# Online learning of activities from video

Luis Patino, François Bremond and Monique Thonnat
INRIA Sophia Antipolis Méditerranée - 2004, route des Lucioles - BP 93
06902 Sophia Antipolis Cedex, FRANCE
{Jose-Luis.Patino_Vilchis, Francois.Bremond, Monique.Thonnat}@inria.fr

## Abstract

*The present work introduces a new method for activity extraction from video. To achieve this, we focus on the modelling of context by developing an algorithm that automatically learns the main activity zones of the observed scene by taking as input the trajectories of detected mobiles. Automatically learning the context of the scene (activity zones) allows first to extract a knowledge on the occupancy of the different areas of the scene. In a second step, learned zones are employed to extract people activities by relating mobile trajectories to the learned zones, in this way, the activity of a person can be summarised as the series of zones that the person has visited. For the analysis of the trajectory, a multiresolution analysis is set such that a trajectory is segmented into a series of tracklets based on changing speed points thus allowing differentiating when people stop to interact with elements of the scene or other persons. Tracklets allow thus to extract behavioural information. Starting and ending tracklet points are fed to a simple yet advantageous incremental clustering algorithm to create an initial partition of the scene. Similarity relations between resulting clusters are modeled employing fuzzy relations. These can then be aggregated with typical soft-computing algebra. A clustering algorithm based on the transitive closure calculation of the fuzzy relations allows building the final structure of the scene. To allow for incremental learning and update of activity zones (and thus people activities), fuzzy relations are defined with online learning terms. We present results obtained on real videos from different activity domains.*

## 1. Introduction

Scene understanding corresponds to the real time process of perceiving, analysing and elaborating an interpretation of a 3D dynamic scene observed through a network of sensors (including cameras and microphones). This process consists usually in matching signal information coming from sensors with a large variety of models which humans are using to understand the scene in the form of activities carried out by people observed in the scene.

Despite some success stories, scene understanding systems remain limited and usually can function only under restrictive conditions (those defined by the models which humans have defined for specific activities of interest). Here, the real challenge is to discover the activity patterns (called also actions, situations,behaviours, scenarios, scripts and chronicles) by exploring efficiently spatial and temporal information of detected objects. Such automatically learned activity patterns allow uderstanding the scene.

Furthermore, here we state that learning meaningful activity patterns rely on the maintenance and update of the activity throughout time. However, not all learning/modelling techniques are well adapted to perform on-line adaptation. On-line learning is indeed an important capability required to perform scene analysis on long-term basis. In this work we aim at designing an unsupervised system for the extraction of structured knowledge from large video recordings that can perform online. We show how meaningful scene activity characterisation can be achieved through trajectory analysis. By employing clustering techniques, we define the context (activity zones) characterising the scene dynamics. Employing learned zones we extract people activities by relating mobile trajectories to the learned zones. The activity of a person can then be summarised as the series of zones that the person has visited. For the analysis of the trajectory, a multiresolution analysis is set such that a trajectory is segmented into a series of tracklets based on changing speed points. Starting and ending tracklet points are then fed to a simple yet advantageous incremental clustering algorithm to create an initial partition of the scene and find an initial set of activity zones. Similarity relations between resulting clusters are modeled employing fuzzy relations. These can then be aggregated with typical soft-computing algebra. A clustering algorithm based on the transitive closure calculation of the fuzzy relations allows building the final structure of the scene. To allow for incremental learning and update of activity zones (and people activities) fuzzy relations are defined with online learning terms. We present results ob-

tained on real videos from different activity domains.

The remainder of this paper is structured as follows. We give first a short overview of the related work (section 2). In section 3 we present an architectural overview of our approach. The detailed methodology for activity extraction is presented in section 4. How we evaluate the proposed approach is explained in section 5. Finally, Section 6 draws the main conclusions and describes some perspectives.

## 2. Related Work

Extraction of the activities contained in the video by applying data-mining techniques represents an emerging field in computer vision. Recently it has been shown that trajectory analysis from mobile objects detected in videos can give meaningful activity information. Trajectory analysis has become a popular approach due to its effectiveness in detecting normal/abnormal behaviours. Several approaches proposed on trajectory clustering include splitting algorithms [12], neural trees [5], Hidden Markov Models (HMM) [3, 10, 13]. Transforming trajectory points to another space employing for instance PCA [1], ICA [2] or DFT [1] has also been employed. The drawback in these approaches is that an activity pattern is basically represented as a single route along which the objects travel repeatedly. Thus the only meaningful information that can be extracted is what are the most common routes. More importantly, no semantic information can be drawn which facilitates activity understanding for end-users. Thus complex activities including changing of route, partially stopping (for interaction with other mobiles or with static objects of the scene) can hardly be captured and meaningfully described to an end-user with these methods.

In order to add semantic information to trajectory clusters, recent approaches describe trajectory information in terms of mobile origin-destination such as Patino et al. [11],Sankaranarayanan et al. [14] where only beginning/ending points of the trajectory are analysed. Alternatively beginning/ending trajectory points serve to learn the entry/exit zones (also called sink and sources) of the scene [17, 8] or any general activity zone [11]. Zhou et al. [18] on the contrary first manually define rough borders of sink and sources to guide the activity clustering employing Random Field Topic models but the learned activities are still single routes. Recently, Hospedales et al. [7] have also proposed Markov Clustering Topic models for more complex activity modeling but the learning of activity zones for better semantic interpretation has not been adressed.

In addition to activity clustering, in order to enable dynamic adaptation to unexpected or newly observed data, we need a system able to learn the activity clusters on line. On-line learning is indeed an important capability required to perform behaviour analysis on long-term basis and to anticipate the human interaction evolutions. An on-line learning algorithm gives a system the ability to incrementally learn new information from datasets that consecutively become available, even if the new data introduce additional classes that were not formerly seen. The previous approaches have not demonstrated to work with on-line learning capabilities. Indeed, many popular classifiers are not structurally suitable for incremental learning; either because they are stable [such as the multilayer perceptron (MLP), radial basis function (RBF) networks, or support vector machines (SVM)], or because they have high plasticity and cannot retain previously acquired knowledge, without having access to old data (such as -nearest neighbor) [9]. Specific algorithms that have been developed to perform on-line incremental learning include the Leader algorithm [6], Adaptive Resonance Theory modules (ARTMAP) [4], Evolved Incremental Learning for Neural Networks [15], leaders-subleaders [16].

Our contribution in this work to the state of the art is thus a trajectory analysis-based algorithm capable to extract complex activities with semantic description capable to perform online. We achieve this by automatically learning the context of the scene (activity zones), then relating mobile trajectories to the learned zones, such that, the activity of a person can be summarised as the series of zones that the person has visited. To guarantee that the algorithm works in on-line way, we have set a memory factor allowing retaining previously acquired knowledge, without having access to old data and yet be able to learn from new incoming data.

## 3. General overview of the activity extraction system

Our proposed system works off-line and is composed of four modules applied sequentially. **Tracklet calculation process**: We study the trajectory speed variations in order to distinguish between the mobile in a stationary state (stop points) or moving state. **Point clustering**: We aim here to automatically discover unknown activity zones. Stop points obtained from the previous tracklet calculation process are clustered together. The different areas occupied by disjoint groups of stop points form the new discovered activity zones. **Zone merging**:It is possible that some discovered activity zones are partially overlapping. When this is the case, both regions are most certainly part of a bigger activity zone. Our algorithm will attempt to merge those overlapping activity zones. **Event discovery and Activity extraction**: One displacement through two different zones is marked as a simple event in our system. This module will find out all different simple events occurring in the video and from them extract the different activities defined by the series of zones that the person has visited.

## 4. Activity Analysis and Clustering

We understand activity as the interactions occurring between mobile objects themselves and those between mobiles and the environment. We employ trajectory-based analysis of mobiles in the video to discover the points of entry and exit of mobiles appearing in the scene and ultimately deduce the different areas of activity. In a second step, mobile objects are then characterised in relation to the learned activity areas.

### 4.1. Trajectory analysis process: Tracklet calculation

In order to discover meaningful activity, it is of prime importance to have available detailed information allowing to detect the different possible interactions between mobiles. As our system is based on trajectory analysis, the first step to prepare the data for the activity clustering methodology is to extract tracklets of fairly constant speed allowing to characterise the displacements of the mobile or its stationary state.

If the dataset is made up of *N* objects, the trajectory $tr_j$ for object $O_j$ in this dataset is defined as the set of points $[x_j(t), y_j(t)]$ corresponding to their position points; x and y are time series vectors whose length is not equal for all objects as the time they spend in the scene is variable. The instantaneous speed for that mobile at point $[x_j(t), y_j(t)]$ is then $v\left(t\right) = \left(\dot{x}\left(t\right)^2 + \dot{y}\left(t\right)^2\right)^{\frac{1}{2}}$. The objective is then to detect those points of changing speed allowing to segment the trajectory into tracklets of fairly constant speed so that the trajectory can be summarised as a series of displacements at constant speed or in stationary state.

The mobile object time series speed vector is analysed in the frame of a multiresolution analysis of a time series function $v\left(k\right)$ with a smoothing function, $\rho_{2^s}\left(k\right) = \rho\left(2^s k\right)$, to be dilated at different scales $s$. In this frame, the approximation $A$ of $v\left(k\right)$ by $\rho$ is such that $A_{2^{s-1}}v$ is a broader approximation of $A_{2^s}v$. By analyzing the time series $v$ at coarse resolutions, it is possible to smooth out small details and select those points associated with important changes.

The speed change points are then employed to segment the original trajectory $tr_j$ into a series of $i$ tracklets $tk$. Each tracklet is defined by two key points, these are the beginning and the end of the tracklet, $[x_j^i(1), y_j^i(1)]$ and $[x_j^i(end), y_j^i(end)]$. By globally reindexing all tracklets, let $m$ be the number of total tracklets extracted, we obtain the following tracklet feature vector :

$$tk_m = [x_m(1), y_m(1), x_m(end), y_m(end)] \qquad (1)$$

### 4.2. Context analysis process

Modeling the spatial context of the scene is essential for recognition and interpretation of activity. Because it is not possible to define a-priori all activity zones, manually-defined contextual zones do not suffice to describe all possible situations or evolving actions in the monitored scene. We thus propose thus to learn the activity zones.

In our approach, we propose to find first high resoluted (small in size) entry/exit activity zones, which in a second step could be further regrouped to obtain broader activity zones (following an hierarchical agglomerative algorithm).

#### 4.2.1 Clustering of tracklet entry/exit points

For this step, we employ the well-known clustering Leader algorithm [6]. It has the advantage to work on-line without needing to specify the number of clusters in advance. In this method, it is assumed that a threshold *T* is given. The algorithm constructs a partition of the input space (defining a set of clusters) and a leading representative for each cluster, so that every object in a cluster is within a distance *T* of the leading object. The threshold *T* is thus a measure of the diameter of each cluster. The algorithm makes one pass through the dataset, assigning each object to the cluster whose leader is the closest and making a new cluster, and a new leader, for objects that are not close enough to any existing leaders.

Let us consider the position of a mobile is *L(x,y)*, its influential zone, *Zn*, is defined by a radial basis function (RBF) centered at the position *L*; and the belonginess of a new point p(x,y) to that zone is given by:

$$Zn(L,p) = \phi(L,p) = exp(-\|p - L\|^2 T^2) \qquad (2)$$

The RBF function has a maximum of 1 when its input is $p = L$ and thus acts as a similarity detector with decreasing values outputted whenever *p* strides away from $L$. An object element will be included into a cluster $Zn$ if $Zn(L,p) \geq 0.5$; the cluster receptive field (hyper-sphere) is controlled by the learnt parameter *T*. In this work we have employed the threshold setting value suggested by [11].

#### 4.2.2 Merging tracklet zones

We find the final activity areas by merging similar entry/exit tracklet zones. We look to establish a similarity relation between the different zones defined by the tracklets. At the end, new zones are given by the fulfillment of different relations. The first relation indicates if zone $Zn_i$ overlaps zone $Zn_j$. This relation is defined as follows:

$R1_{ij}$: Zone $Zn_i$ overlaps Zone $Zn_j$

$$R1_{ij} = \sum_{k=1}^{3} \left[ \sum_{p(x,y)\in(X_{ik},Y_{ik})} Zn_j(L_j, p(x,y)) \right] \quad (3)$$

and $(X_{ik}, Y_{ik}) = \left\{ \frac{(k+1)}{3} T \cos(\theta) + L_i \right\}$ with $\theta = 0, ..., \frac{\pi}{8}, ..., 2\pi$

That is, points $(x, y) \in (X_{ik}, Y_{ik})$ belonging to $Zn_i$ centered at $L_i$ are tested to verify the overlap/similarity between $Zn_i$ and $Zn_j$.

In order to make the algorithm work with on-line learning capabilities, capable of retaining the previously acquired knowledge but able to learn from new data we have introduced a memory factor, $beta$, as follows:

$$R1_{ij}^{t} = (1-\beta)\, R1_{ij} + \beta R_{ij}^{t-1} \quad (4)$$

where $R1_{ij}^{t}$ is the current relationship value and $R_{ij}^{t-1}$ is the relationship value from a previous analysed dataset.

Similar relations that we have introduced are the following

$R2_{ij}^{t}$: zone $Zn_i$ and zone $Zn_j$ are destination zones for mobiles departing from the the same activity zone $Zn_k$

$R3_{ij}^{t}$: zone $Zn_i$ and zone $Zn_j$ are origin zones for mobiles arriving at the the same activity zone $Zn_k$

$R4_{ij}^{t}$: zone $Zn_i$ and zone $Zn_j$ have about the same number of detected mobiles stopping at the zone

$R5_{ij}^{t}$: zone $Zn_i$ and zone $Zn_j$ have about the same mobile interaction time. The mobile interaction time is the mean time a mobile spends in that zone.

All relations can be aggregated employing a soft computing aggregation operator such as

$R = R = R1^t \cap R2^t \cap R3^t \cap R4^t \cap R5^t = \max(0, R1^t + R2^t + R3^t + R4^t + R5^t - 4)$ and made transitive with:

$$R \circ R(x,y) = \max_{z} \min\left[R(x,z), R(z,y)\right] \quad (5)$$

$R$ is then a transitive similarity relation with $R$ indicating the strength of the similarity. If we define a discrimination level $\alpha$ in the closed interval [0,1], an $\alpha-cut$ can be defined such that

$$R^{\alpha}(x,y) = 1 \Leftrightarrow R(x,y) \geqslant \alpha \quad (6)$$

It is thus implicit that $\alpha_1 > \alpha_2 \Leftrightarrow R^{\alpha_1} \subset R^{\alpha_2}$ ; thus, the $R^{\alpha}$ form a nested sequence of equivalence relations, or from the classification point of view, $R^{\alpha}$ induces a partition $\pi^{\alpha} = \{Zni^{\alpha}\}$ such that $\alpha_1 > \alpha_2$ implies $\pi^{\alpha_1}$ is a refinement of $\pi^{\alpha_2}$; that is, learned zones at $\pi^{\alpha_1}$ are a refinement of learned zones at $\pi^{\alpha_2}$.

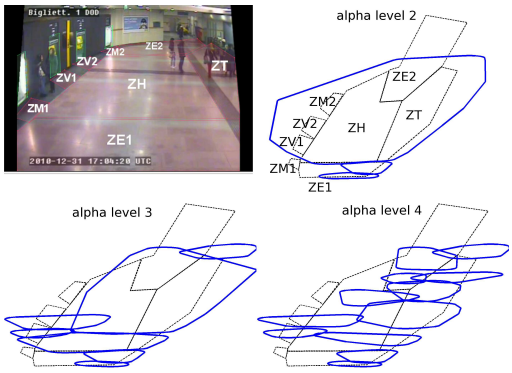


Figure 1. Left top panel: Original underground scene observed by the camera with user-defined areas delimitating the scene. Remaining panels: Learned zones in a 3D top view. They correspond to activity areas as discovered with our algorithm. Different granularity levels allow understanding the activity with different resolutions

### 4.3. Event and Activity extraction process

We aim at creating a system for the recognition and interpretation of human activity and behaviour, and extract new information of interest for end-users. Low-level tracking information is thus expected to be transformed into high-level semantic descriptions conveying useful and novel information. In our application, we transform low-level tracking information into behaviour knowledge by extracting events and activities that give account of mobile displacements through the scene. We do this thanks to the series of tracklets summarising a mobile trajectoty, and thanks to the learned zones and any user-defined contextual zone. Let us assume,

we have in total $k = 1, ..., K$ zones, $Zn_k$, on the scene after the zone learning procedure.

More precisely, we define:

**a simple Event** is the information an object is transferring from one zone $Zn_k$ to a second zone $Zn_{k'}$

**an Activity** is the combination of all simple events associated to the mobile object

Two different kinds of simple events can then be identified:

- Mobile moving from Zone $Zn_k$ to Zone $Zn_{k'}$
- Mobile Inside Zone $Zn_k$

## 5. Experimental Results and Evaluation

To test our approach, we have worked on the extraction of activities from the video recorded at one entrance hall in the Torino (Italy) underground system. The algorithm

for unsupervised learning of activity zones was applied first to a one hour-duration video from the Torino underground system. The final relation $R$ given in equation ( 5), which verifies the transitive closure, is thresholded for different $\alpha - cut$ values going from 0 to 0.9 and with a step value of 0.05. The system gives the possibility to the end-user to select the partition containing the number of activity areas which may suit best his needs. Figure 1 presents those learned zones corresponding to the analysed video. Four different activity levels are calculated from the system. The first level groups all the activity in the scene and thus gives the information as a single activity area where individuals are moving around in the station. The second level gives the information of activity occurring in broad areas. The fourth level gives activity information with smaller and more detailed areas. The third level is a compromise between levels two and four. It can be observed that at the different granularities, those areas being the most employed are: The entry/exit areas to the station, the turnstiles area, and the areas corresponding to the vending machines.

To test the on-line learning capabilities we have analysed four different week days (2 hours/day) from the Torino underground system. We have also worked on the publicly available Edinburgh Informatics Forum Pedestrian Database (http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING/) where we have processed on-line six full working days. Figure 2 shows how most employed activity zones stabilise with the processing of long-term data and the discovered zones are in agreement with the manually-defined zones of interest.

As mentioned in section 4.3, we achieve behaviour carachterisation by linking low-level tracking to the learned zones. The whole activity observed from the scene can then be reported following the behaviours inferred from the learned zones. Some examples of such activities are: at Zone 8 (Stays at VM1), Zone 1 to Zone 8 (ZE1 to VM1), Zone 4 to Zone 15 (Turnstiles to ZE2). The last two shown in figure 3.

When arranging the occurring events according to their semantic, and looking at their frequency of occurrence , it is possible to observe that the most common activities carried out in the station hall are people detected at the north station entry/exit, and people detected at the turnstiles area; the flow of people between these two areas (representing already 50% of the whole station activity). On the other hand, rare activities are for instance those of people going through the station from one entry/exit to the opposite entry/exit or going from one vending machine to the other one (we have depicted these trajectories in figure 4). To test the validity of the activity extraction we analysed one hour video where each trajectory has an annotated activity according to user defined ground-truth zones. When comparing with our discovered activities we obtain the following results: TP:26, FP:3, FN:1, Precision:0.89, Sensitivity:0.96.

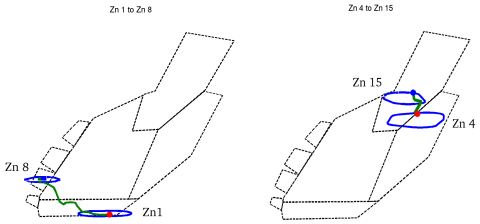


Figure 3. Example of two common activities in the Torino underground.

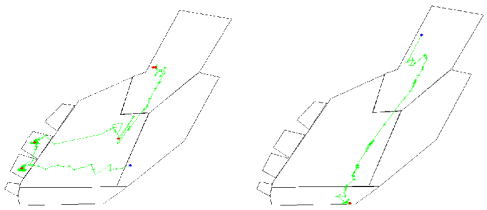

Figure 4. Example of two abnormal trajectories in the Torino underground. Left panel: changing vending machine, right panel: going through the station.

## 6. Conclusions

In this work we have presented an off-line system for human activities extraction with on-line learning (incremental learning) capabilities. So far we have set up a system which starts in a first step by the unsupervised learning of the main activity areas of the scene. In a second step, mobile objects are then characterised in relation to the learned activity areas: either as 'staying in a given activity zone' or 'transferring from an activity zone to another' or a sequence of the previous two behaviours if the tracking persists long enough. This characterisation has allowed us to obtain already informative statistics on the use of the station, and discover what are the main activity areas and main displacements of people in the station. Applying on-line learning, the system is able to continuously process long-term video recordings. From the analysis of different consecutive days, we have shown in this paper how activity zones can be learnt incrementally and refined. Two different application domains have been explored in this regard. Our current evaluation signals encouraging results. Our future work consists in a more complex characterisation of the activities by including more features other than spatial activity zones.

## Acknowledgements

This work was partially funded by the EU FP7 project VANAHEIM with grant no. 248907.

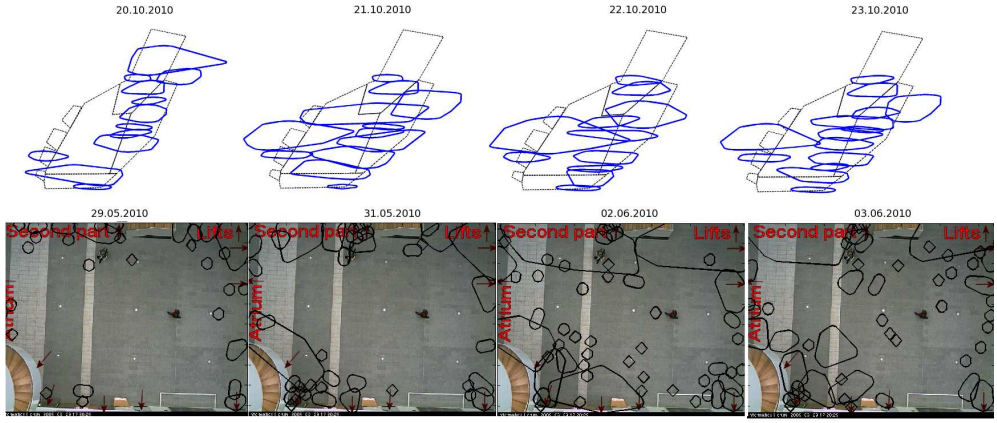


Figure 2. Two examples of on-line learning with the proposed approach. Activity zones are updated as new data is available. Top panel: user-defined ground-truth zones are given in dotted lines. Bottom panel: ground-truth entry/exit zones are given with arrows.

## References


[1] N. Anjum and A. Cavallaro. Single camera calibration for trajectory-based behavior analysis. In *AVSS 2007, IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 147–152, 2007. 2

[2] G. Antonini and J. Thiran. Counting Pedestrians in Video Sequences Using Trajectory Clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 16:1008–1020, 2006. 2

[3] F. Bashir, A. Khokhar, and D. Schonfeld. Object Trajectory-Based Activity Classification and Recognition Using Hidden Markov Models. *IEEE Transactions on Image Processing*, 16:1912–1919, 2007. 2

[4] G. A. Carpenter, S. Grossberg, and J. Reynolds. ARTMAP: supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, 4:565–588, 1991. 2

[5] G. Foresti, C. Micheloni, and L. Snidaro. Event classification for automatic visual-based surveillance of parking lots. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 314–317. IEEE, 2004. 2

[6] J. A. Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., New York, 1975. 2, 3

[7] T. M. Hospedales, S. Gong, and T. Xiang. Video behaviour mining using a dynamic topic model. *International Journal of Computer Vision*, 98(3):303–323, 2012. 2

[8] B. Morris and M. Trivedi. Learning and classification of trajectories in dynamic scenes: A general framework for live video analysis. In *Advanced Video and Signal Based Surveillance, 2008. AVSS '08. IEEE Fifth International Conference on*, pages 154 –161, sept. 2008. 2

[9] M. D. Muhlbaier, A. Topalis, and R. Polikar. Learn++.NC: Combining Ensemble of Classifiers With Dynamically Weighted Consult-and-Vote for Efficient Incremental Learning of New Classes. *IEEE transactions on neural networks*, 20:152–168, 2009. 2

[10] N. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:831–843, 2000. 2

[11] J. L. Patino Vilchis, F. Bremond, M. Evans, A. Shahrokni, and J. Ferryman. Video Activity Extraction and Reporting with Incremental Unsupervised Learning. In *7th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, Boston, États-Unis, Aug. 2010. 2, 3

[12] C. Piciarelli, G. Foresti, and L. Snidaro. Trajectory clustering and its applications for video surveillance. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2005.*, volume 18, pages 40–45. IEEE, 2005. 2

[13] F. Porikli. Learning object trajectory patterns by spectral clustering. In *2004 IEEE International Conference on Multimedia and Expo (ICME)*, volume 2, pages 1171–1174. IEEE, 2004. 2

[14] K. Sankaranarayanan and J. Davis. Learning directed intention-driven activities using co-clustering. In *7th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 400 –407, 2010. 2

[15] T. Seipone and J. Bullinaria. Evolving Neural Networks That Suffer Minimal Catastrophic Forgetting. In *Proceedings of the Ninth Neural Computation and Psychology Workshop*, pages 385–390, Singapore, 2005. 2

[16] P. Vijaya. Leaders Subleaders: An efficient hierarchical clustering algorithm for large data sets. *Pattern Recognition Letters*, 25:505–513, März 2004. 2

[17] X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *ECCV (3)*, pages 110–123, 2006. 2

[18] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3441–3448, Washington, DC, USA, 2011. IEEE Computer Society. 2