HAREM 2005 - International Workshop on Human Activity Recognition and Modelling,
Oxford, UK, September 2005

Segmentation and Classification of Human Activities*

J.C. Nascimento[1] M. A. T. Figueiredo[2] J. S. Marques[3]
jan@isr.ist.utl.pt mtf@lx.it.pt jsm@isr.ist.utl.pt
[1,3]Instituto de Sistemas e Robótica [2]Instituto de Telecomunicações
Instituto Superior Técnico
1049-001 Lisboa
PORTUGAL

Abstract

This paper describes an algorithm for segmenting and classifying human activities from video sequences of a shopping center. These activities comprise entering or exiting a shop, passing, or browsing in front of shop windows. The proposed approach recognizes these activities by using a priori knowledge of the layout of the shopping view. Human actions are represented by a bank of switch dynamical models, each tailored to describe a specific motion regime. Experimental tests illustrate the effectiveness of the proposed approach with synthetic and real data.

Keywords: Surveillance, Segmentation, Classification, Human Activities, Minimum Description Length.

1 Introduction

The analysis of human activities is an important computer vision research topic with applications in surveillance, e.g. in developing automated security applications. In this paper, we focus on recognizing human activities in a shopping center.

In commercial spaces, it is common to have many surveillance cameras. The monitor room is usually equipped with a large set of monitors which are used by a human operator to watch over the areas observed by the cameras. This requires a considerable effort of the human operator, who has to somehow multiplex his/her attention. In recent years a considerable effort was devoted to develop automatic surveillance systems providing information about which activities take place in a given space. With such a system, it would be possible to monitor the actions of individuals, determining its nature and discerning common activities from inappropriate behavior (for example, standing for a large period of time at the entrance of a shop, fighting).

In this paper, we aim at labelling common activities taking place in the shopping space. [1] Activities are recognized from motion patterns associated to each person tracked by the system. Motion is described by a sequence of displacements of the 2D centroid (mean position) of each person's blob. The trajectory is modelled by using multiple dynamical models with a switching mechanism. Since the trajectory is described by its appearance, we compute the statistics for the identification of the dynamical models involved in a trajectory.

The rest of the paper is organized as follows. Section 2 deals with related work. Section 3, describes the statistical activity model. Section 4 derives the segmentation algorithm. Section 5 reports experimental results with synthetic data and real video sequences. Section 6 concludes the paper.

2 Related Work

The analysis of human activities has been extensively addressed in several ways using different types of features and inference methods. Typically, a set of motion features is extracted from the video signal and an inference model is used to classify it into one of *c* possible classes.

For example in [16] the human body is approximated by a set of segments and atomic activities are then defined as vectors of temporal measurements which capture the evolution of the five body parts. In other works the human body is simply represented by the mass center of its active region (blob) in the image plane [12] or the body blob as in [4]. The activity is then represented by the trajectory obtained from the blob center, or from the correspondence of body blob regions respectively.

Other works try to characterize the human activity directly from the video signal without segmenting the active regions. In [2] human activities are characterized by temporal templates. These templates try to convey information about "where" and "how" motion is performed. Two templates are created: a binary motion-energy-image which represents where the motion has occurred in the whole sequence, and a scalar motion-history-image which represents

*This work was partially supported by FCT under project CAVIAR(IST-2001-37540)

[1] This work is integrated in project CAVIAR, which has the general goal of representing and recognizing contexts and situations.An introduction and the main goals of the project can be found in http://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm

how motion occurs for each activity. Motion patterns have also been used in [9] based on the concept of "recency". This work integrates several frames into a single image, assigning higher weights to the most recent frames. In [10], the human motion is characterized by the optical flow.

Several inference techniques have been used for the recognition of human activities using static and dynamic techniques. In [12] a single-person or person-to-person interactions are modelled by Hidden Markov Models (HMMs) and Coupled Hidden Markov Models (CHMMs). Both techniques are used to characterize the evolution of the person mass center along the video sequence. In [4] a Bayesian networks are used to for making inference about the events. In [11] activities are modelled using banks of switched dynamic models each of which tailored to a specific motion regime.

Geometric constraints have also been used e.g., using the layout of the surveillance region [13, 3]. In [1, 3] Finite State Machines (FSM) are used for gesture and activity recognition. The later uses prior knowledge about the scene, where regions of interest are defined (e.g., entrances and exits).

When the human motion is characterized by global features static pattern recognition methods can be used to classify the human activities. In [15] neural networks are used for this purpose.

The previous methods have been used to deal with single pedestrians or a very limited number of pedestrians [12]. To deal with the interaction among multiple pedestrians Bayesian networks have been proposed [8] since they are able to represent the dependencies among several random variables.

3 Statistical Model

We represent the human activity by the trajectory of its centroid. The time evolution of this feature is modelled by a dynamical model. Since a single model may not suffice to describe an entire trajectory, we use multiple dynamical models and a switching mechanism.

In this paper, a trajectory will be represented by a sequence of 2D locations, $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, with $\mathbf{x}_t \in I\!R^2$. We assume that the trajectory is the output of a bank of switched dynamical systems of the form

$$\mathbf{x}_t - \mathbf{x}_{t-1} = \Delta\mathbf{x}_t = \mu_{k_t} + w_t, \tag{1}$$

where $k_t \in \{1, \ldots, c\}$ is the label of the active model at time instant t, μ_{k_t} is a (model-dependent) displacement vector, and the $w_t \sim \mathcal{N}(0, Q_{k_t})$ are independent Gaussian random variable, with covariances Q_{k_t}.

Since the observations are $\{\Delta\mathbf{x}_t; t \in \mathbb{N}\}, \Delta\mathbf{x}_t \in \mathbb{R}^d$ (d is the dimension of the observation vector), instead of $\mathbf{x}_t$, equation (1) describes an independent increment process, given k_t, as shown in Fig. 1

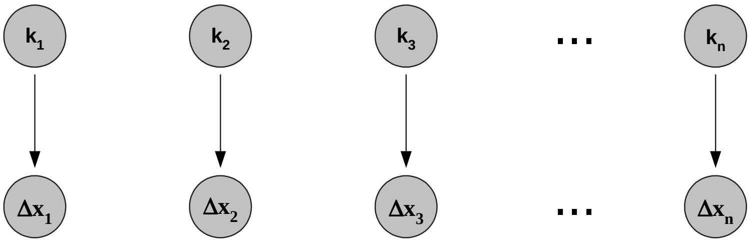


Figure 1: Architecture of the proposed approach.

Finally we assume that the sequence of model labels is composed of T constant segments: $\{k_1, \ldots, k_1, k_2, \ldots, k_2, \ldots, k_T, \ldots, k_T\}$.

4 Segmentation and classification Algorithm

In order to segment and classify the different activities, we first observed that all trajectories concerning a common activity follow a typical route. Fig. 2 shows trajectories corresponding to a person entering a shop (left), leaving a shop (middle) or just passing in front of a shop (right).

This work demonstrates that elementary actions such as: "moving upwards", "stopped", "moving downwards", "moving left" and "moving right" (i.e., $\sharp\mathscr{M} = 5$), are representatives of the trajectories. The underlying idea is: *given a test trajectory* $\mathbf{x}_t = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, *segment it into its elementary actions and classify the activity*. The number of segments will depend on the activity being considered, as described later.

4.1 Model Parameter Estimation

To segment and classify a given trajectory we have to previously obtain the parameters of each dynamic model. To accomplish this, we collect tens of trajectory samples from each model.

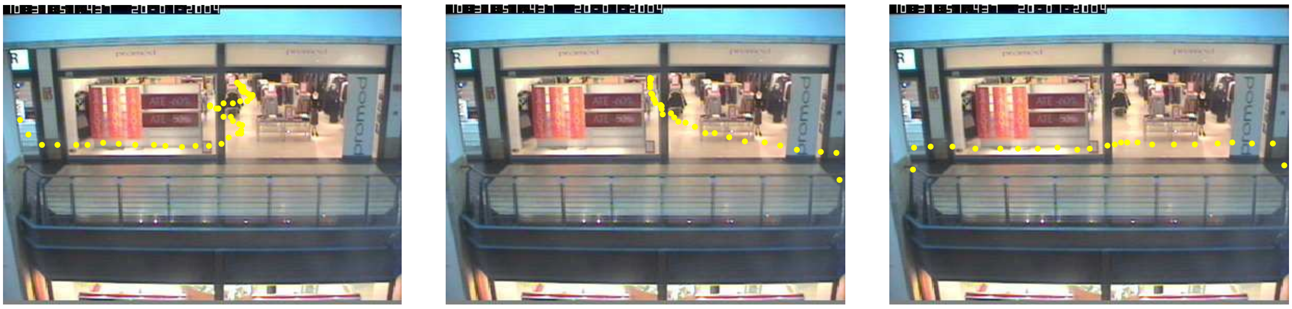


Figure 2: Examples of three different activities (entering, exiting, passing).

From $\mathbf{x}_t$ we can obtain $\Delta\mathbf{x}_t^i$, where $\Delta\mathbf{x}_t^i$ contains the displacements of $\mathbf{x}_t$ known to have been generated by the ith model. Defining $\Delta\mathbf{X}^i = \{\Delta\mathbf{x}_1^i, \Delta\mathbf{x}_2^i, \dots, \Delta\mathbf{x}_N^i\}$ as the vector containing all the displacements in ith model of the training set, we have, for the ith model:

$$\hat{\mu}_i = \frac{1}{\sharp\Delta\mathbf{X}^i}\sum \Delta\mathbf{X}_t^i, \qquad \hat{Q}_i = \frac{1}{\sharp\Delta\mathbf{X}^i}\sum (\Delta\mathbf{X}^i - \hat{\mu}_i)(\Delta\mathbf{X}^i - \hat{\mu}_i)^T, \tag{2}$$

where $\hat{\mu}_i$ and $\hat{Q}_i$ are standard estimates of the mean and the covariance matrix respectively.

4.2 Segmentation and Classification

Having defined the set of models and the corresponding parameters, one can now classify a test trajectory $\mathbf{x}_t$. One way to attain this goal is to compute the likelihood of $\mathbf{x}_t$ into the model space. In this paper, the activity depends on the number of the model switchings. In Fig. 2, we see that “passing” can be described by using just one model. The activities “entering” and “exiting” can be described by using two dynamical models. The fourth activity considered “browsing”, requires three models to be described; we define “browsing” when the person is walking, stop to see the shop-window and restarts walking. This behavior was observed in all the other samples of the activities which come about in this context. This means that we have to estimate the time instants in which the model switching happens.

Assuming that the sequence $\mathbf{x}_t$ has n samples and is described by T segments (and T is known) the log-likelihood is

$$L(m_1,\dots,m_T,t_1,\dots,t_{T-1}) = \log p(\Delta\mathbf{x}_1,\dots,\Delta\mathbf{x}_n \mid m_1,m_2,\dots,m_T,t_1,t_2,\dots,t_{T-1}) \tag{3}$$

where $m_1,\dots,m_T$ is the sequence of model labels describing the trajectory and t_i for $i = 1,\dots,T-1$ is the time instant when switching from model m_i to m_{i+1} occurs. If $T = 1$, there is no switching.

Due to the conditional independence assumption underlying (1), the log-likelihood can be written as

$$\begin{aligned} &L(\Delta\mathbf{x}_1,\dots,\Delta\mathbf{x}_n \mid m_1,\dots,m_T,t_1,\dots,t_{T-1}) \\ &\qquad = \sum_{j=1}^{T}\sum_{i=t_{j-1}}^{t_j} \log p(\Delta\mathbf{x}_i \mid m_j) = \sum_{j=1}^{T}\sum_{i=t_{j-1}}^{t_j} \log \mathcal{N}(\Delta\mathbf{x}_i \mid \mu_{m_j}, Q_{m_j}) \end{aligned} \tag{4}$$

where we define $t_0 = 1$, T is the number of segments and t_j the switch time. Assuming that T is known, we can “segment” the sequence (i.e., estimate $m_1,\dots,m_T$ and $t_1,\dots,t_{T-1}$) using the maximum-likelihood approach:

$$\hat{m}_1,\dots,\hat{m}_T,\hat{t}_1,\dots,\hat{t}_{T-1} = \arg\max L(\Delta\mathbf{x}_1,\dots,\Delta\mathbf{x}_n \mid m_1,\dots,m_T,t_1,\dots,t_{T-1}) \tag{5}$$

This maximization can be performed in a nested way,

$$\hat{t}_1,\dots,\hat{t}_{T-1} = \arg\max_{t_1,\dots,t_{T-1}} \left\{\max_{m_1,\dots,m_T} L(\Delta\mathbf{x}_1,\dots,\Delta\mathbf{x}_n \mid m_1,\dots,m_T,t_1,\dots,t_{T-1})\right\} \tag{6}$$

In fact, the inner maximization can be decoupled as

$$\max_{m_1,\dots,m_T} L(\Delta\mathbf{x}_1,\dots,\Delta\mathbf{x}_n \mid m_1,\dots,m_T,t_1,\dots,t_{T-1}) = \sum_{j=1}^{T}\max_{m_j}\sum_{i=t_{j-1}}^{t_j} \log p(\Delta\mathbf{x}_i \mid m_j) \tag{7}$$

where the maximization with respect to each of m_j is a simple maximum likelihood classifier of sub-set of samples $(\Delta\mathbf{x}_{t_{j-1}},\dots,\Delta\mathbf{x}_{t_j})$ into one of a set of Gaussian classes. Finally, the maximization with respect to $t_1,\dots,t_{T-1}$ is done by exhaustive search (this is never too expensive, since we consider a maximum of three segments).

4.3 Estimating the number of models of the activity

4.3.1 MDL Criterion

In the previous section, we derived the segmentation criterion assuming that the number of segments T is known. As is well known, the same criterion can not be used to select T, as this would always return the largest possible number of segments. We are thus in the presence of a model selection problem, which we address by using the minimum description length (MDL) criterion [14]. The MDL criterion for selecting T is

$$\hat{T} = \arg\min_{T} \Big\{ -\log p(\Delta\mathbf{x}_1, \ldots, \Delta\mathbf{x}_n \mid \hat{m}_1, \ldots, \hat{m}_T, \hat{t}_1, \ldots, \hat{t}_{T-1}) + M(\hat{m}_1, \ldots, \hat{m}_T, \hat{t}_1, \ldots, \hat{t}_{T-1}) \Big\} \tag{8}$$

where $M(\hat{m}_1, \ldots, \hat{m}_T, \hat{t}_1, \ldots, \hat{t}_{T-1})$ is the number of bits required to encode the selected model indeces and the estimated switching times. Notice that we do not have the usual $\frac{1}{2}\log n$ term because the real-valued model parameters (means and covariances) are assumed fixed (previously estimated). Finally, it is easy to conclude that

$$M(\hat{m}_1, \ldots, \hat{m}_T, \hat{t}_1, \ldots, \hat{t}_{T-1}) \approx T \log c + (T-1)\log n \tag{9}$$

where $T \log c$ is the code length for the model indeces $m_1, \ldots, m_T$, since each belongs to $\{1, \ldots, c\}$, and $(T-1)\log n$ is the code length for $\hat{t}_1, \ldots, \hat{t}_{T-1}$, because each belongs to $\{1, \ldots, n\}$; we have ignored the fact that two switchings can not occur at the same time, because $T << n$.

5 Experimental results

This section presents results with synthetic and real data. In the synthetic case, we have performed Monte Carlo tests. We have considered five models ($c = 5$) shown in Fig. 3. The synthetic models shown in Fig. 3(a) were obtained by simulating four activities of a person, using the generation model in (1). Fig. 4 shows examples of activities (the trajectory shape of "Leaving" is the same as "Entering", however with opposite direction). Here, the thin (green) rectangles correspond to areas where the trajectory begins. The first sample of $\mathbf{x}_t$ in these areas is random, because the agent may appears at random places in the scene. The wide (yellow) rectangle is the area in which occurs a model switching. In this figure the trajectories are generated with two segments ("Entering", "Leaving", "Passing") and with three segments ("Browsing").

For each activity we generate 100 test samples using (1) and classify each of them in one of the four classes. Fig. 5 shows the displacements $\Delta\mathbf{x}_t$ (black dots) of the test sequences ("Entering" and "Passing") overlapped with the five models. We can see that the displacements lie on *right-up* clusters ("Entering") and *right* cluster ("Passing"). In this experiment, all the test sequences were correctly classified (%100 accuracy).

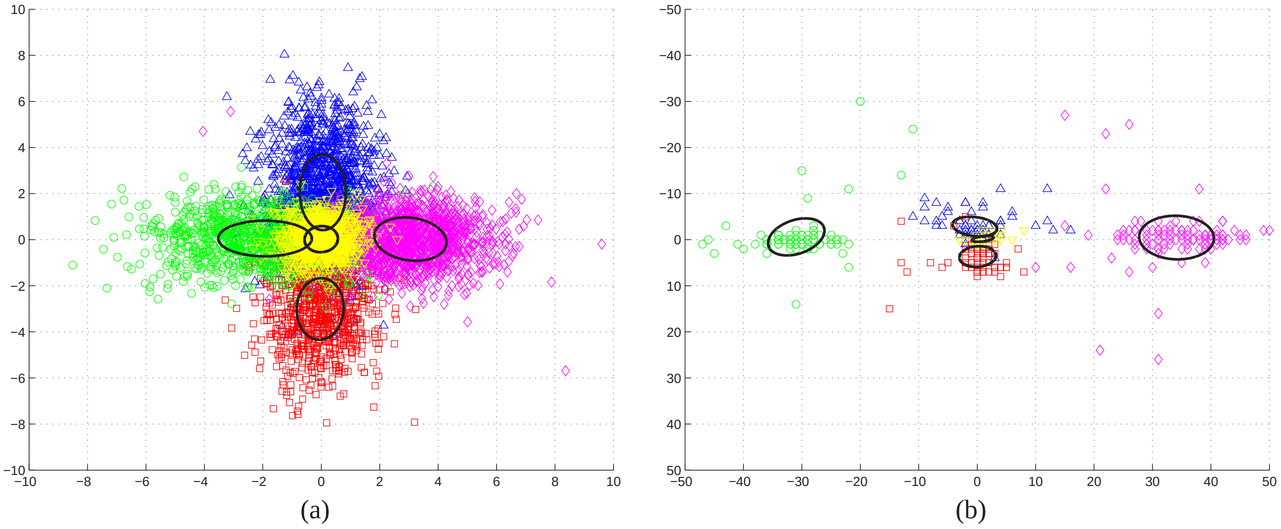


Figure 3: Five models are considered to describe trajectory. Each color corresponds to a different model. Synthetic case (a), real case (b).

We also generated different test trajectories, this is because the exiting and entering may occur in different direction from the ones in Fig. 4. These examples are illustrated in Fig. 6. In this new experiment, the same 100% accuracy was also obtained.

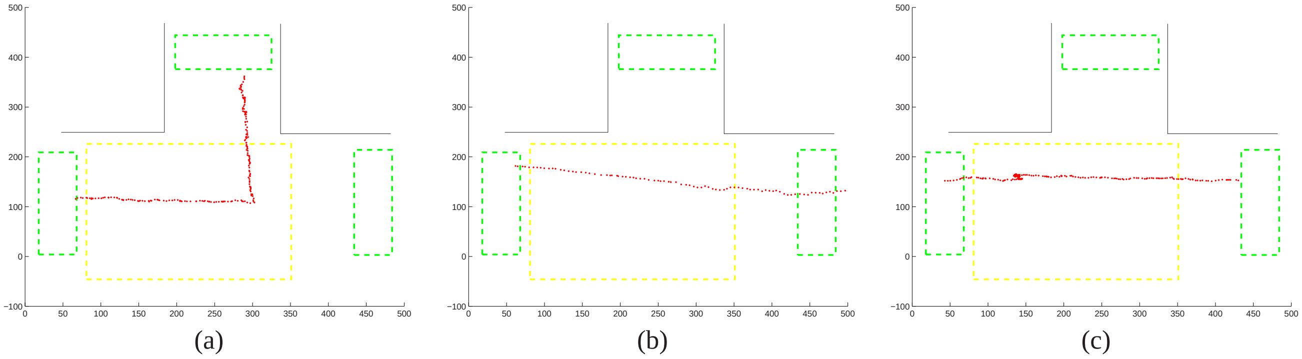


Figure 4: Examples of synthetic activities (performed in left-right direction): (a) entering, (b) passing, (c) browsing.

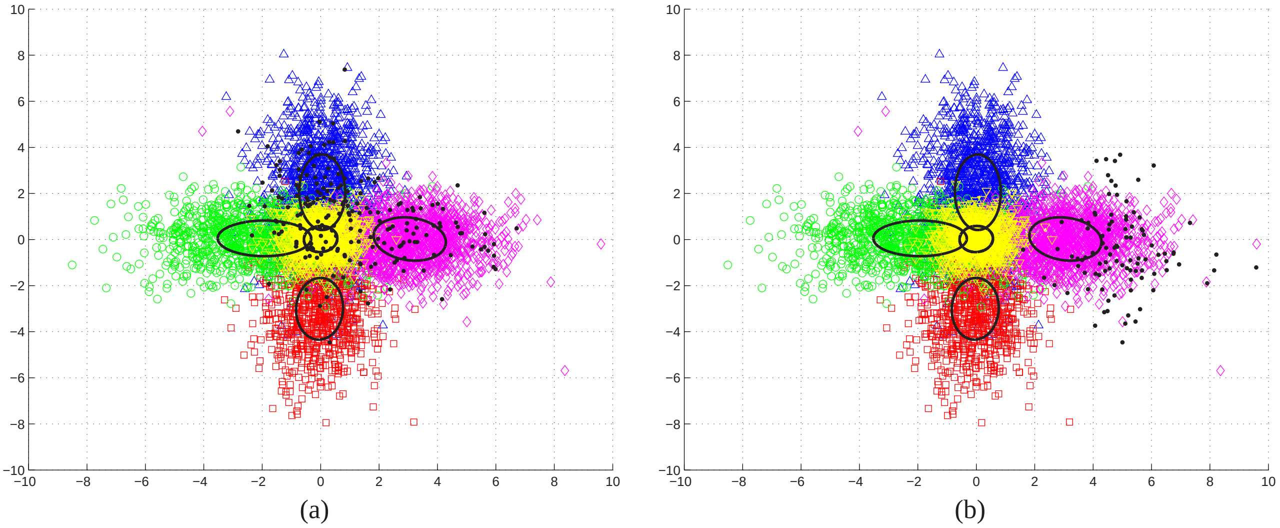


Figure 5: Five models with the displacements (black dots) of the test activities: (a) entering, (b) passing.

The proposed algorithm was also tested with real data. The video sequences were acquired in the context of the EC funded project CAVIAR. All the video sequences comprise human activities in indoor plaza and shopping center observations of individuals and small groups of people. Ground truth was hand-labelled for all sequences[2]. Fig. 7 shows the bounding boxes as well as the centroid, which is the information used for the segmentation.

As in the synthetic case, we also generate the statistics of the considered models. The procedure is the same as in the previous case using training sequences. Fig. 3(b) shows the clusters of the models.

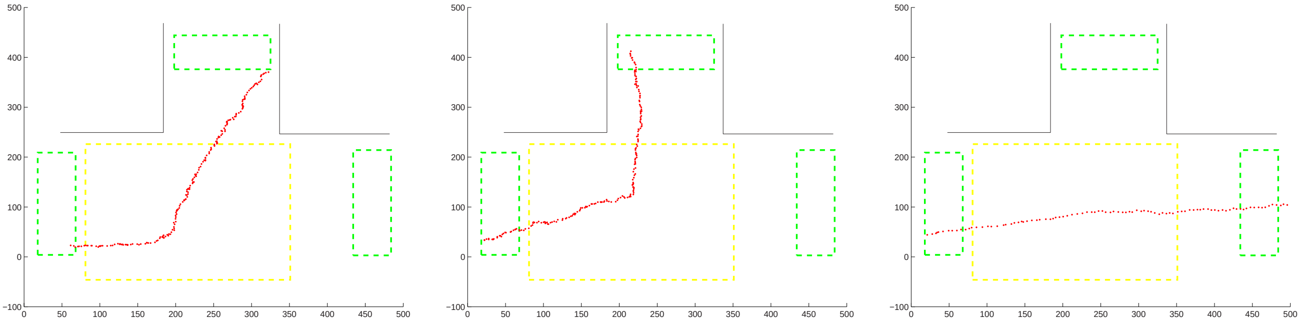
Fig. 8 shows several activities performed at the shopping center with the time instants of the model switching marked with small red circle. From this experiment, it can be seen that the proposed approach correctly determines the switching times between models.

We have tested the proposed approach in more than 40 trajectories from 25 movies of about 5 minutes each. We just present the results of some of those activities in Tables 1 and 2. These Tables show the penalized log-likelihood values (8) of each test sequence. The first table refers to all activities performed in the left-right direction, whilst the second table reports all activities performed in the opposite direction. In the first table the classes referring to entering, exiting, passing and browsing are *right-upwards*, *downwards-right*, *right*, *right-stop-right* respectively, whereas in the second table the classes are *left-upwards*, *downwards-left*, *left* and *left-stop-left*. It can be observed that the output classifier correctly assigns the activities into the corresponding classes, exhibiting good results as in the previous synthetic examples.

6 Conclusions

In this paper we have proposed and tested an algorithm for modelling, segmentation, and classification of human activities in a constrained environment. The proposed approach uses a switched dynamical models to represent the human trajectories. It was illustrated that the time instants are effectively well determined, despite of the significant random perturbations that the trajectory may contain. It is demonstrated that the proposed approach provides good

[2] The ground truth labelled video sequences is provided at `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`.



(a)

(b)

(c)

Figure 6: Synthetic activities with different dynamic models (entering,exiting,passing).

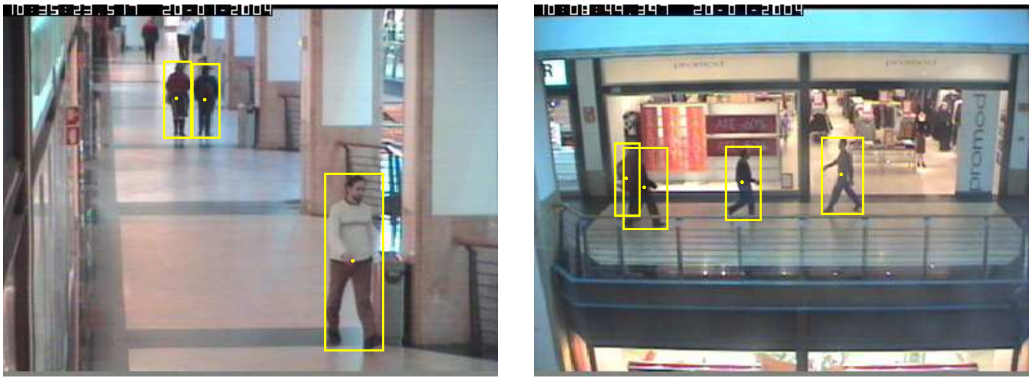


Figure 7: Bounding boxes and centroids of the pedestrians performing activities.

results with synthetic and real data obtained in a shopping center. The proposed method is able to effectively recognize instances of the learned activities. The activities studied herein can be interpreted as atomic, in the sense that they are simple events. Compound actions or complex events can be represented as concatenations of the activities studied in this paper. This is one of the issues to be addressed in the future.

Acknowledgement: We would like to thank Prof. José Santos Victor of ISR and the members of CAVIAR project, for providing video data of human activities with the ground truth information.

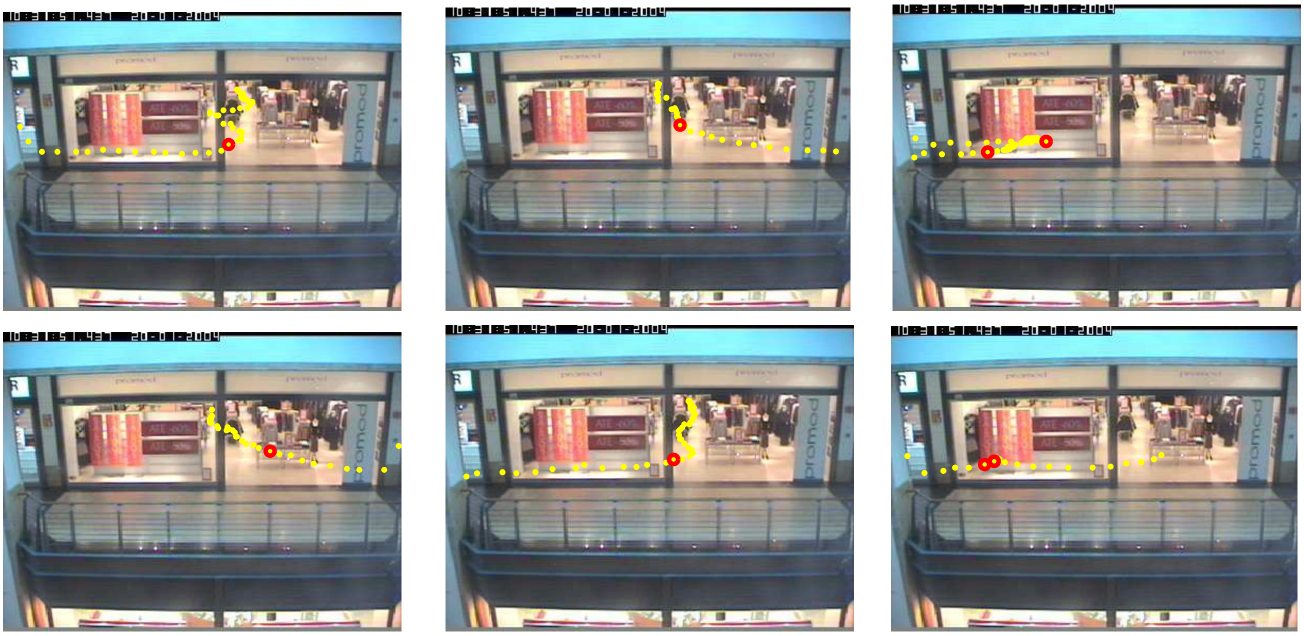


Figure 8: Samples of different activities. The large circles are the computed times instants where the model switches: Entering (first column); exiting (second column); browsing (third column).

	Test trajectories						
Classes	E_1	E_2	Ex_1	Ex_2	P_1	P_2	B
Entering	**187.2**	**157.3**	212.7	217.0	100.3	107.4	169.1
Exiting	401.0	340.0	**116.1**	**102.4**	104.6	93.8	178.7
Passing	359.7	311.0	232.5	183.3	**88.8**	**90.2**	147.7
Browsing	299.1	265.6	196.5	180.0	160.7	156.0	**98.1**

Table 1: Penalized Log-likelihood of several real activities performed in left-right direction: *E*- entering, *Ex*-exiting, *P*- passing, *B*- browsing.

	Test trajectories						
Classes	E_1	E_2	Ex_1	Ex_2	P_1	P_2	B
Entering	**116.2**	**115.0**	337.7	358.2	89.3	90.9	211.7
Exiting	277.6	284.6	**151.0**	**127.4**	98.6	96.6	297.4
Passing	210.0	224.4	350.1	362.0	**63.4**	**64.7**	358.4
Browsing	207.4	197.3	343.2	286.7	188.9	179.0	**170.1**

Table 2: Penalized Log-likelihood of several real activities performed in right-left direction: *E*- entering, *Ex*- exiting, *P*- passing, *B*- browsing.

References

[1] D. Ayers and M. Shah,“Monitoring Human Behavior from Video Taken in an Office Environment”, *Image and Vision Computing*, vol. 19, Issue 12, 1, pp. 833-846, Oct, 2001.

[2] A. Bobick and J. Davis, “The Recognition of Human Movement using Temporal Templates”, in *IEEE Transaction on Pattern Analysis and Machine Intelligence*, pp. 257-267, vol. 23, no. 3, March 2001.

[3] J. Davis and M. Shah, “Visual Gesture Recognition”, *IEE Proc. Vision, Image and Signal Processing*, Vol. 141, No. 2, pp. 101-106, April 1994.

[4] S. Hongeng and R. Nevatia, “Multi-Agent Event Recognition”, in *Proc. of the 8 th IEEE Int. Conf. on Computer Vision (ICCV’01)*, pp. 84-91, vol. 2, 2001.

[5] M. Isard and A. Blake,“A Mixed-state Condensation Tracker with Automatic Model-switching”, *Proc. of the Int. Conf. on Computer Vision*, pp. 107-112, 1998.

[6] J. S. Marques and J. M. Lemos, “Optimal and Suboptimal Shape Tracking Based on Switched Dynamic Models”, *Image and Vision Computing*, pp. 539-550, june, 2001.

[7] N. Johnson and D. Hogg, “Representation and Synthesis of Behaviour using Gaussian Mixtures”, in *Image and Vision Computing*, pp. 889-894, vol. 20, no 12, 2002.

[8] A. J. Abrantes, J. S. Marques, J. M. Lemos, “Long Term Tracking Using Bayesian Networks”, in *Proc. of IEEE Int. Conf. on Image Processing*, Rochester, 609-612, vol. III, Sept. 2002.

[9] O. Masoud and N.P. Papanikolopoulos, “A Method for Human Action Recognition”, in *Image and Vision Computing*, pp.729-743, vol. 21, no. 8, August 2003.

[10] A. Nagai, Y. Kuno and Y. Suirai, “Surveillance Systems based on Spatio-temporal Information”, *Proc. IEEE Int. Conf. Image Processing*, pp. 593-596, 1996.

[11] J. C. Nascimento and M. A. T. Figueiredo and J. S. Marques, “Recognition of Human Activities with Space Dependent Switched Dynamical Models”, *Proc. IEEE Int. Conf. Image Processing*, September, 2005.

[12] N. M. Oliver and B. Rosario and A. P. Pentland, “A Bayesian Computer Vision System for Modeling Human Interactions”, in *IEEE Trans. on Pattern Anal. and Machine Intell.*, pp. 831-843, vol. 22, no. 8, August 2000.

[13] T. J. Olson and F. Z. Brill, “Moving Object Detection and Event Recognition for smart Cameras”, *Proc. Image Understanding Workshop*, pp. 159-175, 1997.

[14] J. Rissanen, “Stochastic Complexity in Statistical Inquiry.”Singapore: World Scientific, 1989.

[15] M. Rosenblum and Y. Yacoob and L. S. Davis, “Human expression recognition from motion using a radial basis function network architecture”, *IEEE Trans. Neural Networks*, no. 7, pp. 1121-1138, 1996.

[16] Y. Yacoob and M. J. Black, “Parameterized Modeling and Recognition of Activities”, in *Computer Vision and Image Understanding*, pp. 232-247, vol. 73, no. 2, February 1999.