On-line Classification of Human Activities *

J. C. Nascimento
jan@isr.ist.utl.pt
ISR-IST

M. A. T. Figueiredo
mario.figueiredo@lx.it.pt
IT-IST

J. S. Marques
jsm@isr.ist.utl.pt
ISR-IST

Abstract. In this paper we address the problem of on-line recognition of human activities taking place in a public area such as a shopping center. We consider standard activities; namely, *entering*, *exiting*, *passing* or *browsing*. The problem is motivated by surveillance applications, for which large numbers of cameras have been deployed in recent years. Such systems should be able to detect and recognize human activities, with as little human intervention as possible.

In this work, we model the displacement of a person in consecutive frames using a bank of switched dynamical systems, each of which tailored to the specific motion regimes that each trajectory may contain.

Our experimental results are based on nearly 20,000 images concerning four atomic activities and several complex ones, and demonstrate the effectiveness of the proposed approach.

1 Introduction and Problem Formulation

In this paper, we address the problem of (on-line) recognition of human activities in video sequences. Recently, this has become an active research area in computer vision, mainly driven by a large number of potential applications, such as video surveillance, computer-human interfaces, and contend-based video retrieval.

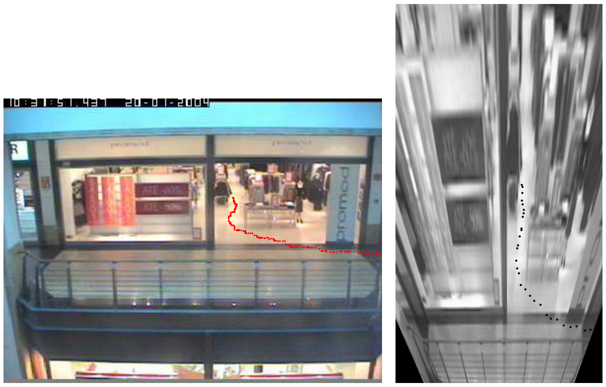
In a surveillance context, the analysis of the human behavior is often split into two parts: tracking and activity recognition [8]. Considering that tracking has seen tremendous recent progress [2,3,5,6,11,14,16], activity recognition has naturally become the next step to be addressed.

Different methods have been used to recognize human activities from the information extracted from video. The most popular techniques rely on *hidden Markov models* (HMM) and *coupled HMM* [12]. Both approaches are used to characterize the evolution of the person's mass center along the video sequence. A model termed *abstract HMM* was used to recognize human indoor motion patterns [10]. Other types of techniques have also been successfully used for gesture and activity recognition; e.g., Bayesian networks [7], neural networks [15], finite state machines (FSM) [1,4] and syntactic recognition [9].

In this work, we consider that a tracking system computes the active region (bounding box) of the person along the video sequence. We also assume that the

* This work was supported by Fundação para a Ciência e a Tecnologia (ISR/IST pluri-anual funding) through the POS Conhecimento Program which includes FEDER funds.

measurements provided by the tracker are corrected using the image to ground plane projective transformation, thus achieving viewpoint invariance and removing perspective distortion. Fig. 1 shows an example of an observed trajectory, before and after the projective transformation.



(a) (b)

Fig. 1. Original (left) and resulting transformed (right) images of the shopping center scenario.

Our fundamental assumption is that the human (motion) activity can be inferred from the sequence of positions of the centroid of the person throughout the video sequence, which is provided by the tracker. After the projective transformation is applied, this sequence is denoted $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, where $\mathbf{x}_t \in \mathbb{R}^2$, for $t = 1, \ldots, n$, is the position at time instant t.

Our approach categorizes human activities using a two-level hierarchical system. At the lower-level, we have *dynamic models*, which are short term coherent units of movement; at the higher level, we consider *activities*, which are linearly ordered sequences of lower level *dynamic models*. In this paper, we consider five low level *dynamic models*: “moving left”, “moving right”, “moving up”, “moving down”, “stopped”. Four activities are considered: “passing”, “entering” “leaving” and “browsing”. Of course, this hierarchy could be extended to more complex arrangements of activities, but this will not be pursued in this paper.

Finally, our problem can be formulated as follows: *given a trajectory* $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$*, observed in a length-n time window, segment it into a sequence of low level dynamic models and classify it into one of the high level activities.*

The paper is organized as follows. Section 2 describes the adopted low level model and the parameter estimation method. Section 3 addresses the segmentation criterion. Section 4 describes the high level classification of sequences. Section 5 describes experimental results and Section 6 concludes the paper.

2 Statistical Model and Parameter Estimation

A trajectory is a sequence of positions, $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ with $\mathbf{x}_i \in \mathbb{R}^2$. This sequence is modeled by a switched dynamical system, which is allowed to switch among the 5 low level models above defined. Formally, the state equation is

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \boldsymbol{\mu}_{k_t} + \mathbf{Q}_{k_t}^{1/2} \mathbf{w}_t, \tag{1}$$

where $\{k_1, \ldots, k_n\}$, with $k_t \in \{1, \ldots, 5\}$, is a sequence of labels indicating the active low level dynamic model at each 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 $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_5\}$, the mean displacements of each model, and $\{\mathbf{Q}_1, \ldots, \mathbf{Q}_5\}$, the corresponding covariances.

The joint probability density of a sequence $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, generated according to (1), given a sequence of model labels $\{k_1, \ldots, k_n\}$ is thus

$$p(\mathbf{x}_1, ..., \mathbf{x}_n | k_1, \ldots, k_n) = \prod_{t=2}^{n} \mathcal{N}(\mathbf{x}_t - \mathbf{x}_{t-1} | \boldsymbol{\mu}_{k_t}, \mathbf{Q}_{k_t}), \tag{2}$$

where $\mathcal{N}(\mathbf{v}|\mathbf{u}, \mathbf{P})$ denotes a multivariate Gaussian density of mean $\mathbf{u}$ and covariance $\mathbf{P}$, computed at $\mathbf{v}$.

Estimation of the parameters of each of the low level models is performed in a supervised fashion using training trajectories which were previously segmented and classified by a human observer. These parameters are set to the standard maximum likelihood estimates, given the training data.

3 Segmentation and Classification

3.1 Segmentation with a known number of segments

For segmentation purposes, we assume that the sequence of labels $\{k_1, \ldots, k_n\}$ is piece-wise constant, with T segments, that is,

$$\{k_1, \ldots, k_n\} = \{m_1, \ldots, m_1, m_2, \ldots, m_2, \ldots, m_T, \ldots, m_T\}. \tag{3}$$

Let us denote as $\{s_1, ..., s_{T-1}\}$ the switching times between segments, where s_j is the time instant where switching from model m_j and m_{j+1} occurs. Obviously, the sequence of models $\{m_1, \ldots, m_T\}$ and switching times $\{s_1, ..., s_{T-1}\}$ contains exactly the same information as the sequence of labels $\{k_1, \ldots, k_n\}$. This allows writing the segmentation log-likelihood, which is simply the logarithm of (2), as

$$\begin{aligned} L(m_1, \ldots, m_T, s_1, \ldots, s_{T-1}) &= \log p(\mathbf{x}_1, ..., \mathbf{x}_n | m_1, \ldots, m_T, s_1, \ldots, s_{T-1}) \\ &= \sum_{j=1}^{T} \sum_{t=s_{j-1}}^{s_j} \log \mathcal{N}(\mathbf{x}_t - \mathbf{x}_{t-1} | \boldsymbol{\mu}_{m_j}, \mathbf{Q}_{m_j}) \end{aligned} \tag{4}$$

where we take $s_0 = 1$.

Assuming that T is known, we can "segment" the sequence (i.e., estimate $\{m_1,\ldots,m_T\}$ and $\{s_1,\ldots,s_{T-1}\}$) by the maximum-likelihood criterion:

$$\{\hat{m}_1,\ldots,\hat{m}_T,\hat{s}_1,\ldots,\hat{s}_{T-1}\} = \arg\max L(m_1,\ldots,m_T,s_1,\ldots,s_{T-1}) \tag{5}$$

The maximization with respect to the switching times can be expressed as

$$\hat{s}_1,\ldots,\hat{s}_{T-1} = \arg\max_{s_1,\ldots,s_{T-1}} \Big\{ \max_{m_1,\ldots,m_T} L(m_1,\ldots,m_T,s_1,\ldots,s_{T-1}) \Big\}. \tag{6}$$

The inner maximization in (6), that is, with respect to $\{m_1,\ldots,m_T\}$, for some fixed $\{s_1,\ldots,s_{T-1}\}$, can be decoupled into

$$\max_{m_1,\ldots,m_T} L(m_1,\ldots,m_T,s_1,\ldots,s_{T-1}) = \sum_{j=1}^{T} \max_{m_j} \sum_{t=s_{j-1}}^{s_j} \log\mathcal{N}(\mathbf{x}_t - \mathbf{x}_{t-1} | \boldsymbol{\mu}_{m_j}, \mathbf{Q}_{m_j}). \tag{7}$$

Notice that the maximization with respect to each of m_j is a simple maximum likelihood classifier of the sub-sequence $(\mathbf{x}_{s_{j-1}-1},\ldots,\mathbf{x}_{s_j})$ into one of the 5 low level models. Finally, the maximization with respect to $s_1,\ldots,s_{T-1}$ is done by exhaustive search, which is never too expensive, since we are considering short segments of the trajectory, with up to a maximum of $T=3$ segments.

3.2 Estimating the number of segments: MDL Criterion

In the previous section, we derived the segmentation criterion assuming that the number of segments T is known. It is well known that 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 [13]. The MDL criterion for selecting T is

$$\begin{aligned}\hat{T} = \arg\min_T \Big\{ & -\log p(\mathbf{x}_1,\ldots,\mathbf{x}_n | \hat{m}_1,\ldots,\hat{m}_T,\hat{s}_1,\ldots,\hat{s}_{T-1}) \\ & + M(\hat{m}_1,\ldots,\hat{m}_T,\hat{s}_1,\ldots,\hat{s}_{T-1}) \Big\}\end{aligned} \tag{8}$$

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

$$M(\hat{m}_1,\ldots,\hat{m}_T,\hat{s}_1,\ldots,\hat{s}_{T-1}) \approx T\log_2 5 + (T-1)\log_2 n \tag{9}$$

where $T\log_2 5$ is the code length for the T model labels $m_1,\ldots,m_T$, since each belongs to $\{1,\ldots,5\}$, and $(T-1)\log_2 n$ is the code length for the $T-1$ switching times, $\hat{s}_1,\ldots,\hat{s}_{T-1}$, because each belongs to $\{1,\ldots,n\}$; we ignore the fact that two switchings can not occur at the same time, a reasonable approximation because $T << n$. The maximization in (8) is solved simply by trying all allowed numbers of segments (1, 2, or 3, in all the experiments below).

In a classical MDL-based segmentation method, we would simply estimate the segment parameters along with the segmentation, and use the MDL criterion in the standard way to select the number of segments. However, without the supervised training scheme, we wouldn't be able to assign a semantic to each model, e.g., *"moving right"*, *"moving left"*. In short, supervised training is needed when, in addition to segmenting, one wishes to classify and interpret activities. This leads to the use of the MDL with fixed parameters, as propose in this work.

4 On-Line Identification of the Sequence

To identify the (high level) activity present in a given sequence of positions, each possible sequence of 1, 2, or 3, low-level models (produced by the segmentation algorithm) is mapped to an activity according to a simple look-up table (which we omit due to lack of space). For example, a sequence segmented into only one segment ($\hat{T} = 1$) with model "walking right" or "walking left" is classified as "passing"; a sequence segmented into two segments ($\hat{T} = 2$), as "walking right" - "walking up", or "walking left" - "walking up", is classified as "entering"; a sequence segmented into three segments ($\hat{T} = 3$), as "walking right" - "stopped" - "walking left" is classified as "browsing". For the 5 considered models, there are 5 possible 1-segment segmentations, $5 \times 4 = 20$ possible 2-segment segmentations, and $5 \times 4 \times 4 = 80$ possible 3-segment segmentations, thus our look-up table has a total of 105 entries.

To perform on-line classification, the segmentation/classification algorithm is not applied to the whole observed trajectory of a given person, but to the positions inside a fixed length sliding window. For each window position, the segmentation and classification algorithm is applied and the system outputs a high-level activity class.

5 Experimental results

The proposed algorithm was tested with real data collected in the context of a EU-funded project [1]. This section shows the performance of the proposed algorithm applied to about 20 movies. The duration of the movies ranges from 30 seconds up to two minutes, with a frame rate of 25 frames/second. The data was collected and the ground truth was hand-labeled for 50 video sequences. These sequences include indoor plaza and shopping center observations of individuals and small groups of people. The sequences are hand labeled with the activity of each track person, frame by frame. The total data consists of nearly 20,000 images.

[1] More information about the CAVIAR project can be found at `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`.

Table 1 shows the confusion matrix for the tested activities. The evaluation is made at every window position. The samples of the trajectories varies from 300 (shorter sequences) to 1200 (longer sequences).

		Output			
		Entering	*Exiting*	*Passing*	*Browsing*
True Classes	*Entering*	81	0	0	0
	Exiting	0	72	5	0
	Passing	1	0	462	3
	Browsing	1	0	1	180

Table 1. Confusion matrix for the classification of high level activities in the shopping.

Figures 2 and 3 show some results obtained using the proposed approach. Each image shows the successive positions of the person up to the time instant in which the classification is being output, the current bounding box. For the sake of clarity, these figures only show the activity class for a single person.

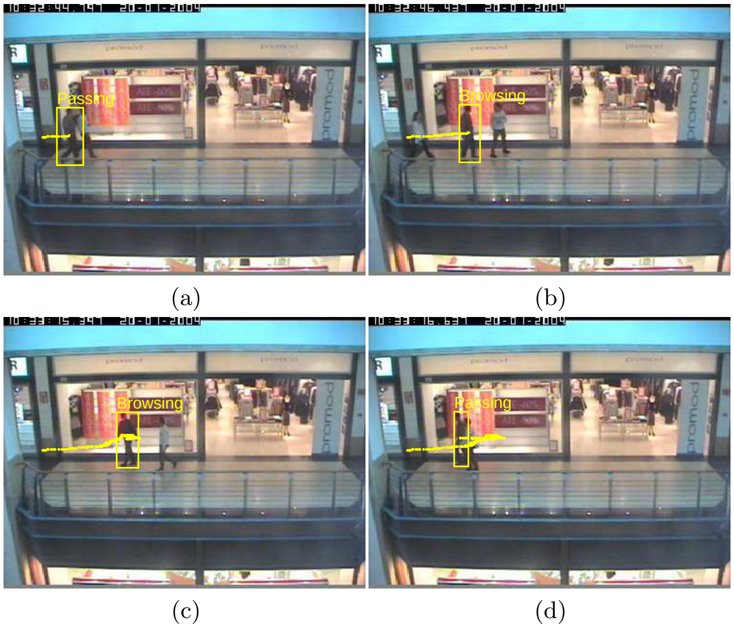
6 Conclusions

In this paper we have proposed and tested an algorithm for online segmentation and classification human motion activities. These activities are classified using a two-level hierarchical system. At a lower-level, we have *dynamic models*, which are short units of movement (such as "moving right") and at the higher level, we have the target *activities*, which are sequences of lower level models. We introduce probabilistic generative low-level models and a minimum description length criterion to segment each observed trajectory into a sequence of low-level models. Each possible sequence of low-level models is then translated into a high-level activity, via a look-up table. We have reported extensive experiments, which testify for the good performance of the proposed methodology.

We plan to extend our work to higher semantic levels. For instance, in the example shown in the Fig. 2, we may hope to infer that the person is waiting for another person, while that other person goes to the shop. In the future, we hope to bring this higher level descriptions (such as "waiting") by considering interactions among the trajectories of different people.

References

1. D. Ayers and M. Shah, "Monitoring human behavior from video taken in an office environment," *Image and Vision Computing*, vol. 19, pp. 833–846, 2001.

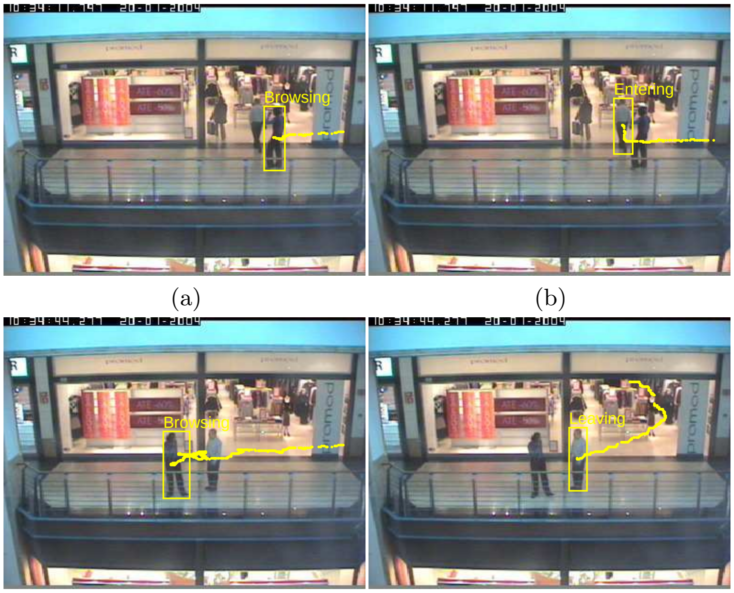


(a) (b)

(c) (d)

Fig. 2. A person "passing" in a front of a shop (a) (frame 21), starts to "browse" while the second one "enters" to the shop (b) (frame 62). He waits ("browsing") until his colleague leaves the shop (c) (frame 786). Finally they go together leaving the scenario ("passing") (frame 817).

2. A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 23, pp. 257–267, 2001.
3. J. Davis, "Sequential reliable inference for rapid detection of human actions," *IEEE Conf. on Advance Video and Signal Based Surveillance*, pp. 169–176, 2003.
4. J. Davis and M. Shah, "Visual gesture recognition," *IEE Proc. Vision, Image and Signal Processing*, vol. 141, pp. 101–106, 1994.
5. A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *IEEE Int. Conf. on Comp. Vision*, pp. 726–733, Nice, France, 2003.
6. I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: real-time surveillance of people and their activities," *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 22, pp. 809–830, 2000.
7. S. Hongeng, R. Nevatia, and F. Bremond, "Video-based event recognition: activity representation and probabilistic recognition methods," *Computer Vision and Image Understanding*, vol. 96, pp. 129–162, 2004.
8. T. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. on Systems and Cybernetics: Applications and Reviews*, vol. 34, pp. 334–352, 2004.
9. Y. Ivanov and A. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 22, pp. 852–872, 2000.



(a) (b)

(c) (d)

Fig. 3. The first row shows a person who starts to “browse” (a) while the second one “enters” to the shop (b) (frame 157). In the bottom row, the first person waits for his colleague (performing a “browsing”) until the second one “leaves” the shop (frame 969).

10. L. Liao, D. Fox, and H. Kautz, “Learning and inferring transportation routines,” in *Proc. National Conf. on Artificial Intelligence*, 2004.
11. K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe, “A bosted particle filter: Multitarget detection and tracking,” *Europ. Conf. Comp. Vision*, LNCS vol. 3021, pp. 28–39, Springer Verlag, 2004.
12. N. Oliver, B. Rosario, and A. Pentland, “A Bayesian computer vision system for modeling human interactions,” *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 22, pp. 831–843, 2000.
13. J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.
14. R. Rosales and S. Sclaroff, “3D trajectory recovery for tracking multiple objects and trajectory guided recognition of actions,” *Proc. IEEE Conf. on Comp. Vision and Patt. Recognition*, vol. 2, pp. 2117–2123, 1999.
15. M. Rosenblum, Y. Yacoob, and L. S. Davis, “Human expression recognition from motion using a radial basis function network architecture,” *IEEE Trans. Neural Networks*, vol. 7, pp. 1121–1138, 1996.
16. T. Zhao and R. Nevatia, “Tracking multiple humans in complex situations,” *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 26, pp. 1208– 1221, 2004.