent samples of a Gaussian mixture with parameters $\boldsymbol{\theta}$, that is, to ignore the Markovian nature of each sequence $\mathbf{z}^{(j)}$, for $j = 1, \ldots, N$. The crucial advantage of this approximation is that it allows direct application of mixture fitting algorithms; in particular, we adopt a technique that simultaneously performs model selection (i.e., estimates the number of low-level models M) [16].

Let $\mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_{N'}\}$ be the set of all the observed increments in all the training trajectories (the total number is $N' = \sum_{j=1}^{N}(n_j - 1)$). Under the assumption explained in the previous paragraph, we model this set of increments as independent samples from a Gaussian mixture with M components and parameters $\boldsymbol{\theta} = \{\mathbf{T}_1, \ldots, \mathbf{T}_M, \mathbf{Q}_1, \ldots, \mathbf{Q}_M\}$

$$\log p(\mathbf{d}_1, \ldots, \mathbf{d}_{N'}) = \sum_{j=1}^{N'} \log \sum_{m=1}^{M} \alpha_m \mathrm{N}(\mathbf{d}_j | \mathbf{T}_m, \mathbf{Q}_m) \quad (7)$$

where α_m is the weight of the mth component in the mixture.

To simultaneously estimate the low-level model parameters $\boldsymbol{\theta} = \{\mathbf{T}_1, \ldots, \mathbf{T}_M, \mathbf{Q}_1, \ldots, \mathbf{Q}_M\}$ and their number M, we use the criterion and algorithm proposed in [16]. Essentially, the method proposed in that paper uses a model selection criterion based on the *minimum message length* (MML) principle [57]. The MML criterion is then directly optimized using a modified version of the EM algorithm which is able to annihilate redundant components; thus, after being initialized with a large number of randomly placed components, that algorithm provides a parsimonious estimate of the mixture parameters. The method also avoids some of the well known drawbacks of the standard EM algorithm for mixtures; namely, the sensitivity to initialization and the need to avoid the boundary of the parameter space (when working with free covariance matrices) where the likelihood is unbounded. For more details, see [16].

An alternative, more classical approach, would be to use an HMM for each activity class, with the number of low level models selected by one of the well known model selection criteria, such as Akaike's *information criterion* (AIC) [2] or the *Bayesian information criterion* (BIC) [52], which is formally equivalent to the basic version of the *minimum description length* (MDL) criterion [45], or even the method [4] In that approach, the HMM parameters are estimated from the data using the EM algorithm (in this case the Baum-Welch algorithm), for a range of numbers of low level models; then, the model leading to the lowest value of the corresponding model selection criterion is selected. In the experimental results section, we will show that the shared low-level models obtained using the algorithm from [16] yield a higher activity classification accuracy, when compared to the standard HMM-based approach, thus demonstrating the adequacy of our approach to the problem addressed in this work.

V. Estimation of Activity Model

After estimating the parameters of the low-level models using the procedure presented in the previous section, it is necessary to estimate the transition matrices of each activity class, $\{\mathbf{B}_1, \ldots, \mathbf{B}_A\}$. These transition matrices are estimated from a set of labeled sequences $\mathcal{T} = \{(\mathbf{x}^{(1)}, a_1), \ldots, (\mathbf{x}^{(N)}, a_N)\}$.

In particular, we obtain the transition matrix estimate $\widehat{\mathbf{B}}_a$, for each $a \in \{1, \ldots, A\}$, from a set of N_a trajectories from class a, using a simplified version of the Baum-Welch algorithm [42]. Let us denote as $\mathcal{T}_a = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_a)}\}$ the set of N_a training trajectories from class a. Notice that, conditionally on the low-level model label sequence, the increments in these trajectories are independent [see (2)]. Thus, instead of the absolute positions in these trajectories, we consider these increments; let $\mathcal{D}_l = \{\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(N_a)}\}$, where (as in the previous section) $\mathbf{d}^{(j)} = \{\mathbf{d}_t^{(j)} = \mathbf{x}_t^{(j)} - \mathbf{x}_{t-1}^{(j)}, t = 2, \ldots, n_j\}$. We consider that the probability density of the first position, $p(\mathbf{x}_1)$ is know for all classes and needs not be estimated. We thus have an M-state hidden Markov model, with Gaussian emission densities with means and covariance matrices $\{\mathbf{T}_1, \ldots, \mathbf{T}_M, \mathbf{Q}_1, \ldots, \mathbf{Q}_M\}$. All these parameters, as well as the number of states M, were estimated using the unsupervised method presented in the previous section. To estimate $\mathbf{B}_a$, we simply use the Baum-Welch algorithm [42], but keeping frozen the emission density parameters.

VI. Classification

The classification problem can be formulated as follows: *given a new observed trajectory* $\mathbf{x}$, *classify it into the set of activities* $\{1, \ldots, A\}$. A trajectory from class a is modeled according to (1), where $\mathbf{z} = \{z_1, \ldots, z_n\}$ is a sample of a Markov model with transition matrix $\widehat{\mathbf{B}}_a$, estimated as explained in the previous section. This model allows computing the class-conditional likelihood of a trajectory $\mathbf{x}$, that is, $p(\mathbf{x}|\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{B}}_a)$, where $\widehat{\boldsymbol{\theta}}$ is the set of low level model parameter estimates, which is common to all the classes. Each class-conditional likelihood term $p(\mathbf{x}|\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{B}}_a)$, for $a = 1, \ldots, A$, is computed by running one forward/backward recursion of the Baum-Welch procedure, with the corresponding model parameter estimates $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{B}}_a$. The classification of the trajectory $\mathbf{x}$ is obtained by the *maximum a posteriori* (MAP) rule, i.e.,

$$\widehat{a} = \arg\max_a \left\{ p(\mathbf{x}|\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{B}}_a) P(a) \right\} \quad (8)$$

where $P(a)$ is the *a priori* probability of the activity a, herein taken simply as $P(a) = 1/A$. Thus, given the trajectory $\mathbf{x}$, the classifier requires running one forward-backward recursion (as in the Baum-Welch algorithm) under all candidate classes $\{1, \ldots, A\}$.

VII. Experiments

This section reports experimental results using both synthetic and real data. The video sequences of the real data were obtained by surveillance cameras located in a shopping center and on a university campus.

A. Synthetic Data

We first illustrate the performance of the algorithm with a synthetic example. This example intends to demonstrate the effectiveness of the approach in the case where the classes share exactly the same low level displacement statistics, only being different at the higher level (i.e., different transition matrices). Consider the trajectories from two classes depicted in red and green in Fig. 1. The low-level statistics of the two classes are roughly the same: 50% of horizontal displacements (right) and 50% of vertical (up) displacements, with the same mean and covariance: $\mathbf{T}_1 = [0.02\ 0]^T$ (horizontal displacement), $\mathbf{T}_2 = [0\ 0.02]^T$ (vertical displacement), $\mathbf{Q}_1 = \mathbf{Q}_2 = 10^{-3}\mathbf{I}$. The

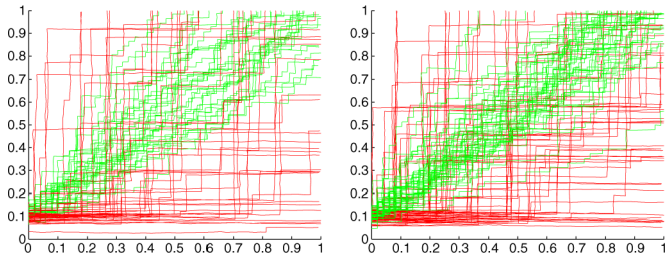


Fig. 1. Two activities sharing the same low level data, with different transition matrices. Training data (left), test data (right).

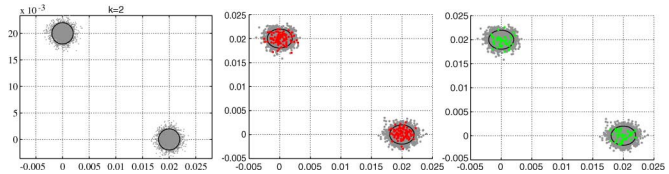


Fig. 2. Fitting Gaussian mixtures to the displacements in the synthetic scenario. Two low level models are correctly estimated in the example (left); type "red" activity (center); type "green" activity (right).

difference between the two classes resides only on the transition matrices which are, respectively, for the red and green trajectories

$$\mathbf{B}_1 = \begin{bmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{bmatrix} \quad \mathbf{B}_2 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}. \tag{9}$$

The trajectories of the red class have a low probability of switching between the two low level models ($\mathbf{B}_1$ is close to identity), while those of the green class have a 0.5 probability of switching at each instant. Notice also that the stationary distributions of these transition matrices are both equal to $[0.5\ 0.5]^T$, meaning that, on average, the trajectories of both classes perform the same number of vertical and horizontal displacements. Using this model, we generate 100 training trajectories (Fig. 1, left) and 100 test trajectories (Fig. 1, right).

The estimation of the low-level model parameters, i.e., $\widehat{\boldsymbol{\theta}} = \{\widehat{\mathbf{T}}_1, \widehat{\mathbf{T}}_2, \widehat{\mathbf{Q}}_1, \widehat{\mathbf{Q}}_2\}$ is shown in Fig. 2. As expected, a mixture of two Gaussian was fitted to the displacements; the corresponding mean vector and covariance estimates have errors of less than 0.1% with respect to the true parameters. The estimated transition matrices are

$$\widehat{\mathbf{B}}_1 = \begin{bmatrix} 0.9499 & 0.0501 \\ 0.0457 & 0.9543 \end{bmatrix} \quad \widehat{\mathbf{B}}_2 = \begin{bmatrix} 0.5412 & 0.4588 \\ 0.5033 & 0.4967 \end{bmatrix} \tag{10}$$

very close to the true underlying matrices. Finally, the classification accuracy obtained on the test data was 100%, showing that it was possible to distinguish trajectories from each class, using only the switching pattern.

B. Real Data

Fig. 3 shows the two scenarios considered in the experiments with real data, where different type of activities (to be detailed below) take place. Fig. 4 shows examples of observed trajectories.

Fig. 3. The two scenarios: (left) shopping center and (right) university campus.

Fig. 4. Trajectories observed in the shopping (top row) and in the campus (bottom row) scenarios.

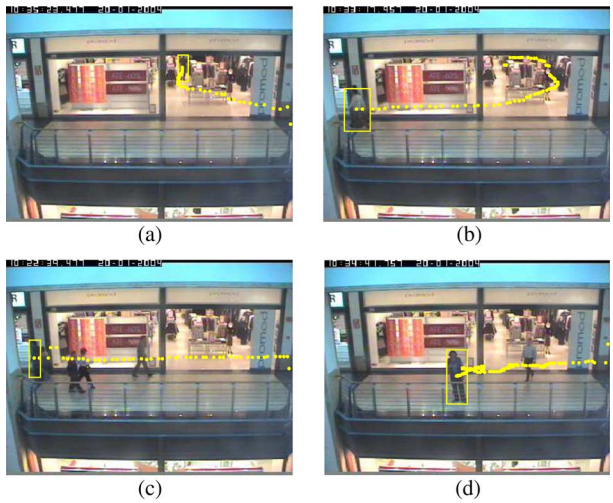


Fig. 5. Examples of the four activities defined for the shopping scenario: (a) entering, (b) leaving, (c) passing, and (d) browsing. All the dots belonging to the trajectory (centroid of the bounding box) as well as the last bounding box are plotted.

After observing many trajectories during several days, we defined a set of high level activities of interest for each scenario. In the shopping scenario, we have defined four main activities: *"entering"* the shop, *"leaving"* the shop, *"passing"* in front of the shop, and *"browsing"*. Fig. 5 shows examples of trajectories of these four classes.

For the university campus, we defined a set of seven high level activities: "entering building", "leaving building", "crossing park up", "crossing park down", "passing through", "walking along", and "wandering". Fig. 6 shows examples of trajectories

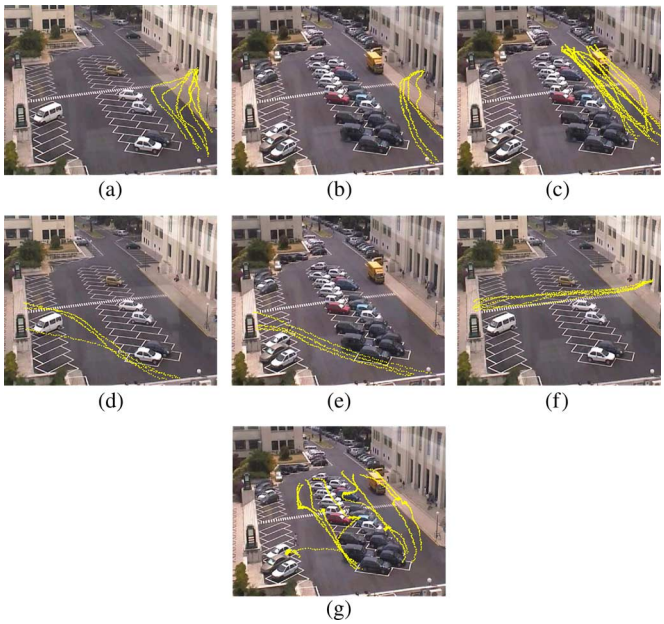


Fig. 6. Examples of trajectories from seven classes defined: (a) "entering building", (b) "leaving building", (c) "walking along", (d) "crossing (diagonally) park up", (e) "crossing (diagonally) park down", (f) "passing through", and (g) "wandering". Here, we just plot the centroid of the pedestrians' blobs.

for each of these seven classes. Each high level class is composed of a sequence of low level motion dynamical models, the parameters of which are estimated by the unsupervised learning scheme presented above. Recall that all of these trajectories may have the corresponding "mirror" activities, i.e., the same activity performed in the opposite direction.

C. Homography and Centroids

When tracking a person across the camera field of view, shape, position and speed are influenced by the perspective effect. This makes the training as well as the classification a space-varying task which is hard to model, since the observations (e.g. centroids) collected from the images depend both on the ongoing trajectory and on the viewing geometry. See, for instance, Fig. 5(a), (b) and (d); in these frames there are very small displacements. This happens due to the following different reasons: in Fig. 5(a) and (b), the small displacements are caused by the camera position, whilst in Fig. 5(d), the small displacements are originated by the low velocity of the pedestrian.

To achieve viewpoint invariance, all image measurements are projected onto a view orthogonal to the ground plane (*bird's eye* view), using a projective transformation (homography) between the image and a plane parallel to the ground. The parameters of this projection were obtained by considering a set of points in the scene with known ground-plane coordinates. The homography is defined as follows:

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{11}$$

where $[X\ Y]^T$ and $[x\ y]^T$ are the coordinates in the real world and in the image plane, respectively. Since the non singular homogeneous matrix H has 8 degrees of freedom, four points are needed to determine them uniquely.

Fig. 7. Two scenarios after the homographic transformation: (a) Homographic projection to the ground plane (Campus); (b) region of interest (Campus); (c) homographic in the shopping scenario.

In this work, we have adopted the centroids of the pedestrians' blobs as features. Ideally, one would like to work with points that lie on the ground plane, for instance, the pedestrians' feet. However, the centroids of the pedestrians' blobs are more stable and robust and less prone to segmentation errors [29]. Besides, the use of the ground plane points requires the assumption of the position and orientation of the camera with respect to the ground. Moreover, using the centroids, the pedestrian heights does not affect the ground plane alignment. Indeed, in our surveillance applications, the distance between the camera to the ground plane is very large in comparison with the first and second order statistics of the heights of the moving objects. Fig. 7(a) shows an image of the campus after this transformation, while Fig. 7(b) shows the region of interest on the transformed image, where all the trajectories take place. Fig. 7(c) shows the transformed image of the shopping.

From this point on, we will use the ground plane homography, to represent the positions of the bounding boxes.

D. Low-Level Model Estimates

Fig. 8 shows several estimates of the mixture components with different numbers of modes, for the shopping data. The mixture estimated by the algorithm from [16] has five components [Fig. 8(e)]; thus, we use five low level dynamical models, which have clear meanings: "stopped", "moving north", "moving south", "moving east", "moving west".

Fig. 9(a) and (b) shows two samples of the "leaving" activity. The dynamical models used to represent this activity are shown in Fig. 9(c) and (d) (red dots), which correspond to "moving south" and "moving east". The trajectory (red dots) are discontinuous due to occlusions. Although these activities belong to the same class, they exhibit significant variability as is clear in the displacements (red dots) in Fig. 9(c) and (d). Fig. 10 shows several trajectories of the "browsing" activity. In Fig. 10(a) and (b), the pedestrian "browses" outside the shop. In Fig. 10(c), the person "browses" outside and inside the shop.

Fig. 11 shows several estimates of the mixture components with different numbers of modes, for the campus data. The mix-

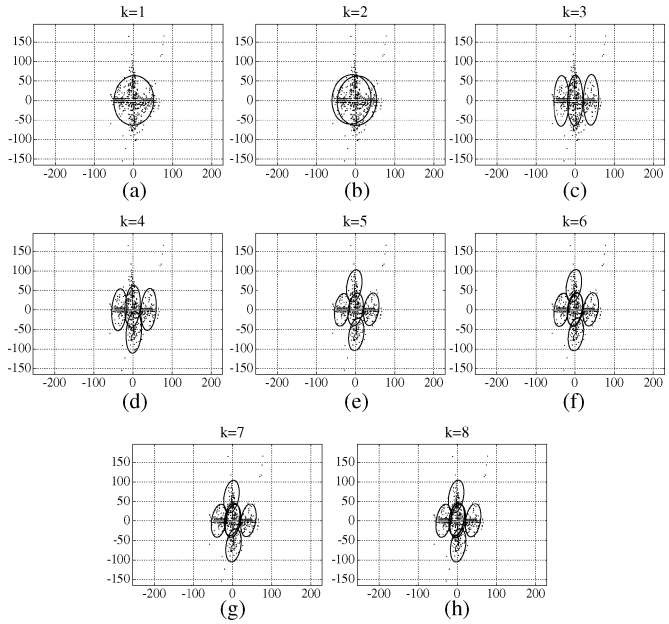


Fig. 8. Fitting a Gaussian mixture over the trajectories displacements in the shopping scenario; (a), (b), (c), (d), (e), (f), (g), and (h) show the estimates with 1, 2, 3, 4, 5, 6, 7, and 8 components, respectively. The dots are the trajectories displacements and the solid elipses are level curves of each component estimate.

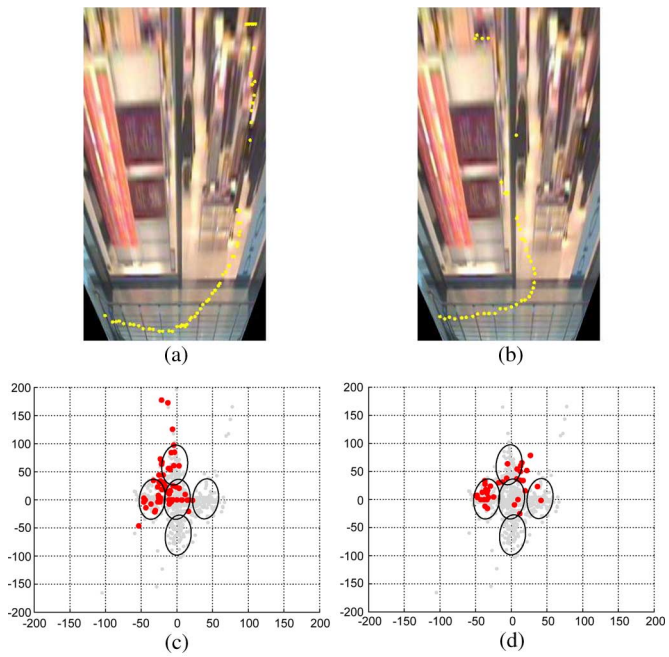


Fig. 9. Two examples of the "leaving" activity (a), (b); corresponding displacements (red dots) superimposed with gaussian mixtures previously estimated by the unsupervised learning scheme (c), (d). Recall that the y-axis are opposed between the image coordinates and the graphic coordinates; thus, the "upper" red dots in (c), (d), correspond to the "down" direction in (a), (b).

ture selected by the algorithm from [16] has nine components [Fig. 11(g)], indicating nine low level dynamical models, which can be labeled as follows: "stopped", "moving north", "moving south", "moving east", "moving west", "moving north-west", "moving north-east", "moving south-west", " moving south-east".

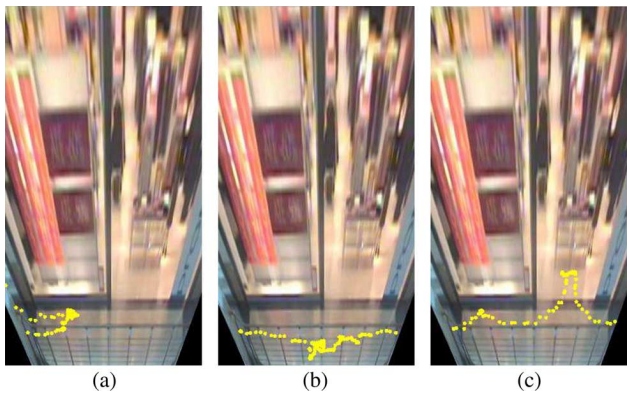


Fig. 10. Several trajectories of the "browsing" activity. Despite the considerable variability, our algorithm correctly recognizes the instances of this activity.

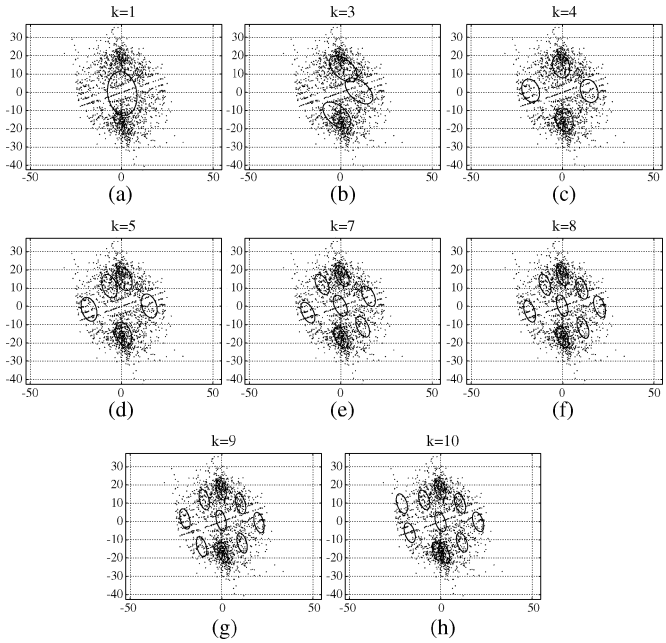


Fig. 11. Fitting Gaussian mixtures to the displacements in the campus scenario with different numbers of components. The dots are the trajectories displacements and the solid elipses are level curves of each component estimate.

E. Estimates of the Transition Matrices

Fig. 12 shows the transition matrices estimates $\widehat{\mathbf{B}}_a$ of the 7 campus activities, obtained with the supervised training procedure described in Section V, for the campus data. The matrices are of size 9×9, since 9 is the number of low level models selected for the campus data. In this figure, higher transition probabilities are represented by darker gray levels. The low level models are numbered from 1 to 9 in the following order: "north", "north-east", "east", "south-east", "south", "south-west", "west", "north-west", "stopped". From Fig. 12, we can see that the "entering" activity visits the "1", "2" and "3" low level models (darker columns). Similarly, the "leaving" activity visits low models "5", "6", and "7", since these activities are more irregular.

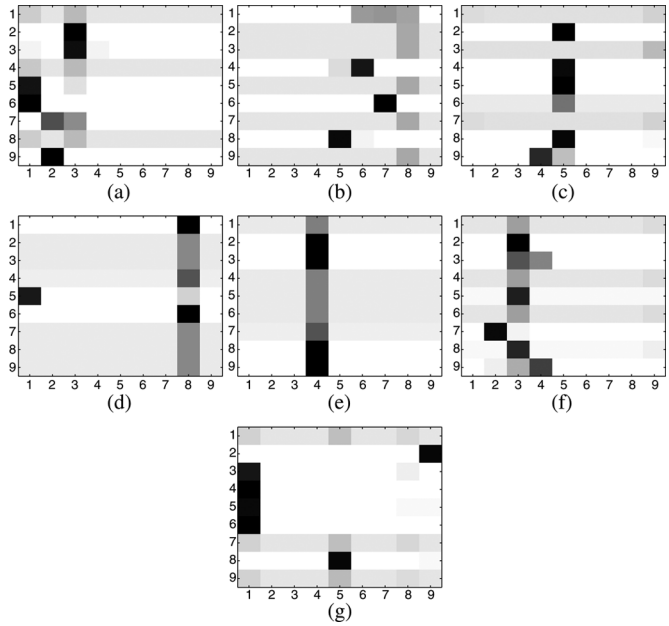


Fig. 12. Estimated transitions matrices for each activity of the SD-HMM: (a) entering; (b) leaving; (c) walking along; (d) crossing park up; (e) crossing park down; (f) passing through cars; (g) browsing.

VIII. Classification Results

A. Training/Testing Partitions

To assess the classification accuracy obtained with our approach, we have considered three different procedures for splitting the available data into training and testing sets: 1) a single training/testing splitting; 2) a complete p-fold cross validation; 3) a random sub-sampling validation.

The first procedure is simple and uses an arbitrary splitting of all the available data set into two disjoint sets. In the second procedure, we consider the set containing all the available trajectories (of all the activities) in a random order; we then perform a full ten-fold cross validation analysis. Finally, in the third procedure, we randomly select 10% of the data from each activity for testing and use the remaining for training.

B. Classification Results for the Shopping Mall Dataset

In the case of the first experimental procedure (a fixed testing/training partition, as described in Section VIII-A), the training set contains three trajectories from each class. The test set is composed of 62 trajectories. Regarding classes, we had to take in consideration the "mirror" activities. A mirror activity is the same activity but with some part(s) of the trajectory performed in the opposite direction. For instance, in the shopping scenario, if the "entering" activity follows the "left" then the "up" directions, the corresponding mirror activity follows the "right" then the "up" directions. For both scenarios, we define mirrors for all classes, with the exception of the "browsing" class, which does not have a particular direction in the image. In Table I, which shows the accuracy obtained in this experiment, we see that some errors occur between the "browsing" and "passing" activities. This may be due to the fact that "browsing" is characterized by periods of very low or null velocity; of course a person passing in front of the shop may have periods of low velocity, thus being confused with a browsing activity.

TABLE I
Confusion Matrix for the Shopping Scenario With the First Strategy

	Classification			
True classes	E	L	P	B
"entering shop" (E)	100%	0	0	0
"leaving shop" (L)	0	100%	0	0
"passing shop" (P)	0	0	94.75%	5.25%
"browsing" (B)	0	8.3%	0	91.7%

TABLE II
Confusion Matrix for the Shopping Scenario With the Third Strategy

	Classification			
True classes	E	L	P	B
"entering shop" (E)	83%	17%	0	0
"leaving shop" (L)	0	100%	0	0
"passing shop" (P)	0	0	100%	0
"browsing" (B)	10.9%	7.6%	0	81.5%

Using the second experimental procedure (ten-fold cross validation) an accuracy of 100% was obtained for the activities "entering", "leaving" and "passing", and 86% accuracy for the "browsing" activity (with all the errors corresponding to misclassifications as "entering" activity).

Finally, with the third experimental procedure (random partition of the data from each class), the results obtained are reported in Table II.

C. Classification Results for the Campus Dataset

In this case, the available data contains 200 trajectories. As above, we applied the three experimental procedures described in Section VIII-A. The results obtained with the first procedure are reported in Table III. The second experimental procedure leads to 100% accuracy for all the activities. Finally, the results obtained with the third strategy are shown in Table IV.

D. Comparison With the HMM Approach

We conclude the experimental study with a comparison between our approach and a standard HMM-based classifier, with the model order selected by the MDL/BIC or AIC criteria. For this comparative study, we adopted the first experimental procedure (fixed splitting of the data into training and testing sets). The overall error rates shown in Table V confirm that the adopted method from [16] yields a set of models with better classification accuracy. This table also shows that BIC/MDL outperforms AIC in this problem. More detail on the performance of the HMM-BIC/MDL approach is found in Tables VI and VII.

IX. Conclusion

In this work, we have presented a framework for modeling and recognition of human trajectories in surveillance applications. The method uses switched dynamical models (low level models), each of which describes a particular type of motion. The low level models are shared by all the classes/activities, thus are learnt in an unsupervised way, using a method which

TABLE III
CONFUSION MATRIX FOR THE UNIVERSITY CAMPUS SCENARIO WITH THE FIRST STRATEGY

	Classification						
True classes	E	L	CPU	CPD	PT	WA	W
"entering building" (E)	100%	0	0	0	0	0	0
"leaving building" (L)	0	100%	0	0	0	0	0
"crossing park up" (CPU)	0	0	100%	0	0	0	0
"crossing park down" (CPD)	0	0	0	95%	5%	0	0
"passing through" (PT)	5.85%	0	0	5.85%	88.3%	0	0
"walking along" (WA)	0	6.67%	0	0	0	93.3%	0
"wandering" (W)	0	0	0	0	0	0	100%

TABLE IV
CONFUSION MATRIX FOR THE UNIVERSITY CAMPUS SCENARIO FOR THE THIRD STRATEGY

	Classification						
True classes	E	L	CPU	CPD	PT	WA	W
"entering building" (E)	93%	0	0	0	0	0	7%
"leaving building" (L)	0	84.6%	0	0	0	0	15.3%
"crossing park up" (CPU)	0	25%	75%	0	0	0	0
"crossing park down" (CPD)	0	0	0	100%	0	0	0
"passing through" (PT)	2.8%	0	0	0	97.2%	0	0
"walking along" (WA)	0	0	0	0	0	100%	0
"wandering" (W)	0	0	0	0	0	0	100%

TABLE V
OVERALL ERROR RATES FOR THE THREE MODEL SELECTION STRATEGIES CONSIDERED

	Shopping	**Campus**
Proposed Approach	3.70%	3.29%
HMM + BIC/MDL	5.56%	7.24%
HMM + AIC	16.67%	8.55%

TABLE VI
PERFORMANCE USING HMM-BIC/MDL FOR THE SHOPPING SCENARIO

	Classification			
True classes	E	L	P	B
"entering shop" (E)	100%	0	0	0
"leaving shop" (L)	0	100%	0	0
"passing shop" (P)	0	0	100%	0
"browsing" (B)	8.33%	16.67%	0	75%

also selects the number of models. The high level models (probabilistic switching) are separately estimated for each class. The overall model resembles an HMM-based classifier, using a bank of common dynamical models at the lower level.

The experimental results reported validate the method by showing that this framework leads to a good classification accuracy for surveillance applications. Future work will include more complex activities and extension to other applications.

One possible work direction is the inclusion of an additional level to deal with complex activities, in the spirit of hierarchical HMMs. Each complex activity would be the concatenation of several activities, each of which described by one SD-HMM presented in this paper. The model would then have three levels: i) low level, (ii) high level (activity model), and iii) a third level which would handle the complex activities. Another interesting issue to be studied concerns the discrimination of the same type of activities occurring in different regions of the image. The low level (displacements) information is the same but the activities may have different meanings. One way to deal with this problem may be to use regions/cells with special meaning.

APPENDIX A
TRACKING PROCEDURE

Although the tracking procedure is beyond the scope of the paper, we briefly describe how the trajectories are obtained. There has been considerable work on tracking systems, namely based on features, edges, and boundaries [24], [28], [50], [58]. However, for our domain of application, the small size of the targets prevents the use of feature-based approaches, especially in the university campus scenario. In order to reach a common procedure for both scenarios, the system herein implemented consists of the following blocks: *i*) region detection, and *ii*) region association.

The first block detects the active regions in the image using the *Lehigh Omnidirectional Tracking System* (LOTS) algorithm [8], since it is considered to be amongst the best for surveillance applications [38].

The tracking block connects active regions detected in consecutive frames. We assume that objects move slowly in the scene, and, therefore, the corresponding regions should overlap. Furthermore, we also assume that the object motion can be predicted and the prediction error should be small.

Both criteria are used to define an association cost for all pairs of regions detected in consecutive frames. Let $\mathbf{x}_{t-1}^{i}, \mathbf{x}_{t}^{j}$ denote the centroids of regions i, j in consecutive frames. If these regions overlap in the image domain, the association cost is given by

$$\mathcal{C}_t(i,j) = \left\| \mathbf{x}_t^j - \left(\mathbf{x}_{t-1}^i + \mathbf{v}_{t-1}^i\right) \right\|$$

where $\mathbf{v}_{t-1}^{i}$ is the velocity estimate of the ith region detected in frame $t-1$. On the other hand, the cost $\mathcal{C}_t(i,j)$ is infinite if there is no overlap.

TABLE VII
PERFORMANCE USING HMM-BIC/MDL FOR THE CAMPUS SCENARIO

	Classification						
True classes	E	L	CPU	CPD	PT	WA	W
"entering building" (E)	94.73%	0	0	0	0	0	5.27%
"leaving building" (L)	0	100%	0	0	0	0	0
"crossing park up" (CPU)	0	0	100%	0	0	0	0
"crossing park down" (CPD)	0	0	0	92.5%	7.5%	0	0
"passing through" (PT)	5.85%	0	0	0	100%	0	0
"walking along" (WA)	0	6.67%	0	0	6.66%	53.33%	40%
"browsing" (B)	0	0	0	0	0	0	100%

After computing the cost matrix $\mathcal{C}_t = [\mathcal{C}_t(i,j)]$, we associate pairs of regions using the *mutual favorite pairing criterion* i.e., we associate a pair of regions (p,q) iff $\mathcal{C}_t(p,q)$ is the minimum cost value in line i and column j. More sophisticated algorithms could be used to perform this task, e.g., using optimization in graphs [49], [56]. The method adopted in this paper is simple and conservative in the sense that it does not associate regions in ambiguous cases in which multiple interpretations are possible (e.g., in region merging and splitting).

REFERENCES

[1] J. Aggarwal and Q. Cai, "Human motion analysis: A review," *Comput. Vis. Imag. Understand.*, vol. 73, no. 3, pp. 428–440, 1999.

[2] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Inf. Theory*, 1973, pp. 267–281.

[3] D. Ayers and M. Shah, "Monitoring human behavior from video taken in an office environment," *Image Vis. Comput.*, vol. 19, no. 12, pp. 833–846, 2001.

[4] M. Bicego, V. Murino, and M. A. T. Figueiredo, "A sequential pruning strategy for the selection of the number of states in hidden Markov models," *Pattern Recognit. Lett.*, vol. 24, no. 9–10, pp. 1395–1407, 2003.

[5] C. Bishop, *Neural Networks for Pattern Recognition*. Cambridge, U.K.: Cambridge Univ. Press, 1995.

[6] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with integrated classification likelihood," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 7, pp. 719–725, Jul. 2000.

[7] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[8] T. Boult, R. Micheals, X. Gao, and M. Eckmann, "Into the woods: Visual surveillance of non-cooperative camouflaged targets in complex outdoor settings," *Proc. IEEE*, vol. 89, no. 10, pp. 1382–1402, Oct. 2001.

[9] C. Bregler, "Learning and recognising human dynamics in video sequences," *IEEE Int. Conf. Computer Vision Pattern Recognition*, pp. 568–574, 1997.

[10] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[11] F. Cheng, W. Christmas, and J. Kittler, "Periodic human motion description for sports video databases," presented at the 17th Int. Conf. Pattern Recognition, 2004.

[12] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. Huang, Facial Expression Recognition From Video Sequences: Temporal and Static Modeling 2003 [Online]. Available: http://citeseer.ist.psu.edu/article/cohen03facial.html

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

[14] D. Comaniciu, V. Ramesh, and P. Meer, "Kernal-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–577, May 2003.

[15] A. Dempster, M. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM-algorithm," *J. Roy. Statist. Soc. B*, no. 39, pp. 1–38, 1977.

[16] M. Figueiredo and A. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.

[17] L. Fuentes and S. Velastin, "People tracking in surveillance applications," presented at the 2nd IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, Kauai, HI, 2001.

[18] I. Gath and B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 773–781, Jul. 1989.

[19] D. Gavrilla, "The visual analysis of human movement: A survey," *Comput. Vis. Imag. Understand.*, vol. 73, no. 1, pp. 82–98, 1999.

[20] I. Haritaoglu, D. Harwood, and L. Davis, "W^4: Real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809–830, Aug. 2000.

[21] S. Hongeng, R. Nevatia, and F. Bremond, "Video-based event recognition: Activity representation and probabilistic recognition methods," *Comput. Vis. Image Understand.*, pp. 129–162, 2004.

[22] T. Horprasert, D. Harwood, and L. Davis, "A robust background subtraction and shadow detection," in *Proc. 4th Int. Conf. ACCV*, 2000, vol. 1, pp. 983–988.

[23] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Syst. Cybern. C, Appl. Rev.*, vol. 34, no. 3, pp. 334–352, 2004.

[24] D. Huttenlocher, J. Noh, and W. Rucklidge, "Tracking nonrigid objects in complex scenes," presented at the Int. Conf. Computer Vision, Berlin, Germany, 1993.

[25] Y. Ivanov and A. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 852–872, Aug. 2000.

[26] N. Johnson and D. Hogg, "Learning the distribution of object trajectories for event recognition," *Image Vis. Comput.*, vol. 14, no. 8, pp. 609–615, Aug. 1996.

[27] P. Jorge, A. Abrantes, J. Lemos, and J. Marques, "Long term tracking of pedestrians with groups and occlusions," in *Bayesian Network Technologies, Applications, and Graphical Models*, A. Mittal and A. Kassim, Eds. New York: Hershey, 2007.

[28] D. Koller and H. Nagel, "Model-based object tracking in monocular image sequences of road traffic scenes," *Int. J. Comput. Vis.*, vol. 10, no. 3, pp. 257–281, 1993.

[29] L. Lee, R. Romano, and G. Stein, "Monitoring activities from multiple video streams: Establishing a common coordinate frame," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 758–767, Aug. 2000.

[30] L. Liao, D. Fox, and H. Kautz, "Learning and inferring transportation routines," presented at the National Conf. Artificial Intelligence (AAAI-04), 2004.

[31] A. Lipton, H. Fujiyoshi, and R. Patil, "Moving target classification and tracking from real-time video," presented at the DARPA Image Understanding Workshop, 1998.

[32] D. Lowe, "Distinctive image features from scale—Invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

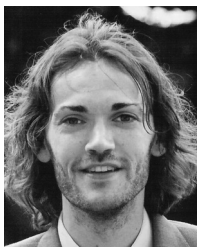
[33] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Comput. Vis. Image Understand.*, vol. 80, no. 1, pp. 42–56, 2000.

[34] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.

[35] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.

[36] T. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comp. Vis. Imag. Understand.*, vol. 104, pp. 90–126, 2006.

[37] F. Muri, "Comparaison d'algorithmes d'identification de chaînes de Markov cachées et application à la déctection de régions homogènes dans les séquences ADN," Ph.D. dissertation, Univ. René Descartes, France, Paris 5.
[38] J. Nascimento and J. Marques, "Performance evaluation of object detection algorithms for video surveillance," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 761–774, Apr. 2006.
[39] J. Nascimento, M. Figueiredo, and J. Marques, "Independent increment processes for human motion recognition," *Comput. Vis. Image Understand.*, vol. 109, no. 2, pp. 126–138, 2008.
[40] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.
[41] I. Pavlidis, V. Morellas, and P. Tsiamyrtzis, "Urban surveillance systems: From the laboratory to the commercial world," *Proc. IEEE*, vol. 89, no. 10, pp. 1478–1497, Oct. 2001.
[42] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[43] A. Raftery, "Bayesian model selection in social research," *Sociol. Meth.*, vol. 25, pp. 111–196, 1995.
[44] D. Ramanan, D. Forsyth, and A. Zisserman, "Tracking people by learning their appearance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 65–81, Jan. 2007.
[45] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific, 1989.
[46] S. Roberts, "Parametric non-parametric unsupervised cluster analysis," *Pattern Recognit.*, vol. 30, no. 2, pp. 261–272, 1997.
[47] M. Rosenblum, Y. Yacoob, and L. Davis, "Human expression recognition from motion using a radial basis function network architecture," *IEEE Trans. Neural Netw.*, vol. 7, no. 5, pp. 1121–1138, May 1996.
[48] D. Rowe, "Towards Robust Multiple-Target Tracking in Unconstrained Human-Populated Environments," Ph.D. dissertation, Univ. Autònoma de Barcelona, Barcelona, Spain, Dec. 2007.
[49] K. Shafique and M. Shah, "A noniterative greedy algorithm for multiframe point correspondence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 1, pp. 51–65, Jan. 2005.
[50] S. Smith and J. Brady, "ASSET-2: Real-time motion segmentation and shape tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, pp. 814–820, 1995.
[51] C. Stauffer, W. Eric, and L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.
[52] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.
[53] T. Truyen, H. Bui, and S. Venkatesh, "Human activity learning and segmentation using partially hidden discriminative models," in *Proc. Workshop Human Activity Recognition and Modelling*, 2005, pp. 87–95.
[54] H. H. Bui, D. Q. Phung, and S. Venkatesh, "Hierarchical hidden Markov models with general state hierarchy," presented at the AAAI, 2004.
[55] N. Nguyen and S. Venkatesh, "Discovery of activity structures using the hierarchical hidden Markov model," in *British Machine Vision Conf.*, 2006.
[56] C. J. Veenman, M. J. T. Reinders, and E. Backer, "Resolving motion correspondence for densely moving points," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 1, pp. 54–71, Jan. 2001.
[57] C. Wallace and P. Freeman, "Estimation and inference via compact coding," *J. Roy. Statist. Soc. B*, vol. 49, no. 3, pp. 241–252, 1987.
[58] J. Woodfill and R. Zabih, "An algorithm for real-time tracking of nonrigid objects," in *Proc. Nat. Conf. AI*, 1991, pp. 718–723.
[59] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, 1997.
[60] T. Xiang and S. Gong, "Beyond tracking: Modelling activity and understanding behaviour," *Int. J. Comput. Vis.*, vol. 1, no. 67, pp. 21–51, 2006.
[61] T. Zhao and R. Nevatia, "Tracking multiple humans in complex scenarios," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1208–1221, Sep. 2004.



Jacinto C. Nascimento (M'06) received the E.E. degree from the Instituto Superior de Engenharia de Lisboa, and the M.Sc. and Ph.D. degrees from the Instituto Superior Técnico, Lisbon, in 1995, 1998, and 2003, respectively.

Presently, he is a Posdoctoral Researcher with the Institute for Systems and Robotics (ISR), IST. His research interests include image processing, shape tracking, robust estimation, medical imaging, and video surveillance. He has coauthored over 40 publications in international journals and conference proceedings (many of which for the IEEE), has served on program committees of many international conferences, and has been a reviewer for several international journals.

Mário A. T. Figueiredo (S'87–M'95–SM'00–F'10) received the E.E., M.Sc., Ph.D., and "Agregado" degrees in electrical and computer engineering, all from the Instituto Superior Técnico (IST), the Engineering School of the Technical University of Lisbon (TULisbon), Portugal, in 1985, 1990, 1994, and 2004, respectively.

Since 1994, he has been with the faculty of the Department of Electrical and Computer Engineering, IST. He is also area coordinator at the Instituto de Telecomunicações, a private not-for-profit research institution. His scientific interests include image processing and analysis, computer vision, statistical pattern recognition, and statistical learning.

Dr. Figueiredo received the 1995 Portuguese IBM Scientific Prize and the 2008 UTL/Santander-Totta Scientific Prize. In 2008, he was elected a Fellow of the International Association for Pattern Recognition. He is a member of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee and is/was associate editor of the following journals: IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MOBILE COMPUTING, *Pattern Recognition Letters, and Signal Processing*. He is/was guest Co-Editor of special issues of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He was a Co-Chair of the 2001 and 2003 Workshops on Energy Minimization Methods in Computer Vision and Pattern Recognition and program/technical committee member of many international conferences.

Jorge S. Marques received the E.E., M.Sc., and Ph.D. degrees, and the aggregation title from the Technical University of Lisbon, Lisbon, Portugal, in 1981, 1984, 1990, and 2002, respectively.

Currently, he is an Associate Professor with the Electrical and Computer Engineering Department, Instituto Superior Técnico, Lisbon, and a Researcher at the Institute for Systems and Robotics. He has published over 140 papers in international journals and conferences and he is the author of the book Pattern Recognition: Statistical and Neural Methods (IST Press, 2005, 2nd ed., in Portuguese). His research interests are in the areas of statistical image processing, shape analysis, and pattern recognition.

Dr. Marques was the Co-Chairman of the IAPR Conference IbPRIA 2005 and President of the Portuguese Association for Pattern Recognition from 2001 to 2003. His research interests are in the areas of statistical image processing, shape analysis, and pattern recognition.