MULTIPLE AGENTS REPRESENTATION USING MOTION FIELDS

Catarina Barata *Jacinto C. Nascimento* *Jorge S. Marques*

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

ABSTRACT

Providing reliable descriptions of the agents in a video scene is an essential task in many applications, such as surveillance. However, most works focus solely on the characterization of pedestrians, which is not sufficient to describe complex scenes, where a variety of vehicles (*e.g.*, bikes and cars) are also present. In this work we address this limitation and propose a framework based on switching motion fields to efficiently characterize the different agents in a scene. Our method achieves a balanced accuracy of 91.9% on the identification of *bikers* and *pedestrian* classes on three challenging scenarios, and provides comprehensive information about their behaviors.

***Index Terms*—** Surveillance, Trajectory Analysis, Multi-agent Identification, Motion Fields

1. INTRODUCTION

The research community has shown a great interest in tasks related with video analysis. This is mainly due to its wide range of potential applications, that spam from sport performance analysis to surveillance systems or self-driving cars [1, 2]. A key factor that is common to most of the applications is the requirement of methods that are able to provide reliable descriptions of the agents in video scenes and/or their movements. Such information is useful many tasks, *e.g*, tracking agents in the video [3, 4], identifying abnormal behaviors in surveillance scenarios [5, 6], and recognizing activities [7].

Most of the works on the aforementioned topics focus solely on describing the behavior of one class of agent, usually pedestrians. Although this is a reasonable assumption in some scenarios (*e.g*, indoor sports), it will not be sufficient to describe complex and crowded scenes such as the recently released *Stanford Drone Dataset* [3], where there are several types of agents (bikers, skaters, cars, pedestrians) all interacting at the same time (see Fig. 1). Some works addressed the multi-agent problem considering that the pedestrians represent normal examples and that any other agent will be an abnormality [5, 6]. However, abnormality is a subjective concept that incorporates other (abnormal) behaviors such as fighting, jaywalking, or vehicles that move on the wrong direction. Moreover, it is not expected for pedestrians and vehicles to exhibit similar motions. Additionally, different types of vehicles (*e.g.*, bicycles and cars) may also show variability. Hence, it is important to address the representation of multiple agents in complex scenes.

Fig. 1. Image from the *Stanford Drone Dataset* [3]. The bounding box colors identify different agents.

In this work we aim to address this issue, and demonstrate that a switching probabilistic model based on motion fields [8] is able to separately describe different agents in a surveillance scene. Our main contributions are the estimation of agent-specific motion models and the proposal of suitable metrics to recognize the class of the agent based on its movement.

2. RELATED WORK

Methods for video analysis may be divided into two groups: those that rely on a pixel-based analysis of the video frames and those that are trajectory based [9]. Both groups have pros and cons. Pixel based do not require explicit tracking of the agents in the video, thus being robust to tracker failures. However, they require the analysis of images and may be less appropriate to deal with far field scenarios, such as the *Stanford Drone Dataset* [3] (see Fig. 1), where the camera is too far way from the agents to provide detailed information of their poses and body motions. On the other hand, trajectory based methods are able to deal with far field scenarios, provide relevant information about movement patterns in a scene, and

This work was supported by the FCT project [PTDC/EEIPRO/0426/2014] and supported by FCT plurianual funding [UID/EEA/50009/2013].

are interpretable. Moreover, trajectory based methods do not require the actual videos, since they rely on sequences of positions, which can be conveyed by other types of data (*e.g.*, GPS [10]).

The switching motion model used in this work belongs to the category of trajectory models. In particular, it fits in the class of models that assume that the trajectories of the agents are constrained by their interactions with the geometry of the scenes (*e.g.*, a pedestrian will go around an obstacle and a car will circulate on the road and not on a sidewalk). Under this assumption, it is possible to summarize the motions into a finite set of patterns. An alternative to the aforementioned methods are those that rely on the social forces model, where the trajectories are governed by attractive and repulsive forces between agents [11].

It is also possible to further separate the trajectory models into deterministic and generative groups. The former comprise methods based on clustering (*e.g.*, vector field k-means [10]) and supervised classification, such as deep learning methods [12], while the latter includes methods that assume the existence of a dynamical equation that governs the movement. Examples of generative methods are those based on Gaussian processes [13], Dirichlet processes [14], motion fields [8], and more recently circular distributions [4]. The proposed model uses motion fields, thus it fits in the category of generative methods.

3. SWITCHED MOTION MODEL

A complex outdoor scene usually comprises several agents that may belong to different classes, such as pedestrians or cars. We postulate that different agents exhibit distinct motion patterns, which may be used to identify them. We also assume that these motion regimes can be represented using motion fields [8], where $T_k^c : [0,1]^2 \rightarrow \mathbb{R}^2$ is the $k-th$ motion field associated with agent class $c \in \{1, ..., C\}$, $[0,1]^2$ denotes the image plane and $k \in \{1, \ldots, K^c\}$ is the type of motion regime. Each agent class c is associated with a set of possible motion fields, but only one motion field governs the trajectory at each time instant. Nonetheless, the model is flexible enough to allow the switching between motions at specific positions in the scene.

Based on this formulation, the position x_t^c is given by the following dynamical equation

$$x_t^c = x_{t-1}^c + T_{k_t}^c(x_{t-1}^c) + w_{k_t}^c, \tag{1}$$

where $T_{k_t}^c(x_{t-1}^c)$ is the class-specific active motion field that governs the movement of the agent at time instant t and $w_{k_t}^c \sim N(0, \Sigma_{k_t}^c)$ is white noise perturbation, which defines the uncertainty associated with the position. The transition between motion fields is modeled as a first order Markov process, with space-varying probabilities

$$P(k_t = j | k_{t-1} = i, x_{t-1}^c) = b_{ij}^c(x_{t-1}^c), \tag{2}$$

where $b_{ij}^c(x_{t-1}^c)$ is the element ij of a stochastic transition matrix $B^c(x_{t-1}^c)$, computed at position x_{t-1}^c.

The class-specific motion models are discretized over a regular grid of $\sqrt{n} \times \sqrt{n}$ nodes, such that each node i of the grid is associated with a set of displacement vectors $T_k^{c,i}$, a set of transition matrices $B^{c,i}$, and a set of noise covariances $\Sigma_k^{c,i}$. In positions outside the nodes, these parameters are obtained by bilinear interpolation

$$\begin{aligned} T_k^c(x) &= \sum_{i=1}^{n} T_k^{c,i} \phi^i(x) \\ B^c(x) &= \sum_{i=1}^{n} B^{c,i} \phi^i(x) \\ \mathbf{\Sigma}_k^c(x) &= \sum_{i=1}^{n} \mathbf{\Sigma}_k^{c,i} \phi^i(x), \end{aligned} \tag{3}$$

where $T_k^{c,i}$ is the motion vector from the $k-th$ field, $B^{c,i} \in \mathbb{R}^{K \times K}$ is the switching matrix, and $\mathbf{\Sigma}_k^{c,i}$ is the $k-th$ covariance matrix, all of them associated with node g^i for the $c-th$ class. The scalar $\phi^i(x)$ is the interpolation coefficient of the $i-th$ node.

Given C separate sets of trajectories, $\mathcal{X}^c = \{x^{(c,1)}, ..., x^{(c,S)}\}$, one per agent class, it is possible to estimate the class-specific model parameters $\theta = (\mathcal{T}^c, \mathcal{B}^c, \mathbf{\Sigma}^c)$ using the EM-algorithm [8, 15, 16], where the hidden variables are the sequences of active motion fields $\mathcal{K}^c = \{k^{(c,1)}, ..., k^{(c,S)}\}$.

4. TRAJECTORY CLASS IDENTIFICATION

In this work we assume that different agents in an outdoor scene will exhibit different behaviors and movement patterns that can be captured using motion fields. Under this assumption, we will start by estimating a separate motion model for each of the possible classes, using the formulation introduced in Section 3. Given the set of C models, it is possible to analyze new trajectories, for which the class is unknown, as we describe next.

i) Trajectory Sampling: Our strategy is suitable to analyze the trajectory either as a whole, or by dividing it into segments. In the second case, we use a moving window strategy to analyze consecutive portions of a trajectory $(x_{t_o}, x_{t_o+1}, ..., x_{t_o+\Delta})$, with an overlap of $\frac{\Delta}{2}$. Both formulations are admissible, however, the analysis by segments allows us to deal with localized abnormalities in a trajectory.

ii) Trajectory Analysis: We assume that by estimating a motion model for each class of agents, we will be able to capture their specific patterns of movement and discriminate between classes. The motion model allows us to define the following set of metrics, which can be used to characterize the whole trajectory or a segment. It is important to stress that while the following metrics are formulated for a trajectory segment, they can easily be extended the entire trajectory.

1. The log-likelihood

$$\log p(x_{t_o},...,x_{t_o+\Delta}|\theta^c) = \sum_k \log \left[\prod_{t=t_o+1}^{t_o+\Delta} p(x_t|k_t, x_{t-1}, \theta^c) \times b^c_{k_{t-1},k_t}(x_{t-1}) \right], \quad (4)$$

where $p(x_t|k_t, x_{t-1}, \theta^c) = N(x_t|x_{t-1} + T^c_{k_t}(x_{t-1}), \mathbf{\Sigma}^c_{k_t}(x_{t-1}))$ and the outer sum accounts for all the possible label sequences $k = (k_{t_o+1}, ..., , k_{t_o+\Delta})$. The computation of this metric becomes more computationally demanding as the number of motion fields and the length of the trajectory segment increase.

2. The complete log-likelihood

$$p(x_{t_o},...,x_{t_o+\Delta}; \hat{k}_{t_o},...,\hat{k}_{t_o+\Delta}|\theta^c), \quad (5)$$

which is computed as (4), but without performing the outer sum. In this case, the sequence of active fields $(\hat{k}_{t_o},...,\hat{k}_{t_o+\Delta})$ is estimated using the Viterbi algorithm [17].

3. The auxiliary function of the EM-algorithm [15]

$$U(\theta^c, \theta'^c) = E\left\{\log p(\mathcal{X}, \mathcal{K}|\theta^c)|\mathcal{K}, \theta'^c\right\} + \log p(\theta^c). \quad (6)$$

4. The minimum prediction error ϵ, defined using

$$\epsilon^c = \min_k \epsilon^c_k$$
$$\epsilon^c_k = \sum_{t=t_o+1}^{t_o+\Delta} \|x_t - \hat{x}_{t-1} - T^c_k(\hat{x}_{t-1})\|^2_2, \quad (7)$$

where $k \in \{1,...,K\}$, $\hat{x}_{t-1}$ is the estimated position, and $\hat{x}_{t_o} = x_{t_o}$. In this case we set the same k for the whole segment, meaning we do not consider the possibility of switching between motion regimes.

5. The representation error given by

$$\varepsilon^c = \sum_{t=t_o+1}^{t_o+\Delta} \|x_t - \hat{x}_{t-1} - T^c_{\hat{k}_t}(\hat{x}_{t-1})\|^2_2, \quad (8)$$

where the sequence $(\hat{k}_{t_o},...,\hat{k}_{t_o+\Delta})$ is estimated using the Viterbi algorithm. Contrary to (7), this metric considers the possibility of switching.

These metrics are computed for each of the C sets of class-specific motion models. This means that a trajectory segment will be characterized by C sets of metrics, each associated with one class model.

iii) Trajectory Classification: The last step is to classify the trajectory into one of the possible agent classes. This corresponds to comparing the metrics of the different classes and selecting the class that maximizes $\log p(x_{t_o},...,x_{t_o+\Delta}|\theta)$, $p(x_{t_o},...,x_{t_o+\Delta}; \hat{k}_{t_o},...,\hat{k}_{t_o+\Delta}|\theta)$, or $U(\theta, \theta')$; and the class that minimizes ϵ or ε.

The whole trajectory is classified by averaging the labels across all segments.

Fig. 2. Experimental scenes (1st row) and trajectories (2nd and 3rd rows - pedestrians in blue and bikers in magenta): *Death-circle* (left), *Gates* (mid), and *Little* (right).

Table 1. Experimental datasets - $*$ identifies the reference video for the spatial transformation.

Scene	#Bikers	#Ped.	Train Videos ID	Test Videos ID
Death-circle	1548	917	0*,2,4	1
Gates	458	335	1,3,5	4*
Little	385	200	1,2,3	0*

5. EXPERIMENTAL RESULTS

The experiments were carried on the *Stanford Drone Dataset* [3], which comprises eight different scenes with multiple videos, recorded using a quadcopter platform with a 4k camera. For each of the scenes, the authors have tracked and annotated the agents according to one of six classes $c \in \{pedestrian, bike, skater, car, cart, bus\}$.

In our experiments, we have selected three scenes (see Fig. 2) and divided their videos in training and test sets as summarized in Table 1. These scenes were selected due to their complexity in terms of types of movement and proportion of agents. The different videos were not recorded from the same point of view, thus it was necessary to align them through a spatial transformation, using one of the videos as reference. Regarding the number of classes, we focused on the distinction between *pedestrians* and *bikers*, due to the high number of trajectories for each of these classes.

For each of the scenes we have estimated $C = 2$ motion models, corresponding to the two agent classes. Each model comprises $K = 4$ motion fields, roughly representing the directions *North-South*, *South-North*, *East-West*, *West-East*.

The discriminative power of the models is evaluated using the the balanced accuracy metric (BACC), which averages the recall (Re) of the two classes

$$Re = \frac{TP^c}{N^c}, \quad (9)$$

where TP^c is the number of true positives of class c and N^c is the total number of elements.

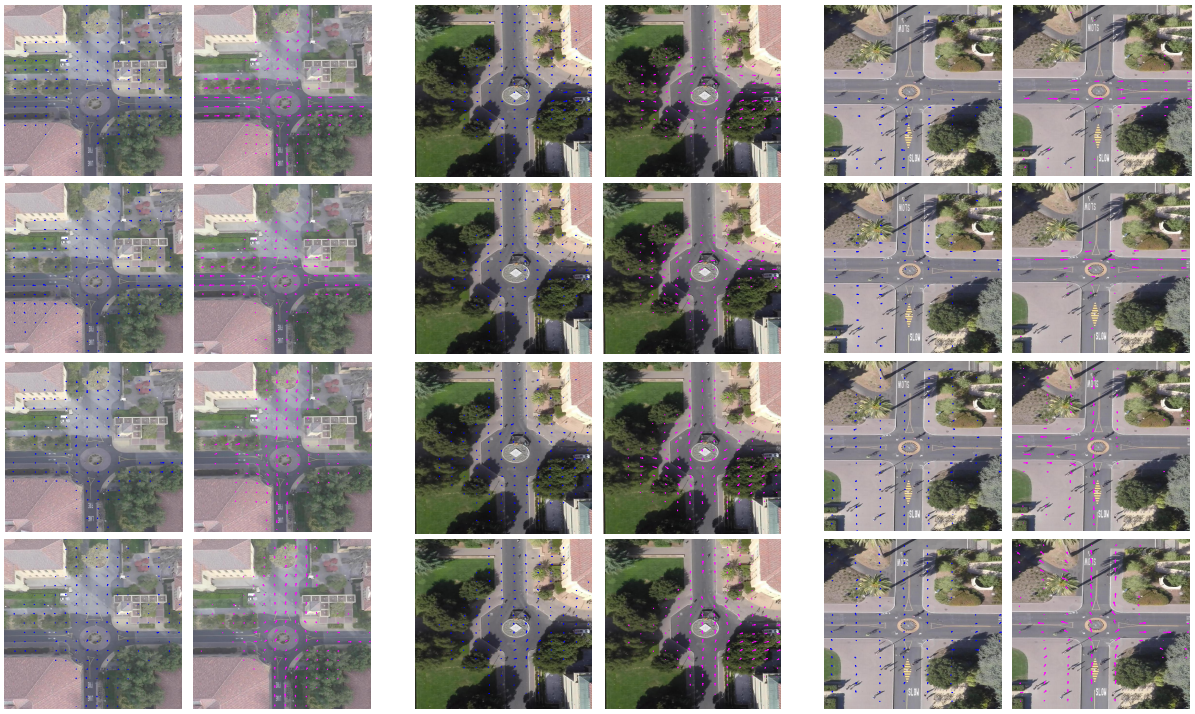

Fig. 3. Estimated motion fields (pedestrians in blue and bikers in magenta): *Death-circle* (left), *Gates* (mid), and *Little* (right).

Table 2. Experimental results in terms of BACC using the *whole* and sampled (*samp.*) trajectories. In **bold** we highlight the best results.

	Death-Circle		*Gates*		*Little*	
Metric	*Whole*	*Samp.*	*Whole*	*Samp.*	*Whole*	*Samp.*
Log-like.	82.7%	83.1%	81.9%	82.8%	81.9%	82.2%
Comp-like.	74.7%	82.0%	81.3%	82.8%	75.6%	80.8%
$U(\theta, \theta')$	82.2%	82.7%	82.6%	83.7%	80.0%	83.7%
ϵ	58.3%	80.2%	61.1%	79.3%	48.7%	83.7%
ε	70.9%	77.3%	69.5%	79.3%	74.1%	85.6%
Log-like+ε	88.2%	**93.4%**	86.9%	88.2%	89.3%	89.3%
U$(\theta, \theta') + \varepsilon$	90.5%	**93.3%**	86.9%	**91.3%**	91.1%	**91.1%**

Fig. 3 shows the estimated fields. As expected, the motion fields are different for the two classes of agents, especially on the case of the *Little* dataset. On one hand, there are certain regions of the scenes, such as the walkways or the roads that are mainly used by only one of the classes. In regions where both pedestrians and bikers trajectories were observed, the velocities of the bikers are usually much higher, as exemplified by the length of the arrows.

Table 2 (rows 1-5) shows the performance of the model in the task of agent recognition, using both the entire trajectory and the trajectory sampling process with $\Delta = 10$, as described in Section 4. All of the classification metrics lead to good performances, with the *log-likelihood* and the auxiliary function $U(\theta, \theta')$ being slightly better across the different scenes. The results show that it is preferable to separately analyze the trajectory segments (*samp.*), as it leads to better results. This was expected from the ϵ metric, since in this case we assume that the motion field is always the same. Nonetheless, this result shows that switching is an important part of the model. The performance of the *complete-likelihood* and ε are also considerably worse when we analyze the whole trajectory, which may be due to the propagation of an error in the estimation of the sequence of active fields using the Viterbi algorithm. The ability of the model to recognize the class of the agent solely using a portion of its trajectory is an important finding, which may have relevant applications in many areas, such as multi-target tracking [3].

We have also investigated the fusion of metrics by averaging their scores. The best combinations are shown in Table 2 (bottom rows). This fusion allows us to improve the results, achieving an average BACC of 91.9% for the three scenes. It also shows that some of the metrics provide complementary information.

6. CONCLUSIONS

This paper proposes a methodology based on switching motion fields to characterize and distinguish multiple agents in surveillance scenarios. The proposed method was evaluated on three challenging scenes from the *Stanford Drone Dataset*, achieving a promising balanced accuracy of 91.9%, on the recognition of *pedestrians* and *bikers*. The estimated motion fields provide comprehensive information about the movement patterns of each class, and the model is able to identify the class of the agent, even when it only has access to a portion of the trajectory.

7. REFERENCES

[1] E. Maggio and A. Cavallaro, *Video tracking: theory and practice*, John Wiley & Sons, 2011.

[2] H. Yao, A. Cavallaro, T. Bouwmans, and Z. Zhang, "Guest editorial introduction to the special issue on group and crowd behavior analysis for intelligent multi-camera video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 405–408, 2017.

[3] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *European conference on computer vision*, 2016, pp. 549–565.

[4] P. Coscia, F. Castaldo, F. A. N. Palmieri, A. Alahi, S. Savarese, and L. Ballan, "Long-term path prediction in urban scenarios using circular distributions," *Image and Vision Computing*, vol. 69, pp. 81–91, 2018.

[5] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2014.

[6] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *Image Processing, 2017 IEEE International Conference on*, 2017, pp. 1577–1581.

[7] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1971–1980.

[8] J. C. Nascimento, M. A. T. Figueiredo, and J. S. Marques, "Activity recognition using a mixture of vector fields," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1712–1725, 2013.

[9] S. Coşar, G. Donatiello, V.a Bogorny, C. Garate, L. O. Alvares, and F. Brémond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 683–695, 2017.

[10] N. Ferreira, J. T. Klosowski, C. E. Scheidegger, and C. T. Silva, "Vector field k-means: Clustering trajectories by fitting multiple vector fields," in *Computer Graphics Forum*, 2013, vol. 32, pp. 201–210.

[11] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 1345–1352.

[12] A. Alahi, V. Ramanathan, K. Goel, A. Robicquet, A. A. Sadeghian, L. Fei-Fei, and S. Savarese, "Learning to predict human behaviour in crowded scenes," *Group and Crowd Behavior for Computer Vision*, pp. 183–207, 2017.

[13] K. Kim, D. Lee, and I. Essa, "Gaussian process regression flow for analysis of motion trajectories," in *ICCV*, 2011, pp. 1164–1171.

[14] V. Bastani, L. Marcenaro, and C. S. Regazzoni, "Online nonparametric bayesian activity mining and analysis from surveillance video," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2089–2102, 2016.

[15] C. Barata, J. C. Nascimento, and J. S. Marques, "A sparse approach to pedestrian trajectory modeling using multiple motion fields," in *Image Processing (ICIP), 2017 IEEE International Conference on*, 2017, pp. 2538–2542.

[16] C. Barata, J. M. Lemos, and J. S. Marques, "Estimation of space-varying covariance matrices," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 4003–4007.

[17] A. J. Viterbi, "A personal history of the viterbi algorithm," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 120–142, 2006.