CLASSIFICATION OF COMPLEX PEDESTRIAN ACTIVITIES FROM TRAJECTORIES

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

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

ABSTRACT

We propose a method to classify human trajectories, modeled by a set of motion vector fields, each tailored to describe a specific motion regime. Trajectories are modeled as being composed of segments corresponding to different motion regimes, each generated by one of the underlying motion fields. Switching among the motion fields follows a probabilistic mechanism, described by a field of stochastic matrices. This yields a space-dependent motion model which can be estimated using an *expectation-maximization* (EM) algorithm. To address the model selection question (how many fields to use?), we adopt a discriminative criterion based on classification accuracy on a held out set. Experiments with real data (human trajectories in a shopping mall) illustrate the ability of the proposed approach to classify complex trajectories into high level classes (client versus non-client).

1. INTRODUCTION AND PRIOR WORK

Activity recognition is a central topic in video surveillance tasks [3, 7]. The methods proposed to address this problem depend on the type of environment and application. The most recent approaches consider essentially two scenarios: *short range activities* (SRA) and *long range activities* (LRA). For SRA, the human body occupies a significant fraction of the image; accordingly, shape features, such as silhouettes, can be used to classify the activities [15, 10]. In LRA scenarios, *i.e.*, in surveillance of wide areas, pedestrians usually occupy a very small image area, sometimes just a few pixels, precluding any accurate estimation of shape features. In this case, pedestrian trajectories are the commonly used feature for activity recognition. These trajectories (usually of the center of mass of the pedestrian) are obtained by tracking algorithms; examples of such trajectories are depicted in Fig. 1. The work reported in this paper addresses the problem of trajectory recognition for LRA scenarios.

Several trajectory analysis problems (namely classification and clustering) have been addressed using pairwise (dis)similarity measures between trajectories; these include Euclidean [2] and Hausdorff distances [13]. However, since the duration of the trajectories is not constant, sequence alignment techniques are needed for meaningful comparisons; this has been done using dynamic time warping [4] and other techniques [12]. However, distance between trajectories may not be appropriate to describe the nature of the underlying activities; *i.e.*, two trajectories may be spatially close and still correspond to quite different activities. Alternatively, trajectory modeling can be viewed as a problem of semantic scene modeling, since knowledge about the scene context plays an important role in activity recognition. In methods that exploit this idea, semantically meaningful events, such as trajectory intersections or crossing entry/exit points are used [6, 13].

This work was supported by project the FCT (ISR/IST plurianual funding) through the PIDDAC Program funds and by project PTDC/EEA-CRO/098550/2008 (ARGUS).

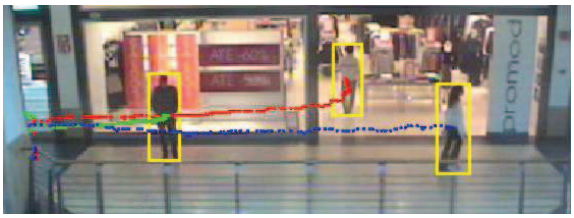


Fig. 1. Examples of pedestrian's trajectories performing different activities.

The approach herein proposed is an extension of the framework presented in [8], in which we introduced a novel approach for modeling trajectories in natural image sequences. In this paper, we show how those (generative) models can be used to deal with trajectory classification in a LRA surveillance setting.

We model the trajectories as being generated by a set of motion (vector) fields, corresponding to different motion regimes. Switching between these fields is allowed and controlled by a space-varying probabilistic mechanism; specifically, a field of stochastic matrices. This model (*i.e.*, the motion and switching fields) can be learnt from a set of observed trajectories using an expectation-maximization (EM) algorithm [8], in which the label of the active field at each time instant is treated as missing/hidden data.

The paper is organized as follows. Section 2 describes the generative model while the learning algorithm is presented in Section 3. Activity classification is described in Section 4. Experimental results are presented in Section 5. Section 6 concludes the paper.

2. GENERATIVE MOTION MODEL

We denote the set of vector motion fields as $\mathcal{T} = \{\mathbf{T}_1, \ldots, \mathbf{T}_K\}$, where each $\mathbf{T}_k : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a vector (velocity, *i.e.*, displacement in one time unit) field. The velocity vector at point $\mathbf{x} \in \mathbb{R}^2$ of the k-th field is denoted as $\mathbf{T}_k(\mathbf{x})$. At each time instant, one of these velocity fields is *active*, *i.e.*, is driving the motion (generating the trajectory); accordingly, the motion model is

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{T}_{k_t}(\mathbf{x}_{t-1}) + \mathbf{w}_t, \quad t = 2, ..., L, \qquad (1)$$

where $k_t \in \{1, ..., K\}$ is the label of the active field at time t, $\mathbf{w}_t \sim \mathcal{N}(0, \sigma^2_{k_t}\mathbf{I})$ is white Gaussian noise with zero mean and variance $\sigma^2_{k_t}$ and L is the length of the trajectory. The initial position is assumed to follow some known distribution $p(\mathbf{x}_1)$. The conditional probability density of a trajectory $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$, given the sequence of active models $\mathbf{k} = \{k_1, \ldots, k_L\}$ is thus

$$p(\mathbf{x}|\mathbf{k}, \mathcal{T}, \boldsymbol{\sigma}) = p(\mathbf{x}_1) \prod_{t=2}^{L} p(\mathbf{x}_t|\mathbf{x}_{t-1}, k_t), \tag{2}$$

where $\boldsymbol{\sigma} = (\sigma^2_1, ..., \sigma^2_K)$ and

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, k_t) = \mathcal{N}(\mathbf{x}_t|\mathbf{x}_{t-1} + \mathbf{T}_{k_t}(\mathbf{x}_{t-1}), \sigma^2_{k_t}\mathbf{I}), \tag{3}$$

with $\mathcal{N}$ denoting, as usual, a Gaussian density.

We model the sequence of active fields $\mathbf{k} = (k_1, \ldots, k_L)$ as a realization of a first order Markov process, with some initial distribution $P(k_1)$, and a space-varying transition matrix, *i.e.*,

$$P(k_t = j|k_{t-1} = i, \mathbf{x}_{t-1}) = B_{ij}(\mathbf{x}_{t-1}),$$

where $\mathbf{B} : \mathbb{R}^2 \to \mathbb{R}^{K\times K}_+$ is a field of stochastic matrices,

$$\mathbf{B}(\mathbf{u}) = \begin{bmatrix} B_{1,1}(\mathbf{u}) & \cdots & B_{1,K}(\mathbf{u}) \\ \vdots & \ddots & \vdots \\ B_{K,1}(\mathbf{u}) & \cdots & B_{K,K}(\mathbf{u}) \end{bmatrix}, \tag{4}$$

i.e., such that $\sum_{j=1}^{K} B_{ij}(\mathbf{u}) = 1$, for any $\mathbf{u}$ and any $i \in \{1, ..., K\}$. We assume that the initial label follows some known initial distribution $P(k_1)$.

Finally, the joint distribution of a trajectory $\mathbf{x}$ and the underlying hidden sequence of active field labels, $\mathbf{k}$, is

$$p(\mathbf{x}, \mathbf{k}|\mathcal{T}, \mathbf{B}, \boldsymbol{\sigma}) = p(\mathbf{x}_1)P(k_1) \prod_{t=2}^{L} p(\mathbf{x}_t|\mathbf{x}_{t-1}, k_t)P(k_t|k_{t-1}, \mathbf{x}_{t-1}), \tag{5}$$

where $p(\mathbf{x}_t|\mathbf{x}_{t-1}, k_t)$ is as given in (3) and $P(k_t|k_{t-1}, \mathbf{x}_{t-1}) = B_{k_{t-1},k_t}(\mathbf{x}_{t-1})$.

An interesting feature of this motion model, for surveillance applications, is its interpretability. Each velocity field describes a different type of motion in the scene, a piece of information which can be used, for example, by the manager of a public area to characterize the typical ways in which people move in that area.

3. LEARNING THE MODEL

The model learning problem consists in estimating the set of velocity fields $\mathcal{T}$, the field of transition matrices $\mathbf{B}$, and the set of noise variances $\boldsymbol{\sigma} = \{\sigma^2_1, ..., \sigma^2_K\}$, from a set of observed trajectories. Specifically, we assume that we have a training set of S independent trajectories $\mathcal{X} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(S)}\}$, where $\mathbf{x}^{(j)} = (\mathbf{x}^{(j)}_1, ..., \mathbf{x}^{(j)}_{L_j})$ is the j-th observed trajectory, with length L_j. Naturally, the corresponding set of sequences of active fields, $\mathcal{K} = \{\mathbf{k}^{(1)}, ..., \mathbf{k}^{(S)}\}$, is not observed. We denote the complete set of fields and parameters to be estimated as $\boldsymbol{\theta} = (\mathcal{T}, \mathbf{B}, \boldsymbol{\sigma})$.

3.1. Model Estimation Criterion: Marginal MAP (MMAP)

The fact that the active field labels $\mathcal{K}$ are missing suggests the use of an EM algorithm to find a *marginal maximum a posteriori* (MMAP) estimate of $\boldsymbol{\theta}$ under some prior $p(\boldsymbol{\theta}) = p(\mathcal{T})p(\mathbf{B})p(\boldsymbol{\sigma})$; formally, the set of parameters $\boldsymbol{\theta}$ are given as

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} p(\mathcal{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) \sum_{\mathcal{K}} p(\mathcal{X}, \mathcal{K}|\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) \prod_{j=1}^{S} \sum_{\mathbf{k}^{(j)}\in\{1,\ldots,K\}^{L_j}} p(\mathbf{x}^{(j)}, \mathbf{k}^{(j)}|\boldsymbol{\theta}) \end{aligned} \tag{6}$$

where each $p(\mathbf{x}^{(j)}, \mathbf{k}^{(j)}|\boldsymbol{\theta})$ has the form given in (5). Clearly, this maximization can not be solved in closed form; next, we present an EM algorithm for solving it.

3.2. The EM algorithm

The complete log-likelihood is given by

$$\log p(\mathcal{X}, \mathcal{K}|\boldsymbol{\theta}) = \sum_{j=1}^{S} \log p(\mathbf{x}^{(j)}, \mathbf{k}^{(j)}|\boldsymbol{\theta}), \tag{7}$$

where each term $p(\mathbf{x}^{(j)}, \mathbf{k}^{(j)}|\boldsymbol{\theta})$ has the form (5). Let us introduce binary indicator variables to represent the model labels (which are missing data, in the EM framework): each label $k^{(j)}_t \in \{1, ..., K\}$ (the active field at time t of trajectory j) is represented by a binary vector $\mathbf{y}^{(j)}_t = (y^{(j)}_{t,1}, ..., y^{(j)}_{t,K}) \in \{0, 1\}^K$, where $y^{(j)}_{t,l} = 1 \Leftrightarrow k^{(j)}_t = l$. With this notation, the complete log-likelihood becomes

$$\begin{aligned} \log p(\mathcal{X}, \mathcal{K}|\boldsymbol{\theta}) &= \sum_{j=1}^{S}\sum_{t=2}^{L_j}\sum_{l=1}^{K}\sum_{g=1}^{K} y^{(j)}_{t-1,g}\, y^{(j)}_{t,l} \log B^{(j)}_{g,l}(\mathbf{x}^{(j)}_{t-1}) \\ &+ \sum_{j=1}^{S}\sum_{t=2}^{L_j}\sum_{l=1}^{K} y^{(j)}_{t,l} \log \mathcal{N}(\mathbf{x}^{(j)}_t|\mathbf{x}^{(j)}_{t-1} + \mathbf{T}_l(\mathbf{x}^{(j)}_{t-1}), \sigma^2_l\mathbf{I}) + C, \end{aligned} \tag{8}$$

where C is an irrelevant constant.

The E-step computes the conditional expectation (with respect to the missing variables $\mathcal{K}$) of the complete log-likelihood (8), given the current estimates of the parameters $\widehat{\boldsymbol{\theta}}$ and the observations $\mathcal{X}$:

$$Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}) \equiv \mathbb{E}_{\mathcal{K}}\left[\log p(\mathcal{X}, \mathcal{K}|\boldsymbol{\theta})|\mathcal{X}, \widehat{\boldsymbol{\theta}}\right].$$

Given the linearity of $\log p(\mathcal{X}, \mathcal{K}|\boldsymbol{\theta})$ with respect to the binary indicators $y^{(j)}_{t,l}$ and the (also binary) switching indicators $y^{(j)}_{t-1,g}\, y^{(j)}_{t,l}$, computing this conditional expectation corresponds to computing the conditional expectations of these binary variables, which are then plugged into $\log p(\mathcal{X}, \mathcal{K}|\boldsymbol{\theta})$. Notice that these conditional expectations can be obtained by a simple modified forward-backward procedure [9]; the modification involves taking into account a varying transition matrix.

In the M-step, the model estimates are updated according to

$$\widehat{\boldsymbol{\theta}}_{\text{new}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}) + \log p(\boldsymbol{\theta}). \tag{9}$$

This maximization is not trivial and requires an adequate finite-dimensional representation of the motion and switching fields. For full details on this, the reader is referred to our previous work [8].

4. ACTIVITY CLASSIFICATION

The generative motion model introduced in Section 2 can be easily used to build a *maximum a posteriori* (MAP) trajectory classifier. For that purpose, consider a collection of sets of trajectories, $\{\mathcal{X}^{(a)},\ a = 1, ..., A\}$, each assumed to have been produced by one of the A activities. A direct approach to using the proposed generative models is to simply estimate A generative models, $\{\widehat{\boldsymbol{\theta}}^{(a)},\ a = 1, ..., A\}$, from these sets of trajectories, using the EM algorithm described in Section 3.2. With these model estimates, and given a new trajectory $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$, the MAP activity classifier is given by

$$\widehat{a}(\mathbf{x}) = \arg \max_{a \in \{1,\ldots,A\}} p(\mathbf{x}|\widehat{\boldsymbol{\theta}}^{(a)})\ P(a), \tag{10}$$

where $P(a)$ is the prior probability for activity a (in this paper, we take $p(a) = 1/A$), and $p(\mathbf{x}|\boldsymbol{\theta}^{(a)})$ is the probability density function of the trajectory $\mathbf{x}$ under the model with parameters $\widehat{\boldsymbol{\theta}}^{(a)}$, which can be easily computed using the forward-backward procedure [9].

In this paper, we consider a variant of the approach described in the previous paragraph, motivated by the following observation. In some scenarios, the types of motion regimes underlying the classes tend to be the same; what distinguishes the classes is essentially the way the trajectories switch among those motion regimes. When this is a valid hypothesis, it makes sense to estimate a common set of motion fields (shared by all the classes), but a switching field for each class. The motion fields estimates thus obtained benefit from being based on a larger number of trajectories. Moreover, this choice requires a very minor change in the EM algorithm. As explained in Section 5, the experiments reported in this paper use this option of shared motion fields.

One last question that has to be faced is that of choosing the number of vector fields for the generative model. We resort to a *discriminative model selection* criterion [11], where instead of resorting to generative criteria, such as minimum description length (MDL), minumum message length (MML), or the Akaike information criterion (AIC) [5], we select the set of generative models achieving the best performance in terms of classification. This model selection is carried out on a held out training subset. Of course, this can be seen as a simplified (and cheaper) version of cross validation.

5. EXPERIMENTAL RESULTS

In this section we present results with real data, concerning typical trajectory classes in a shopping mall. In the following, two main classes are considered for the trajectories in front of, and inside, a store: *client* and *non-client*. The trajectories of these two classes are obviously different. In the first one (*client*), the pedestrian enters into the mall, stays for a considerable amount of time, during which he/she enters the store and then leaves the scene. Fig. 2 shows some examples of the *client* class. In typical trajectories of the *non-client* class, the pedestrian enters and leaves the scene, maybe entering the store but not for a long time, or simply passing in the front of the store. Fig. 3 shows several trajectories of the *non-client* class. In Fig. 3 (a) and (d), the pedestrian enters the scene, then enters the store for a short time and finally leaves the scene; in (b) and (e), the pedestrian never fully enters the store, simply browsing at the entrance; in (c) and (f), the pedestrian first browses at the entrance but then takes a quick walk inside the store and finally leaves.

Notice the difference between the scene view in Fig. 1 and the view in Figs. 2 and 3 (and also in Fig. 4). This difference is due to the fact that the trajectories are shown on a so-called *bird's eye view*, obtained by a projective transformation (homography) between the image and a plane parallel to the ground. This is a common procedure in computer vision, which is done to compensate for the apparent speed variation with the distance from the camera.

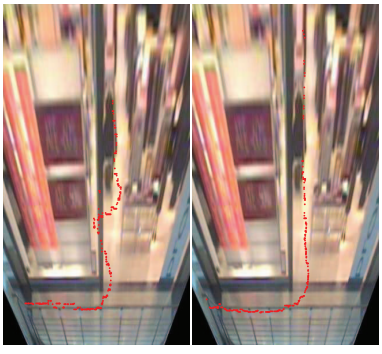
It is clear that, in this problem, it is the switching pattern, not the underlying motion regimes, that distinguish the two classes. Accordingly, as described in Section 4, we use a common set of motion fields and a different switching field for each class. According to the model selection criterion described above, the best performance was obtained using four motion fields, which is the number adopted.

We had a total of 58 trajectories, from which we used 10 for training and the remaining ones for testing. Fig. 4 shows the estimates of the four vector fields obtained by the EM algorithm. From this figure, we see that the four vector fields represent essentially the following four motion regimes: inwards motion; outwards motion; left-to-right motion; and right-to-left motion.

Table 1. Confusion matrix of the proposed classifier on the *Client* and *Non-Client* classes.

	Classifier output	
	Client	Non-Client
True class = client	93.0%	7.0%
True class = non-client	9.8%	90.2%

Table 1 show the performance of the proposed approach, where 93% and 90.2% accuracy is achieved for the *Client* and *Non- client* classes, respectively. These results allow concluding that the proposed approach achieves a good performance in identifying complex activities, with high level semantic meaning (in this case, being or not a store client), using only low level data (trajectories).



(a) (b)

Fig. 2. Examples of trajectories belonging to client class; (a) the person enters in the right direction, (b) the person leaves the mal in the opposite direction.

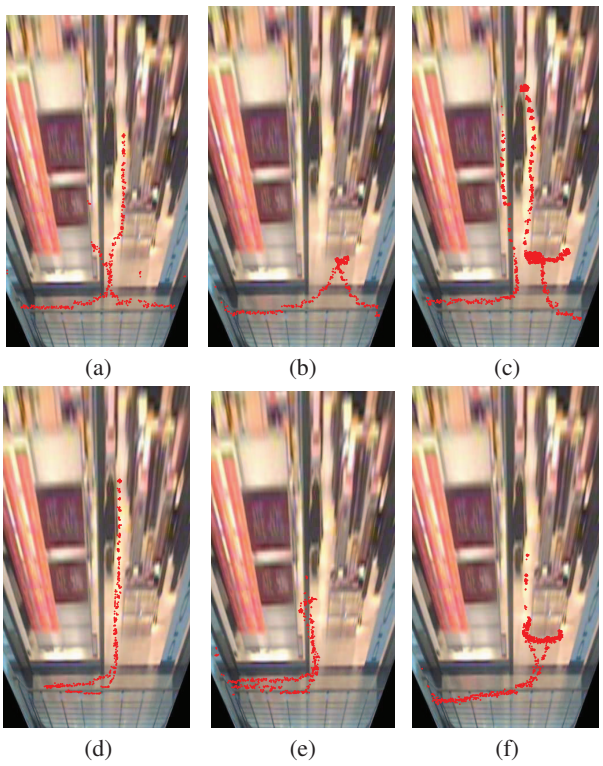


Fig. 3. Examples of trajectories belonging to non-client class. Notice the high variability exhibited by these trajectories.

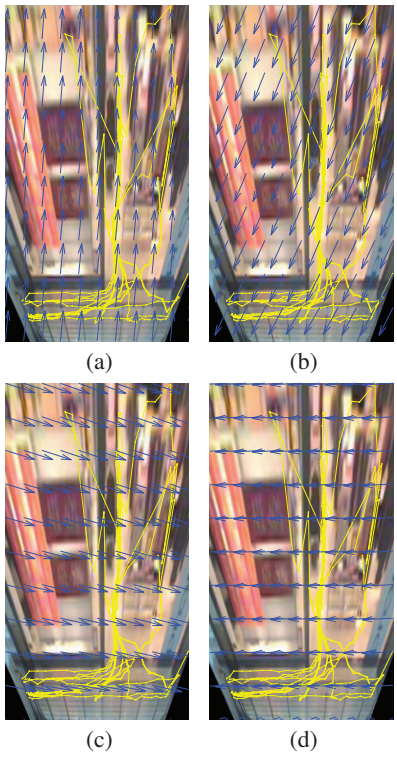


Fig. 4. Vector fields estimates provided by the EM algorithm.

6. CONCLUSIONS

We have presented a new method for classification trajectories in tailored to LRA surveillance settings. The method is based on mixtures of vector fields that allow modeling space-dependent motions. We have presented an EM algorithm to estimate the underlying motion fields along with the space-dependent switching probabilistic model. Experiments using real data have shown that the proposed approach is able to classify complex activities, with high level semantic meaning, based only on low level trajectory data.

Directions of further work will include: (i) study of other model selection strategies; (ii) application of the proposed approach to other types of data, such as crowds [1] and traffic flows.

7. REFERENCES

[1] S. Ali, and M. Shah, "A Langragian particle dynamics approach for crowd flow segmentation and stability analysis," *IEEE Proc. CVPR.*, 2007. 4

[2] Z. Fu, W. Hu, and T. Tan, "Similarity based vehicle trajectory clustering and anomaly detection," in *Proc. of ICIP*, 2005. 1

[3] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviours" *IEEE Trans. on Systems and CyberneticsPart C: Applications and Reviews* 34 (3):334-352, 2004. 1

[4] E. Keogh and M. Pazzani, "Scaling up dynamic time warping for datamining application," in *Proc. of Int. Conf. on Knowledge Discovery and Data Mining*, 2000, pp. 285–289. 1

[5] A. D. Lanterman, "Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model selection," *International Statistical Review*, vol. 69, pp. 185–212, 2001. 3

[6] D. Makris and T. Ellis, "Automatic learning of an activity based semantic scene model," in *Proc. of AVSBS*, 2003. 1

[7] T. Moeslund, A. Hilton and V. Kruger, "A survey of advances in vision-based human motion capture analysis" *Comp. Vision Image Understanding*, vol. 104, 90-126, 2006. 1

[8] J. Nascimento and M. Figueiredo and J. Marques, "Trajectory analysis in natural images using mixtures of vector fields," in *Proc. of ICIP*, pp. 4353-4356, 2009. 1, 2

[9] L. 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. 2, 3

[10] C. Rao, A. Yilmaz, M. Shah, "View-invariant representation and recognition of actions", *Int. Journal of Computer Vision*, 203-226, 2004. 1

[11] B. Thiesson and C. Meek, "Discriminative model selection for density models," in *9th Int. Workshop on Artif. Intell. and Statistics – UAI*, 2003. 3

[12] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering similar multidimensional trajectories," in *Proc. of Int. Conf. on Data Engineering*, 2002, pp. 673–685. 1

[13] X. Wang, K. Tieu, and E. Grimson, "Learning semantic scene models by trajectory analysis," in *Proc. of ECCV*, 2006. 1

[14] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson, "Trajectory analysis and semantic region modeling using a nonparametric Bayesian model," in *Proc. of CVPR*, 2008.

[15] R. Wren, A. Azarbayejani, T. arrell, A. Pentland, "Pfinder: real-time tracking of the human body", *IEEE Trans. on PAMI*, vol. 19, no. 7, pp. 780-785, 1997. 1