SEMI-SUPERVISED LEARNING OF SWITCHED DYNAMICAL MODELS FOR CLASSIFICATION OF HUMAN ACTIVITIES IN SURVEILLANCE APPLICATIONS

Jacinto C. Nascimento [(a)] *Mário A. T. Figueiredo* [(b)] *Jorge S. Marques* [(a)]

[(a)]Instituto de Sistemas e Robótica [(b)]Instituto de Telecomunicações
Instituto Superior Técnico
1049-001 Lisboa,
Portugal

ABSTRACT

This work introduces a semi-supervised approach for learning generative models for classification/recognition of human trajectories, with application to surveillance. The classifier is based on switched dynamical models, with each model describing a specific motion regime. We present a semi-supervised modified version of the classical Baum-Welch algorithm, which is able to take into account a subset of known model labels. The experimental results reported, using both synthetic and real data, show that the classifier learned with semi-supervision leads to a higher classification accuracy than the fully unsupervised version, thus validating the proposed approach.

Keywords: hidden Markov models, switched dynamical models, EM algorithm, semi-supervised learning, surveillance.

1. INTRODUCTION

One of the main tasks of video surveillance systems is to recognize and monitor human activities. These tasks are difficult, mainly due to the complexity of the scene and the human actions. To tackle the uncertainty inherent to the data, current approaches to activity recognition typically use probabilistic temporal models, such *hidden Markov models* (HMM) and *conditional hidden Markov models* (CHMM) [1], modeling single-person or person-to-person interactions. Alternative methods have been recently proposed, including the *abstract HMM* (AHMM) [2] and the *hierarchical HMM* (HHMM) [3], which models the high-level behavior of persons in indoor environments, using images from multiple cameras. These approaches are generative, since the relationship between the activity and the observations is modeled via a joint probability function.

Discriminative techniques have also been considered. Two recent examples are *conditional random fields* (CRF) and *maximum entropy Markov models* (MEMM); for details and performance comparison in the context of video surveillance see [4].

The main feature of this paper is the use a bank of switched dynamical systems to describe the trajectory of a pedestrian in a video sequence. We have previously used this type of models for activity recognition in [5]. The novelty in this paper lies on the use of a semi-supervised learning framework, in which some of the model labels are observed. More specifically, we show how the classical Baum-Welch (BW), that is, the expectation-maximization (EM), algorithm for learning the parameters of a switched dynamical model can be modified to incorporate the observation of some labels.

(a) (b)

Fig. 1. (a) Examples of activities (Browsing, leaving) unrolled in the context of our application; (b) sensors located in the scenario.

The context of application of the present work is the recognition of typical human activities in a shopping center. The activity classes are "passing", "entering" (a shop), "exiting" (a shop), and "browsing"; see Fig. 1(a) for an illustration. The trajectories in each activity class are described by a switched dynamical model of several motion regimes (such as "moving left", "stopped"). When only trajectories are observed, the parameters of these switched dynamical models are estimated by an EM algorithm, which in this case coincides with the BW algorithm. When some model labels are observed (*e.g.*, obtained manually) the EM/BW algorithm has to be modified; this modification is the main topic of this paper.

The remainder of this paper is organized as follows. Section 2 describes the adopted model. Sections 3 and 4 present the proposed approach based on semi-supervised learning. Experiments are reported and discussed in Section 5. Finally, Section 6 presents some concluding remarks.

2. THE MODEL

It is assumed that the human motion activities of interest are represented by the trajectory of a person mass center in the video sequence. The evolution of the mass center is modeled by a bank of switched dynamical models.

Let $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_N)$ be the sequence of positions of the mass center, $\mathbf{x}_i \in \mathbb{R}^2$. The switched dynamical system considered is

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{T}_{k_t} + w_t, \quad (1)$$

where $k_t \in \{1, \ldots, m\}$ is the label of the active model at time instance t; $\mathbf{T}_{k_t}$ is the mean displacement which depends on the active model; and $w_t \sim \mathcal{N}(0, \mathbf{R}_{k_t})$ is a white Gaussian noise with zero mean and covariance matrix $\mathbf{R}_{k_t}$.

We assume that the sequence of model labels $\mathbf{k} = (k_1, \ldots, k_N)$ is a sample of a Markov chain, with $(m \times m)$ transition matrix $\mathbf{B} = [B(i,j)]$ and initial distribution $\boldsymbol{\pi}$. The sequence $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is observed and $\mathbf{k} = (k_1, \ldots, k_N)$ is partially hidden: we observe the labels at some regions of the image. See Fig. 1(b) where the small areas correspond to "sensors" where we know the model labels. For each activity, the model parameters are those of an classical HMM: $\boldsymbol{\theta} = (\mathbf{T}_1, \mathbf{R}_1, \ldots, \mathbf{T}_m, \mathbf{R}_m, \mathbf{B}, \boldsymbol{\pi})$.

We now describe the main variables of the problem, assuming that we have M trajectories. Let $\mathbf{x}^{(s)} = (\mathbf{x}_1^{(s)}, \ldots, \mathbf{x}_{N_s}^{(s)})$, for $s = 1, ..., M$, denote the s-th observed sequence of positions (trajectory) and $\mathcal{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}\}$ the set of observed sequences. Let $\mathbf{k}^{(s)} = (k_1^{(s)}, \ldots, k_{N_s}^{(s)})$ be the label sequence of the s-th trajectory and $\mathcal{K} = \{\mathbf{k}^{(1)}, \ldots, \mathbf{k}^{(M)}\}$ the set of label sequences. Finally, let $\mathbf{v}^{(s)} = (v_1^{(s)}, \ldots, v_{N_s}^{(s)})$ the binary sequence indicating whether $k_t^{(s)}$ is visible ($v_t^{(s)} = 1$) or hidden ($v_t^{(s)} = 0$) and $\mathcal{V} = \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(M)}\}$; of course, $\mathcal{V}$ is known.

3. MODEL PARAMETER ESTIMATION

3.1. The EM Algorithm

To learn the parameters of the model presented in the previous section, we use a Baum-Welch-type algorithm [6]. Just as the standard BW algorithm is an instance of EM [7] to estimate the parameter of an HMM (which the switched dynamical model is), the algorithm herein presented is the EM algorithm for the semi-supervised learning setting above described.

If both $\mathcal{X}$ and $\mathcal{K}$ were observed, the complete log-likelihood could be written as

$$\begin{aligned}&\log p(\mathcal{X}, \mathcal{K}|\boldsymbol{\theta}) = \\ &\sum_{s=1}^{M}\sum_{t=2}^{N_s}\{\log p(\mathbf{x}_t^{(s)}|\mathbf{x}_{t-1}^{(s)}, k_t^{(s)}) + \log B(k_{t-1}^{(s)}, k_t^{(s)}) \\ &+ \log p(\mathbf{x}_1^{(s)}|k_1^{(s)}) + \log \pi(k_1^{(s)}))\}\end{aligned} \quad (2)$$

where the terms depending on $\mathbf{x}_1, k_1$ are assumed known; to simplify the notation, we have omitted explicit dependency of the model parameters $\boldsymbol{\theta}$.

The EM algorithm produces a sequence of parameter estimates $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_r, \widehat{\boldsymbol{\theta}}_{r+1}, \ldots$, by maximizing the so-called Q-function. Letting $\widehat{\boldsymbol{\theta}}$ denote the current parameter estimate and $\widehat{\boldsymbol{\theta}}_{\text{new}}$ its updated value, we have

$$\widehat{\boldsymbol{\theta}}_{\text{new}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}), \quad (3)$$

where $Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}})$ is the conditional expectation of the complete log-likelihood, with respect to the hidden elements of $\mathcal{K}$, given the current parameter estimate $\widehat{\boldsymbol{\theta}}$, and the set of observed sequences $\mathcal{X}$. Formally,

$$Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}) = E\Big[p(\mathcal{X}, \mathcal{K}|\boldsymbol{\theta}) \mid \mathcal{X}, \widehat{\boldsymbol{\theta}}_r\Big]. \quad (4)$$

The computation of this conditional expectation constitutes the E-step and is carried out via forward and backward recursions [6].

3.2. The E-step

The forward/backward recursions, applied to the s-th sequence, yield estimates of $P(k_t^{(s)}|\boldsymbol{x}_t^{(s)})$, $P(k_{t-1}^{(s)}|\mathbf{x}^{(s)})$, and $P(k_{t-1}^{(s)}, k_t^{(s)}|\mathbf{x}^{(s)})$, where $\boldsymbol{x}_t^{(s)} = (\mathbf{x}_1^{(s)}, \ldots, \mathbf{x}_t^{(s)})$, denotes the set of samples of $\mathbf{x}^{(s)}$ up to instant t. All probabilities that depend on $\widehat{\boldsymbol{\theta}}$ are written as $\widehat{P}$.

Forward recursion: the *prediction step* is given by

$$\widehat{P}(k_t^{(s)} = j|\boldsymbol{x}_{t-1}^{(s)}) = \sum_{i=1}^{m} \widehat{B}(i,j)\widehat{P}(k_{t-1}^{(s)} = i|\boldsymbol{x}_{t-1}^{(s)}), \quad (5)$$

The *filtering step* is given by

$$\widehat{P}(k_t^{(s)} = j|\boldsymbol{x}_t^{(s)}) \propto \widehat{P}(\mathbf{x}_t^{(s)}|k_t^{(s)} = j, \mathbf{x}_{t-1}^{(s)})\widehat{P}(k_t^{(s)} = j|\boldsymbol{x}_{t-1}^{(s)}). \quad (6)$$

Backward recursion: this recursion produces estimates of $P(k_{t-1}^{(s)}, k_t^{(s)}|\mathbf{x}^{(s)})$ and $P(k_t^{(s)}|\mathbf{x}^{(s)})$, as follows:

$$\begin{aligned}\widehat{P}(k_{t-1}^{(s)} = i, k_t^{(s)} = j|\mathbf{x}^{(s)}) &= \widehat{P}(k_{t-1}^{(s)} = i|k_t^{(s)} = j, \mathbf{x}^{(s)}) \\ &\quad \times \widehat{P}(k_t^{(s)} = j|\mathbf{x}^{(s)}) \\ &= \widehat{B}(i,j)\frac{\widehat{P}(k_{t-1}^{(s)} = i|\boldsymbol{x}_{t-1}^{(s)})\widehat{P}(k_t^{(s)} = j|\mathbf{x}^{(s)})}{\widehat{P}(k_t^{(s)} = j \mid \boldsymbol{x}_{t-1}^{(s)})}.\end{aligned} \quad (7)$$

and

$$\begin{aligned}\widehat{P}(k_{t-1}^{(s)} = i|\mathbf{x}^{(s)}) &= \sum_{j=1}^{m} \widehat{P}(k_{t-1}^{(s)} = i, k_t^{(s)} = j|\mathbf{x}^{(s)}) \\ &= \widehat{P}(k_{t-1}^{(s)} = i|\boldsymbol{x}_{t-1}^{(s)}) \sum_{j=1}^{m} \frac{\widehat{B}(i,j)\widehat{P}(k_t^{(s)} = j|\mathbf{x}^{(s)})}{\widehat{P}(k_t^{(s)} = j|\boldsymbol{x}_{t-1}^{(s)})}.\end{aligned} \quad (8)$$

3.3. Semi-Supervision

The standard BW algorithm, which assumes that all elements of $\mathcal{K}$ are hidden, defines a set of "weights" $w_{t\,i}^{(s)}$, where $w_{t\,i}^{(s)} = \widehat{P}(k_t^{(s)} = i|\mathbf{x}^{(s)})$, that is, the current estimate of the probability that at time t of sequence s, the active model is i. Similarly, the BW algorithm also defines transition weights $w_{t\,ij}^{(s)} = \widehat{P}(k_{t-1}^{(s)} = i, k_t^{(s)} = j|\mathbf{x}^{(s)})$. These weights are the only information which is needed to compute the Q-function.

In our scenario, we assume that if $v_t^{(s)} = 1$, then $k_t^{(s)}$ is not hidden, but an observed label. This requires defining new modified "weights" $\overline{w}$ as follows

$$\overline{w}_{t\,i}^{(s)} = \begin{cases} w_{t\,i}^{(s)} & \Leftarrow \quad v_t^{(s)} = 0 \\ \delta(i - k_t^{(s)}) & \Leftarrow \quad v_t^{(s)} = 1, \end{cases} \quad (9)$$

where is δ is the Kronecker delta function, i.e., $\delta(a - b) = 1$, if $a = b$, and zero otherwise. Notice that if $v_t^{(s)} = 1$, then $k_t^{(s)}$ is an observed variable. Similarly,

$$\overline{w}_{t\,ij}^{(s)} = \begin{cases} w_{t\,ij}^{(s)} & \Leftarrow (v_{t-1}^{(s)} = 0) \wedge (v_t^{(s)} = 0) \\ \delta(i - k_{t-1}^{(s)})\delta(j - k_t^{(s)}) & \Leftarrow (v_{t-1}^{(s)} = 1) \wedge (v_t^{(s)} = 1) \\ \left\langle \delta(i - k_{t-1}^{(s)}) w_{t\,ij}^{(s)} \right\rangle_j & \Leftarrow (v_{t-1}^{(s)} = 1) \wedge (v_t^{(s)} = 0) \\ \left\langle \delta(j - k_t^{(s)}) w_{t\,ij}^{(s)} \right\rangle_i & \Leftarrow (v_{t-1}^{(s)} = 0) \wedge (v_t^{(s)} = 1), \end{cases}$$

where $\langle\cdot\rangle_u$ denotes normalization such that the sum with respect to u equals one.

3.4. The M-step

The Q-function is simply obtained as in the standard BW algorithm, but using the semi-supervised weights $\overline{w}_{t\,i}^{(s)}$ and $\overline{w}_{t\,ij}^{(s)}$ instead of $w_{t\,i}^{(s)}$ and $w_{t\,ij}^{(s)}$. The parameter estimates are updated according to standard rules of the BW algorithm (see [6] for full details).

In summary, our semi-supervised EM algorithm comprises three steps in each iteration: the standard E-step, which yields the probabilities/weights $w_{t\,i}^{(s)}$ and $w_{t\,ij}^{(s)}$, under the assumption that all the labels are hidden; a "forcing" step (described in Section 3.3) in which the known labels are used to modify these weights; a standard M-step which updates the model parameter estimates.

4. CLASSIFICATION

The classification problem can be stated as follows: *given an observable trajectory* $\mathbf{x}$ *we want to classify into the set of activities* $\mathcal{A} = \{A_1, \ldots, A_L\}$. Each activity A_l is characterized by a model of the form (1), *i.e.*, by the corresponding parameter estimate $\widehat{\boldsymbol{\theta}}_{(l)}$, previously obtained using the EM algorithm above described.

The classification of the sequence $\mathbf{x}$ is obtained by the *maximum a posteriori* (MAP) rule as

$$j = \arg\max_{l} \{p(\mathbf{x}|\widehat{\boldsymbol{\theta}}_{(l)})p(A_l)\}. \tag{10}$$

where $p(A_l)$ is the *a priori* probability of activity A_l, herein taken as $p(A_l) = 1/L$. Each likelihood term $p(\mathbf{x}|\widehat{\boldsymbol{\theta}}_{(l)})$, for $l = 1, ..., L$, can be obtained by running one forward/backward recursion, with the corresponding model parameter estimate $\widehat{\boldsymbol{\theta}}_{(l)}$.

5. EXPERIMENTAL RESULTS

5.1. Synthetic Data

The considered scenario is shown in Fig. 2, which shows typical patterns of trajectories that a person may perform in a corridor of a shopping center. The thin rectangles correspond to areas where the trajectory begins/ends which correspond to the position of the sensors (known model labels). The first sample of the trajectory is random, simulating that the person may appear randomly in the scene. The trajectories are generated according to (1).

The first stage of the algorithm is to estimate $\widehat{\boldsymbol{\theta}}_{(l)}$, for $l = 1, ..., 4$ ("passing", "entering", "exiting", "browsing"). For this purpose, we ran both the standard and the semi-supervised EM (BW) algorithms, with a set of training trajectories, using 5 models ($m = 5$) for all activities. In this experiment we have used about 200 samples per trajectory, with 5% of known labels. Fig. 3 (a) shows the initialization of the EM algorithm; we plot level curves of the 5 Gaussian densities of means $\mathbf{T}_k$ and covariances $\mathbf{R}_k$, for $k = 1, ..., 5$, and the dots represent the observed displacements $\mathbf{x}_t - \mathbf{x}_{t-1}$. Fig. 3(b) and (c) shows the estimates obtained by the standard and the semi-supervised algorithms, respectively. It's clear that semi-supervision allowed the correct estimation of five models which have the following semantics: "moving left", "moving right", "moving up", "moving down", "stopped". To assess the classification accuracy, we have generated another 210 trajectories from each activity, and have classified them according to (10). The parameter estimates obtained in the semi-supervised mode lead to 100% classification accuracy, while the fully unsupervised estimates lead to 88% accuracy.

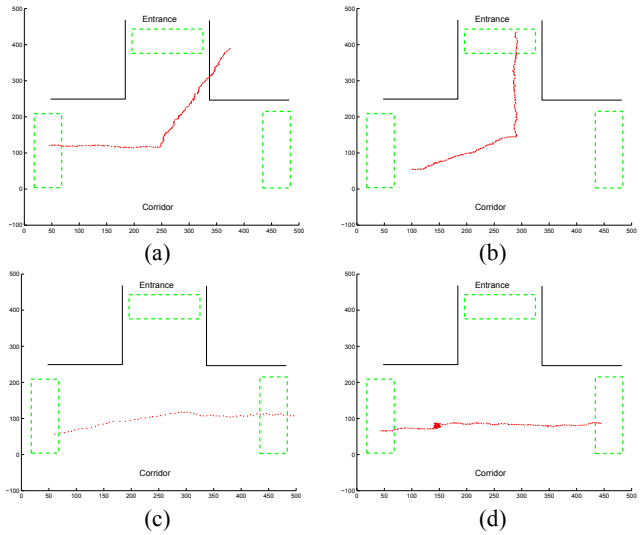

Fig. 2. Several synthetic activities: (a) entering, (b) leaving, (c) passing, (d) browsing.

5.2. Real Data

The proposed approach was also tested in real data collected in the context of CAVIAR project [8]: 64 video sequences hand labeled with the ground truth (this data is available at [8]). These sequences include indoor shopping center observations of individual and groups of pedestrians. Fig. 4 (a) and (b) show the results obtained without (a) and with (b) semi-supervised training. As in the synthetic case, the semi-supervision allowed estimating five underlying models which approximately correspond to five motion patterns: "moving left", "moving right", "moving up", "moving down", "stopped".

Fig. 5 shows several real trajectories, i.e, evolution of the centroid of the bounding box, as well as the corresponding activity classifier output. To assess the classification accuracy, we have classified 51 trajectories, using the MAP criterion (10). Using the parameter estimates obtained with semi-supervision, the accuracy obtained was was 90.0%, while without semi-supervision, the accuracy dropped to 80.3%.

6. CONCLUSIONS

In this work we have presented a semi-supervised framework for modeling and recognition of human trajectories with application to surveillance. The method uses switched dynamical models, with each model describing a specific motion regime. We have shown how to modify the classical Baum-Welch (BW) algorithm to take into account a subset of known model labels, leading to a semi-supervised BW algorithm. The experimental results reported, with both synthetic and real data, validate the method by showing that semi-supervision leads to a higher classification accuracy than the unsupervised version. Future work will include more complex events containing more switching times and extension to other applications.

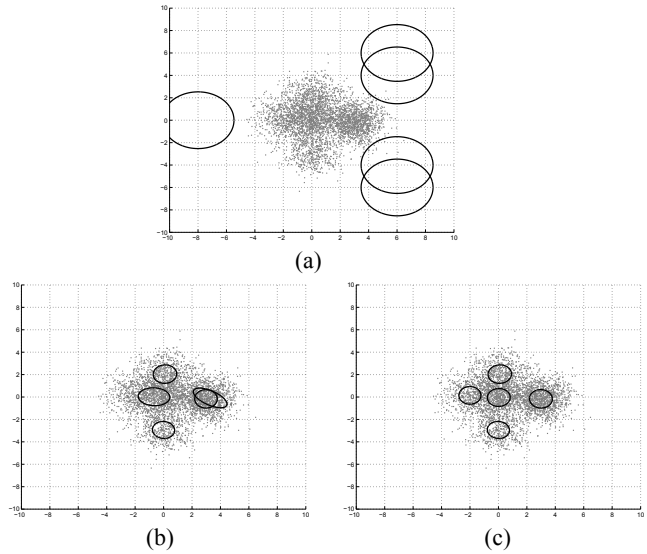

Fig. 3. Estimates of the dynamical models in the synthetic case: (a) initialization; (b) estimates without semi-supervision (after 20 iterations); (c) estimates with semi-supervision (after 10 iterations).

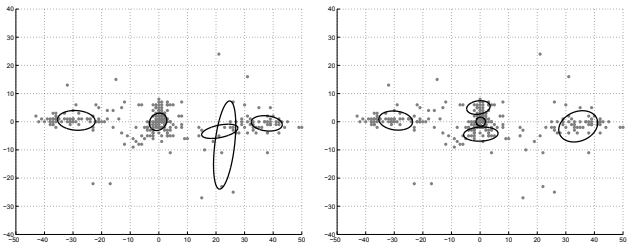

Fig. 4. Estimates of the dynamical models in the real case: estimates without (left) and with (right) semi-supervision.

Fig. 5. Several synthetic activities: (a) entering, (b) leaving, (c) passing, (d) browsing.

7. REFERENCES

[1] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 22, no. 8, pp. 831–843, 2000.

[2] H. Bui, S. Venkatesh, and G. West, "Policy recognition in the abstract hidden Markov model," *Journal of Artificial Intelligence Research*, vol. 17, pp. 451–499, 2002.

[3] T. T. Truyen, H. H. Bui, and S. Venkatesh, "Human activity learning and segmentation using partially hidden discriminative models," in *Workshop on Human Activity Recognition and Modelling HAREM'2005*, Oxford, UK, pp. 87–95, 2005.

[4] N. T. Nguyen and S. Venkatesh, "Discovery of activity structures using the hierarchical hidden markov models," in *British Machine Vision Conference*, Oxford, UK, pp. 409–418, 2005.

[5] J. C. Nascimento, M. A. T. Figueiredo, and J. S. Marques, "Segmentation and classification of human activities," in *British Machine Vision Conf., Int. Workshop on Human Activity Recognition and Modelling - HAREM 2005*, Oxford, UK, pp. 79–86, 2005.

[6] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[7] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm." *Jour. Royal Statist. Soc. (B)*, vol. 39, pp. 1–38, 1977.

[8] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm