# Activity discovery from video employing soft computing relations

Jose Luis Patino Vilchis, François Bremond, Monique Thonnat

# Activity discovery from video employing soft computing relations

Luis Patino, Francois Bremond and Monique Thonnat

*Abstract*— **The present work presents a novel approach for activity extraction and knowledge discovery from video. Spatial and temporal properties from detected mobile objects are modeled employing fuzzy relations. These can then be aggregated employing typical soft-computing algebra. A clustering algorithm based on the transitive closure calculation of the fuzzy relations allows finding spatio-temporal patterns of activity. 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 the scene. These areas can be reported as activity maps with different granularities thanks to the analysis of the transitive closure matrix of the mobile fuzzy spatial relations. Discovered activity zones and spatio-temporal patterns of activity can be labeled in a human-like language. We present results obtained on real videos corresponding to apron monitoring in the Toulouse airport in France.**

## I. Introduction

Current video monitoring systems aiming at safety and security issues, on large infrastructure sites, are able to record huge amounts of data and store this in standard databases. Most vision systems specialize on recognizing predefined events (or behaviours), mostly with the aim of raising an alarm. However, the raw video data stored in the database may represent a rich source of new information still unknown to human operators. The complexity related to manipulate such large amounts of data makes impossible for the user to achieve a methodological and systematic exploration of the database. Knowledge discovery systems (KDS) aim at helping the human operator on this aspect. Although knowledge discovery systems have become a central part on many domains where data is stored in a database, little research has been done in the field of video data-mining to discover the behaviours stored on large video recordings and give a comprehensive analysis of the ongoing activity. It must be said the challenge is particularly difficult not only because of the difficulty in identifying the interesting patterns of activity in the video but also in describing them in a concise and meaningful manner. Both tasks are crucial in the field of data mining and knowledge discovery, and particularly important for activity discovery from video. Soft computing methodologies are particularly adapted for these tasks because they provide capability to process uncertain or vague information, as well as a more natural framework to cope with linguistic terms and produce natural language-like interpretable results. Fuzzy sets [1] are the ground stone of soft computing, which has led to a wide acceptance of soft computing together with other techniques such as neural networks and genetic algorithms. The relation between different existing fuzzy sets can be graded with the use of fuzzy relations [2]. Various fuzzy-based soft computing systems have been developed for different applied fields of data mining; surprisingly only a few systems employ soft computing techniques to characterize video activity patterns [3], [4]; the use of fuzzy relations for their discovery from video, as presented in this paper, is to our knowledge, a premier. The remaining of the paper is structured as follows. Next section gives a short overview of related work from the literature, then in section III we present our approach. In section IV we give a generic explanation on how mobile objects are first detected and tracked. We explain the procedure we perform on trajectories obtained from the tracking of mobile objects in section V and ultimately the activity extraction process is presented in section VI. The experimental results are given in section VII while our main conclusions and future work are presented in section VIII.

Authors are with Institut National de Recherche en Informatique et en Automatique / Centre de recherche Sophia Antipolis - Méditerranée, 2004 route des Lucioles - 06902 Sophia Antipolis Cedex , France (phone: +33 4 92 38 77 77; email: {Jose-Luis.Patino_Vilchis,Francois.Bremond,Monique.Thonnat}@sophia.inria.fr).

## II. Related Work

Extraction of the activity contained in the video by applying data-mining techniques represents a field that has only started to be addressed. Although the general problem of unsupervised learning has been broadly studied in the last couple of decades [5], there are only a few systems which apply them in the domain of behaviour analysis. Because of the complexity to tune parameters or to acquire knowledge, most systems limit themselves to object recognition [6], [7], [8]. For behaviour recognition, three main categories of learning techniques have been investigated.

• The first class of techniques learns the parameters of a video understanding program. These techniques have been widely used in case of event recognition methods based on neural networks [9], naïve Bayesian classifiers [10], [11] and HMMs [12], [13], [14].

• The second class consists in using unsupervised learning techniques to deduce abnormalities from the occurring events [15], [16], [17], [18].

• The third class of methods focuses on learning behaviour based on trajectory analysis. This class is the most popular learning approach due to its effectiveness in detecting normal/abnormal behaviours. For example, Piciarelli et al. [19] employ a splitting algorithm applied on very structured scenes (such as roads) represented as a zone hierarchy. Foresti et al. [9] employ an adaptive neural tree to classify an event occurring on a parking lot (again a highly structured scene) as normal/suspicious/dangerous. Anjum et al. [20] employ PCA to seek for trajectory outliers. Similarly, Naftel et al. [21] first reduce the dimensionality of the

trajectory data employing Discrete Fourier Transform (DFT) coefficients and then apply a self-organizing map (SOM) clustering algorithm to find normal behaviour. Antonini et al. [22] transform the trajectory data employing Independent Component Analysis (ICA), while the final clusters are found employing an agglomerative hierarchical algorithm. Hidden Markov Models (HMM) have also been employed to detect different states of predefined normal behaviour [23], [24], [25]. All these techniques are interesting, but little has been said about the semantic interpretability of the results. Indeed, more than trajectory clusters, we are interested in extracting meaningful activity clusters from the operational point of view; learn not only the main trajectory (i.e. routes) and their characteristics (e.g. speed, proxemic information...) but activities with semantic content, which can be interpreted, and as a complement normal/abnormal behaviour extraction. This work comes thus into the frame of behaviour extraction from trajectory analysis, however we have in addition a higher semantic level that employs proximity relations between resulting clusters of detected mobiles as well as between clusters and contextual elements from the scene to, first, build the structure of the scene and, then, characterise the ongoing different activities of the scene. By including temporal information we are able to find spatio-temporal patterns of activity.

In addition, we need a system able to learn the activity clusters in an on-line way. 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 of incrementally learning new information from datasets that consecutively become available, even if the new data introduce additional classes that were not formerly seen. This kind of algorithm complies with three rules 1) does not require access to previously used datasets, 2) it is capable to retain the previously acquired knowledge and 3) has no problem to accommodate any new classes that are introduced in the new data [26]. Specific algorithms have been developed to perform on-line incremental learning, such as Leader [18], Adaptive Resonance Theory modules (ARTMAP) [27], leaders-subleaders [28], or the BIRCH algorithm [29]. The last two techniques can be considered as an extension of the leader algorithm. However, all these approaches rely on a manually-selected threshold to decide whether the data is too far away from the clusters. In this work we propose also to improve this aspect by controlling the learning rate with coefficients indicating how flexible the cluster can be when updated with new data.

## III. Proposed Approach

### A. System Architecture

Our proposed system is mainly composed of two different processing components (shown in Figure 1). The first one is a real time analysis subsystem for the detection and tracking of objects. This is a processing that goes on a frame-by-frame basis. The second subsystem works off-line and achieves the extraction of activity patterns from the video. This subsystem is composed of two modules: The trajectory analysis module, and the activity analysis module. The former is aimed at obtaining behavioural displacement patterns indicating the origin and destination of mobile objects observed in the scene. We achieve this through trajectory-based analysis of mobiles in the video to discover the points of entry and exit of mobiles appearing in the scene. The latter is aimed at extracting more complex patterns of activity, which include spatial information (coming from the trajectory analysis) and temporal information related to the interactions of mobiles observed in the scene, either between themselves or with contextual elements of the scene. We achieve this by aggregating soft-computing relations defined on the mobile objects. Spatial and temporal properties from detected mobile objects are modeled employing fuzzy relations. These can then be aggregated employing typical soft-computing algebra. A clustering algorithm based on the transitive closure calculation of the fuzzy relations allows finding spatio-temporal patterns of activity.

For the storage of video streams and the trajectories obtained from the on-line video processing, a relational database has been set up. The off-line modules read the trajectories from the database and return the different trajectory types identified; the discovered activities on the video; and resulting statistics calculated from the activities.

Streams of video are acquired at a speed of 10 frames per second. The on-line (real time) analysis subsystem takes its input directly from the data acquisition component; the video is stored in the DB parallel to the real time processing.

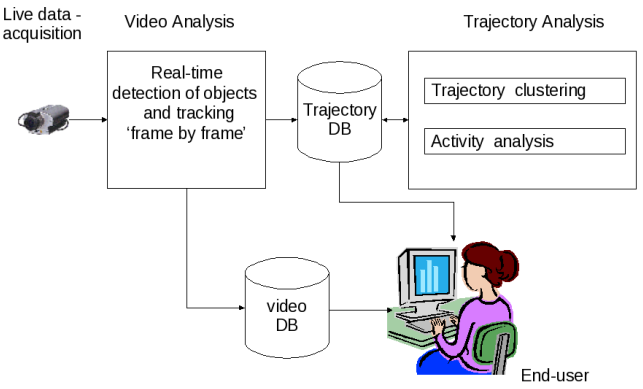


Fig. 1. General architecture.

The off-line analysis subsystem is dedicated to the manager or designer who wants to get global and long-term information from the monitored site. The user can specify a period of time where he/she wishes to retrieve and analyse stored information. In particular the user can access all database and visualize specific events, streams of video and off-line information.

### B. Defining the scene model

Modelling the spatial context of the scene is essential for recognition and interpretation of activity. By contextual

areas we understand those semantic regions of the scene where people activities are expected to be different from one another. Contextual areas in the scene have thus a central role to understand activities as they allow analysing possible interactions between mobile and environmental objects of the scene and thus establish a semantic meaning. In our current application 12 specific areas have been defined for the ground personal in the airport to carry out specific operations. These specific contextual zones, *Zn*, are defined in figure 2. The relevant scenario areas are: Entry/Exit areas and Parking/Servicing areas.

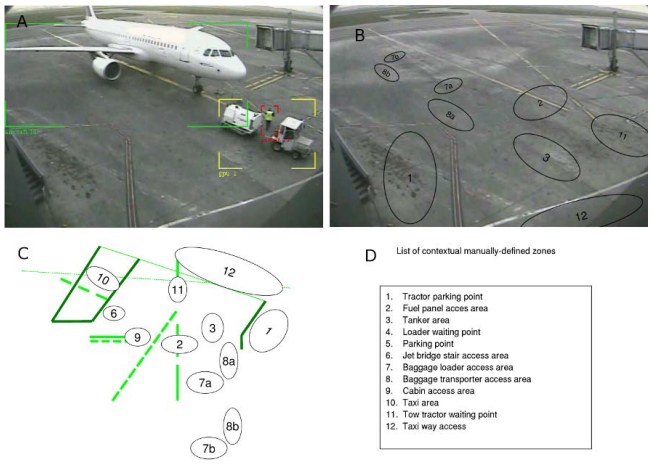


Fig. 2. Contextual zones manually defined: a) monitored apron area with aircraft b) empty apron area with some user-defined contextual areas c) top view of the monitored apron with all user-defined contextual areas d) list of contextual areas.

## IV. Detection and Tracking of Objects

We have developed specific algorithms [30] to detect in real time objects of interest in the video. This is however not the main topic of this paper and therefore only a general description will be given. Object detection is made through a motion analysis algorithm; it segments, from a background reference image,the foreground pixels which belong to a moving object by means of a simple thresholding operation. The foreground pixels are then spatially grouped into moving regions (and enclosed by bounding boxes). These moving regions constitute the detected object . A frame to frame tracker then links the list of detected objects in each pair of successive frames. The output of the frame to frame tracker is a graph of linked detected objects. This graph provides all the possible trajectories that a mobile object may take. After a small period of time, it is possible to build a temporary model of the path a mobile may follow. A path is composed of a temporal sequence of detected objects fulfilling two conditions: 1) detected objects included in the path must be coherent regarding to the 3D size; 2) detected objects included in the path must be coherent regarding to their sequential distance from each other on the ground plane and relative speeds. A path is created when a mobile object is detected. An existing path is updated when a new detection occurs. A path is deleted if the two coherency criteria are not met. The remaining paths constitute the trajectories of the detected mobile objects; each trajectory corresponding to one mobile object.

## V. Trajectory analysis

Consider a dataset made up of $N$ objects, the trajectory 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. Two key points defining these time series are the beginning and the end, [$x_j(1)$, $y_j(1)$] and [$x_j(end)$, $y_j(end)$] as they define where the object is coming from and where it is going to. We build a feature vector from these two points. Additionally, we also include the directional information given as [$cos(\theta_j)$, $sin(\theta_j)$], where $\theta_j$ is the angle which defines the vector joining [$x_j(1)$, $y_j(1)$] and [$x_j(end)$, $y_j(end)$]. A mobile object seen in the scene is thus represented by the feature vector

$$v_j = [x_j(1),\ y_j(1), x_j(end), y_j(end), cos(\theta_j), sin(\theta_j)] \quad (1)$$

To characterise trajectories and to enable dynamic adaptation to newly observed data, we employ an on-line clustering algorithm, the Leader algorithm [31]. Given a distance $D$ between any pair of objects, and a threshold $T$, the algorithm constructs a partition of the input space (defining a set of clusters) and a leading representative for each cluster, such 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 leading object representative associated with cluster $CL_i$ is denoted by $L_i$. 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 to any existing leaders. However, the algorithm is extremely sensitive to the threshold defining the minimum activation of a cluster $CL$. Defining $T$ is application-dependent. It can be supplied by an expert with a deep knowledge of the data or employing heuristics. In this work we propose to learn this parameter employing a training set and Machine learning.

Let each cluster $CL_i$ be defined by a radial basis function (RBF) centred at the position given by its leader $L_i$:

$$CL_i(v) = \Phi\left(L_i, v, T\right) = \exp(-\|v - L_i\|^2 T^2) \quad (2)$$

The RBF function has a maximum of 1 when the difference between its leader $L_i$ and the input $v$ is 0 and thus acts as a similarity detector with decreasing values outputted whenever $v$ strides away from $L_i$. We can make the choice that an object element will be included into a cluster if $CL_i(v) \geqslant 0.5$ , which is a natural choice. The cluster receptive field (hyper-sphere) is controlled by the parameter $T$. Now, consider $C = \{CL_1 \cdots CL_k\}$ is a clustering structure of a data set $X = \{v_1, v_2, ..., v_N\}$; $\{L_1,\ldots,L_k\}$ are the leaders in this clustering structure and P = $\{P_1 \cdot \cdot$

$\cdot P_s\}$ is the 'true' partition of the data (Ground-truth) and $\{M_1,\ldots,M_s\}$ are the main representatives (or Leaders) in the 'true' partition. We can define an error function given by

$$E = \frac{1}{2N}\sum_{j=1}^{N} E_j^2 \tag{3}$$

$$E_j = \hat{\Phi}\left(L(v_j), v_j, T\right) - \Phi\left(M(v_j), v_j, T\right) \tag{4}$$

$L(v_j)$ is the Leader associated to $v_j$ in the clustering structure C. $M(v_j)$ is the Leader associated to $v_j$ in the 'true' partition P. The error gives thus an indication of how many elements are misclassified according to the partition P. Minimising this error is equivalent to refine the clustering structure C or equivalently adjusting the parameter T controlling the cluster receptive field. A straightforward way to adjust T and minimise the error is employing an iterative gradient-descent method:

$$T(t+1) = T(t) - \eta\frac{\partial E(t)}{\partial T} \tag{5}$$

With the purpose of tuning parameter *T*, we have defined a training dataset containing 23 trajectory classes, which we also call trajectory types (see Figure 3). Each trajectory type contains triplets of trajectories.

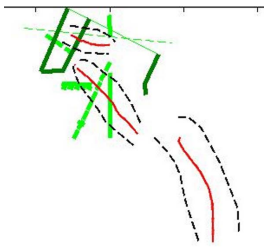

Fig. 3. Ground-truth for different trajectory types. The main trajectory of each trajectory type is represented by a thick line while broken lines represent complementary trajectories.

The proposed gradient-descent methodology was applied to the training dataset. The threshold *T*, in the leader algorithm, is initially set to a large value (which causes a merge of most trajectory types). The error decreases monotonically until obtaining a null error for a value of T=0.7964, which is then selected for our analysis. The set of Leaders defined from this process will also guide the further partition of the incoming data.

## VI. Activity analysis

Our system aims 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 establish a semantic meaning from the scene model presented in section III-B. The behaviour knowledge can be thus expressed with semantic concepts, instead of using quantitative data, thanks to the defined contextual zones. Let us assume we have defined $q$ contextual zones on the scene model. Two different kinds of behaviours can then be identified:

- From Zone $Zctx_q$ to Zone $Zctx_{q'}$
- At Zone $Zctx_q$

In order to cope with the uncertainty aspects, contextual zones are modelled as elliptical shapes with a Gaussian probability density function being associated. Each ellipse,$\varepsilon(a,b,c)$ , is thus defined by its major and minor axis a, b respectively and its centre c. The membership degree that a point p (x, y) can have to a defined zone, Zctx, is then given by

$$Zctx_q(x',y') = \exp\left(-\left(\frac{x'}{\sigma_{a(Zctx_q)}}\right)^2\right)\exp\left(-\left(\frac{y'}{\sigma_{b(Zctx_q)}}\right)^2\right) \tag{6}$$

where (*x*', *y*') is the image point *p*' after projection of *p* into the major and minor axes which define the elliptical zone, *Zn*. That is $p' = A(p-c)$ and A is the rotation matrix defined by the major and minor axis of the ellipse.

The likelihood that the entry/exit points belonging to a trajectory cluster $CL_i$ can be associated with the semantic given by a zone $Zctx_q$ is the mean value of the membership degree of these points to that zone.

### A. Scene model update

Because it is not possible to define a-priori all activity zones, the manually defined Contextual zones do not suffice to describe all possible situations or evolving actions in the monitored scene, but only those matching the previously modelled zones of interest. We thus learn the complementary activity zones from the results obtained on trajectory clustering. We employ the entry/exit (beginning/end) spatial zone of influence $Zcl_i$ of a trajectory cluster $CL_i$.

$$Zcl_i(x,y) = \Phi\left(L_i(1), x, T\right)\Phi\left(L_i(2), y, T\right) \tag{7}$$

Remark that in this case employing $L_i(1)$ and $L_i(2)$ means $Zcl_i(x,y)$ is built from the entry points of trajectory cluster $CL_i$. We then look to establish a similarity relation between the different zones defined by the clusters. Quantitatively the relation is given by measuring the degree of fitness of data belonging to cluster $CL_i$ to the model (Gaussian density function) defined from cluster $CL_j$. On the end, new zones are given by the fulfilment of two relations: cluster $CL_i$ influential zone $Zcl_i$ is similar to cluster $CL_j$ influential zone $Zcl_j$ and cluster $CL_j$ influential zone $Zcl_j$ does not overlap an a-priori defined contextual Zone $Zctx_q$. These relations are defined:

$R1_{ij}$: Zone $Zcl_i$ is similar to Zone $Zcl_j$

$$R1_{ij} = \sum_{k=1}^{3}\left[\sum_{(x,y)\in(X_k,Y_k)} Zcl_j(x,y)\right] \tag{8}$$

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

That is, points belonging to concentric circles to $L_i$ are employed for the similarity comparison between $CL_i$ and $CL_j$. This allows avoiding equity problems with clusters defined on sparse regions (some clusters may be defined with a much larger number of points than others).

$R2_{iq}$: Zone $Zcl_i$ overlaps Zone $Zctx_q$

$$R2_{iq} = \sum_{k=1}^{3} \sum_{p \in (X_i k, Y_i k)} Zn_q(x', y') \tag{9}$$

It is possible to transform R2 into a new relation, R3, which links $CL_i$ and $CL_j$ if both clusters are related to the same Zone $Zctx_q$ through the fulfilment of R2. The relation between $CL_i$ and $CL_j$ is then given by

$$R3_{ij} = \max_q \min\left[R2_{iq}, R2_{qj}\right] \tag{10}$$

Remark that $\overline{R3}$, the complement to $R3$ given by $\overline{R3} = -R3$, represents the relation linking $CL_i$ and $CL_j$ if both clusters are not related to any contextual Zone ($Zctx_q$). $R1$ and $\overline{R3}$ can be aggregated employing a soft computing aggregation operator such as $R = R1 \cap \overline{R3} = \max\left(0, R1 + \overline{R3} - 1\right)$. It is interesting to verify whether the resulting relation is symmetric, $R(x,y) = R(y,x)$, reflexive $R(x,x) = 1$, which make of R a compatibility relation and occurs in most cases when establishing a relationship between binary sets. R is however not a transitive relation. $R(x,y)$ is a transitive relation if $\exists\, z \in X, z \in Y / R(x,y) \geqslant \max_z \min\left[R(x,z), R(z,y)\right]$

If *R* is not transitive, it can be made transitive and furthermore closure transitive following the next steps

Step1. $R' = R \cup (R \circ R)$
Step2. If $R' \neq R$, make$R = R'$ and go to step1
Step3. $R = R'$ Stop. R is the transitive closure where

$$R \circ R(x,y) = \max_z \min\left(R(x,z), R(z,y)\right) \tag{11}$$

R is now 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

$$R^\alpha(x,y) = 1 \Leftrightarrow R(x,y) \geqslant \alpha \tag{12}$$

any relation R can be represented by its resolution form

$$R = \bigcup_\alpha \alpha R^\alpha \tag{13}$$

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 $\pi^\alpha$ of $X \times Y$ (or $X^2$ ) such that $\alpha_1 > \alpha_2$ implies $\pi^{\alpha_1}$ is a refinement of $\pi^{\alpha_2}$.

At this point, the difficulty comes down to select the appropriate $\alpha - cut$ such that $\pi^\alpha$ from $R^\alpha$ represents the best partition of the data. This is still a difficult and open issue that we choose to approach by selecting the alpha-values, which induce a significant change from $\pi^{\alpha_k}$ to $\pi^{\alpha_{k+1}}$.

To monitor those significant partition changes we choose to study the cluster area and number of clusters induced at each partition $\pi^\alpha$

Let $\pi^\alpha = \{C_1^\alpha, ..., C_i^\alpha, ..., C_{n^\alpha}^\alpha\}$ be the partition for the $\alpha - cut$ level. $C_i^\alpha$ is one cluster of the induced partition or in the frame of our application a new discovered zone. Let $A_i^\alpha$ be the area corresponding to $C_i^\alpha$ and which can be calculated from the convex hull enveloping all elements in $C_i^\alpha$.

The mean cluster area for the $\alpha - cut$ level, $A^\alpha = \frac{1}{n_\alpha} \sum_{i=1}^{n_\alpha} A_i^\alpha$ , the range value of the cluster areas, $\max(A_i^\alpha) - \min(A_i^\alpha)$, and number of clusters for that level, $|\pi^\alpha|$ are analysed in the frame of a multiresolution analysis of a time series function $f(k)$ with a smoothing function, $\phi_{2^j}(k) = \phi\left(2^j k\right)$ , to be dilated at different scales j. In this frame, the approximation *S* of *f(k)* by $\phi$ is given by

$$S_{2^j} f = \sum_{u=-\infty}^{\infty} f(k)\, \phi_{2^j}\left(k - 2^{-j} u\right) = \langle f, \phi_{2^j} \rangle \tag{14}$$

such that $S_{2^{j-1}} f$ is a broader approximation of $S_{2^j} f$ . By analysing the time series *f* at coarse resolutions, it is possible to smooth out small details and select the $\alpha - cut$ levels associated with important changes. From the monitored scene, it would be useful to distinguish among different information levels: (i) grouped activity on large spaces, (ii) very detailed individual activity, (iii) somewhere meaningful in-between the last two. For this reason, when performing activity zone discovery, we select the three highest change-inducers $\alpha - cut$ levels from the previous analysis. The result is then that we end up with a three levels hierarchy of activity zones.

### *B. Semantic update*

It is important to observe from equation 13 that our approach allows establishing an inclusion (parent-child) relationship between discovered activity zones; however, no particular semantic information can be drawn. To solve this problem, we rely again on the semantic that can be deduced from the contextual areas of the scene, as we know that this is the link to establish possible interactions between mobile and environmental objects of the scene. To this end we consider two new relations: R4, The comparison of areas between discovered and contextual areas, and R5, The distance relationships between discovered and contextual areas:

$R4_{iq}$ = Zone $Zcl_i$ is similar in area to Zone $Zctx_q$
$R5_{iq}$ = Zone *Zcl(i)* is near to Zone *Zctx(q)*

$$R = R4 \cap R5 = \max(0, R4 + R5 - 1) \tag{15}$$

From R, we know for each discovered zone what is the 'best' contextual zone to refer to. As mentioned before, the zone areas are calculated from the convex hull enveloping either $C_i^\alpha$ or $Zctx_q$, and the distance between zones from the nearest vertex points of each convex hull.

### C. Spatio-temporal decomposition

We now look into adding a temporal component to obtain activity patterns reflecting spatio-temporal similarities. With this aim, and for our current application we define the following relations:
$R6_{ij}$: mobile object *O(i)* enters Zone *Zn(j)*
$R7_{ij}$: mobile object *O(i)* exits Zone *Zn(j)*
and *Zn(j)* is either a known contextual zone *Zctx(j)* or a discovered zone *Zcl(j)* corresponding to a cluster $C_j^\alpha$ of the induced partition $\pi^\alpha$. For each of these cases we can calculate the membership of the mobile to the corresponding area:

$$R6_{ij} = Zn_j(x, y) \tag{16}$$

$R8$: mobile object *O(i)* starts equal to mobile object *O(j)*
$R8 = 1 - abs(start_time(i) - start_time(j))$

$R9$: mobile object *O(i)* starts equal to mobile object *O(j)*
$R9 = 1 - abs(duration_time(i) - duration_time(j))$

Obtaining the patterns of activity is made by aggregating the above spatio-temporal relations. First $R10 = R6 \cup R7 = \min(1, R6 + R7)$ aggregates mobile objects according to spatial zone relations. $R10 \circ R10$ finds the transitive relation between mobiles according to their spatial distribution.

$R = R8 \cup R9 \cup R10$ aggregates temporal similarity relations between mobiles. We calculate again the transitive closure of this new relation. Analogically to section VI-A an $\alpha - cut$ can be defined such that $R^\alpha(x, y) = 1 \Leftrightarrow R(x, y) \geqslant \alpha$ and $R^\alpha$ induces a new partition $\pi^\alpha = \{C_1^\alpha, ..., C_i^\alpha, ..., C_{n^\alpha}^\alpha\}$; each $C_i^\alpha$ represents a discovered spatio-temporal activity pattern.

## VII. Main Results

We have applied our algorithm to five video datasets corresponding to different monitoring instances of an aircraft in the airport docking area (In the following, these video datasets are to be named: cof1, cof2, cof3, cof4 and cof8). This corresponds to about five hours of video analysed, this means the analysis of about 8000 trajectories. The system was first tuned and initialised as described in section V (i.e. employing the defined parameter T and set of initial Leaders learned from the mentioned Training dataset). Figure 4 shows the evolution of the system with on-line learning as the different video sequences are processed.

The scene model is then updated according to the fulfilment of relations $R1$, $\overline{R3}$ given in section VI-A. The final relation $R$, 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 $\alpha - cut$ values defining the different granularities (or information levels) for the scene are then obtained from the multiresolution analysis of the mean cluster area, range value of cluster areas, and number of clusters, which are obtained at each $\alpha - cut$ level. Figure 5 shows how these three parameters change on the sequence 'cof1' depending on the $\alpha - cut$ level. Remember from section VI-A that from these parameters we are looking the three highest change-inducers $\alpha - cut$ levels, which will define three information levels for scene activity reporting. Figure 6 shows the $\pi^\alpha$ partitions corresponding to the selected $\alpha - cut$ levels. The first granularity level is set for $\alpha - cut$=0, which merges all activity outside the user-defined contextual zones and thus creates one single broad new zone of global activity outside contextual zones. The second granularity level corresponds to grouped activity on large spaces and is defined as the lowest $\alpha - cut$ value from those three highest change-inducers $\alpha - cut$ levels. In contrast, the fourth granularity level, which is the most detailed activity corresponds thus to the partition induced by the highest change-inducer $\alpha - cut$ level. The third granularity level, corresponds to a compromise between detailed and large activity description (and is defined by the remaining $\alpha - cut$ level). In this way, the different partitions can be seen as activity maps with different granularity levels.

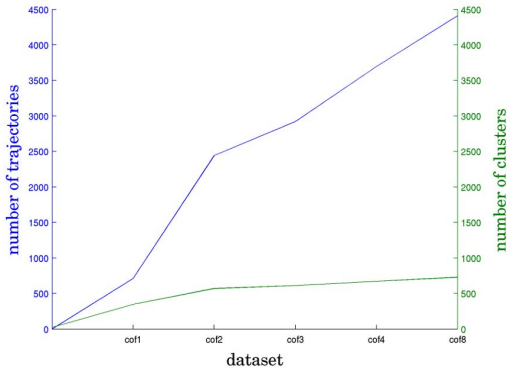


Fig. 4. Number of trajectories processed (blue curve) and number of trajectory clusters created by the on-line system as the different datasets are sequentially treated.

As mentioned in section VI-B, it is important to attach a semantic meaning to each of the new discovered zones. This is achieved, as mentioned before, through fulfilment of relations $R4$ and $R5$ linking the new discovered zones to the user defined contextual zones by their area similarity and their spatial closeness. For instance, for the numerotated zones in Figure 6, the deduced semantics are given in the figure legend.

When introducing temporal information and following the procedure given in section VI-C, it is possible to find for each sequence a series of spatio-temporal clusters. In order to have a better homogeneity of the spatio-temporal activity clusters, we set the $\alpha - cut$ value to 0.9. Because of this precision, the number of spatio-temporal clusters obtained is relatively high: 'cof1'=96 clusters; 'cof2'=117; 'cof3'=93; 'cof4'=126; 'cof8'=118. In order to look for the meaningful

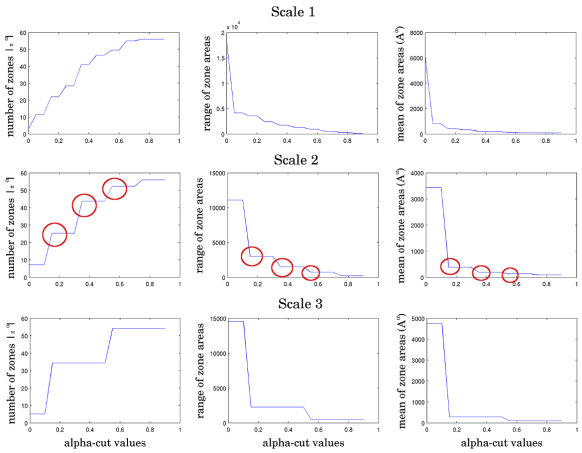


Fig. 5. Change in the number of zones (clusters in $|\pi^\alpha|$), range value of areas, and mean area of zones ( $A^\alpha$), on the sequence 'cof1' for the partition induced by an alpha-cut. The points corresponding to a brisk change of these parameters and selected to define the different granularities (or information levels) for the scene are circled in red.

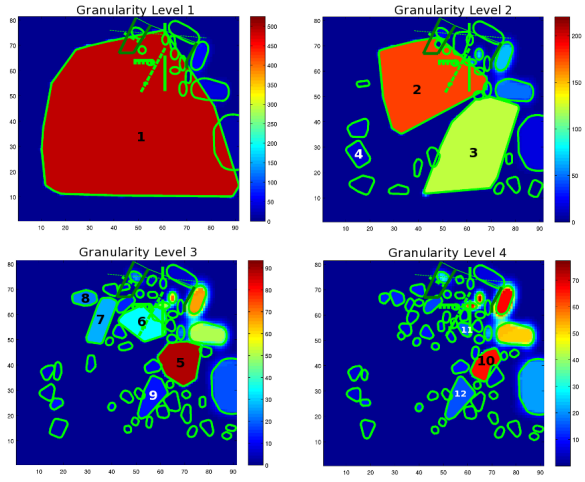


Fig. 6. Activity maps for the sequence 'cof1'. Numbers in zones indicate the most frequently employed discovered zones. **1.** ERA and large surrounding **2.** just west of and inside ERA **3.** just south of and inside ERA **4.** 52 meters away south-west of Cabin access area **5.** 9 meters away south of Rear unload area **6.** 7 meters away south of Cabin access area **7.** 17 meters away south-west of Jet bridge stairs access area **8.** 20 meters away south-west of Taxi parking area **9.** 23 meters away south of Rear unload area **10.** 8 meters away south of Rear unload area **11.** 5 meters away north-west of Rear unload area **12.** 23 meters away south of Rear unload area

common activities we look to correlate the activity clusters over all available video sequences. Briefly speaking, for each couple of activity clusters we take into account the spatial correlation of area occupancy, temporal similarity for the start of the activity, and the correlation of trajectory types involved. Figure 7 shows the time-line of activities with highest correlation, that is, with more reproducibility over the different video sequences.

The sequence 'cof1' contains ground-truth conveying seven activities: 'GPU arrival', the baggage disposal related activities 'unloading' and 'loading',the aircraft related movements 'Aircraft arrival' and 'Aircraft departure' and the Jet

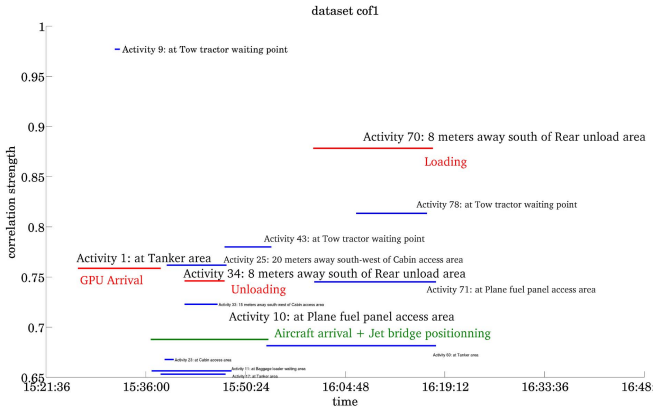


Fig. 7. Spatio-temporal activity patterns discovered in the sequence 'cof1'.

bridge related movements 'Jet bridge positioning', and 'Jet bridge parking'. With our approach we found patterns of activity in 'cof1' corresponding to the first three events and having strong repetitiveness on the other video sequences. These are marked in Figure 7 in red. We found that 'Aircraft arrival' and 'Jet bridge positioning' were merged in one cluster: 'Activity 10: at Plane fuel panel access area' (marked in green in Figure 7). This cluster still contained some more people activity occurring in the same area and because of the transitive spatial and temporal relations included in our approach all these mobiles were merged into one cluster. A similar situation appeared with the remaining events signalled by the ground-truth, although the reproducibility of the resulting cluster is low and is thus less visible.

## VIII. Conclusions

We have presented a fuzzy relation analysis-based approach for unsupervised activity pattern discovery. The approach contains relevant cognitive functions such as perception, learning, dynamic context discovery, recognition, and reasoning, and their integration in a complete artificial cognitive vision system. The proposed approach allows monitoring and processing large periods of time (large amount of data), and thus perform analysis on long-term basis. The system employs a simple, yet advantageous incremental algorithm, the Leader algorithm, to find trajectory clusters from new incoming data. Generally, incremental approaches rely on a manually-selected threshold to decide whether the data is too far away from the clusters. We solve the difficulty of tuning the system by employing a training set and Machine learning.

We employ the information given by trajectory clusters to complement what we know from the scene topology and discover unknown activity zones. This gives a big flexibility to the system contrary to other approaches working with static predefined topologies. Moreover, we study the scene activity at different granularities which give the activity description in broad terms, or with detailed information thus managing different information levels. We elaborate activity maps with a semantic description of the discovered zones (and thus of

the evolving activities) which is closer to a natural language thanks to a series of relationships which allow to describe inferred zones/activities in association to contextual (static) areas of the scene. By adding temporal information such as tracking start time and duration, patterns of activities can be discovered for mobile objects sharing the same spatial and temporal relationships. On this aspect our current results are encouraging as the final patterns of activity are given with coherent spatial and temporal information, which is understandable for the end-user. From the analysed sequences, the first one, 'cof1', contained explicit ground-truth for seven activities. From the discovered activity patterns, we had a perfect match with three of them. The remaining activities were not recognised as a single event because of their spatial and temporal closeness with other mobiles. We will thus be looking in our future work to differentiate between mobiles by including object type information. Our approach has the capability to recognise common repetitive activities but we still need to determine the meaningfulness (or abnormality) of single activity patterns not found recursively on the video analysis. This is also part of our future work, which we plan to address by adding more temporal constraints allowing to better characterize the activity patterns.

## REFERENCES


[1] L. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, pp. 338–353, 1965.

[2] ——, "Similarity relations and fuzzy ordering," *Information sciences*, vol. 3, pp. 159–176, 1971.

[3] A. Doulamis, "A fuzzy video content representation for video summarization and content-based retrieval," *Signal Processing*, vol. 80, no. 6, pp. 1049–1067, juin 2000.

[4] S.-W. Lee and K. Mase, "Activity and Location Recognition Using Wearable Sensors," *IEEE pervasive computing*, vol. 1, no. 03, pp. 24–32, 2002.

[5] Z. Ghahramani, "Unsupervised Learning," *Lecture Notes in Computer Science*, no. 3176, pp. 72–112, 2004.

[6] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," *Lecture Notes in Computer Science*, vol. 3952, p. 428, 2006.

[7] I. Laptev, "Improvements of object detection using boosted histograms," in *British Machine Vision Conference*, vol. 3, 2006, pp. 949–958.

[8] S. Munder and D. Gavrila, "An Experimental Study on Pedestrian Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1863–1868, November 2006.

[9] 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.*, vol. 3. IEEE, 2004, pp. 314–317.

[10] F. Lv, X. Song, B. Wu, V. Singh, and R. Nevatia, "Left luggage detection using bayesian inference," *Proceedings of the 9th IEEE International Workshop on*, 2006.

[11] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, pp. 1778–1792, 2005.

[12] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Hidden markov models for optical flow analysis in crowds," *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1, 2006.

[13] A. Galata, N. Johnson, and D. Hogg, "Learning Variable-Length Markov Models of Behavior," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 398–413.

[14] A. D. Wilson and A. F. Bobick, "Hidden Markov models for modeling and recognizing gesture under variation," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, pp. 123–160, 2001.

[15] G. L. Foresti, G. Giacinto, and F. Roli, "Detecting dangerous Behaviors of Mobile Objects in Parking Areas," *Multisensor Surveillance Systems: The Fusion Perspective*, pp. 199 FG – 0, 2003.

[16] T. Xiang and S. Gong, "Video behaviour profiling and abnormality detection without manual labelling," *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*, vol. 2, 2005.

[17] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted hmms for unusual event detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, vol. 1, 2005.

[18] T. Xiang and S. Gong, "Incremental and adaptive abnormal behaviour detection," *Computer Vision and Image Understanding*, vol. 111, pp. 59–73.

[19] C. Piciarelli, G. Foresti, and L. Snidaro, "Trajectory clustering and its applications for video surveillance," in *Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.* Ieee, 2005, pp. 40–45.

[20] N. Anjum and A. Cavallaro, "Single camera calibration for trajectory-based behavior analysis," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007*. IEEE, 2007, pp. 147–152. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4425301

[21] A. Naftel and S. Khalid, "Classifying spatiotemporal object trajectories using unsupervised learning in the coefficient feature space," *Multimedia Systems*, vol. 12, pp. 227–238, 2006.

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

[23] F. Bashir, A. Khokhar, and D. Schonfeld, "Object Trajectory-Based Activity Classification and Recognition Using Hidden Markov Models," *IEEE Transactions on Image Processing*, vol. 16, pp. 1912–1919, 2007. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4237188

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

[25] F. Porikli, "Learning object trajectory patterns by spectral clustering," in *2004 IEEE International Conference on Multimedia and Expo (ICME)*, vol. 2. IEEE, 2004, pp. 1171–1174.

[26] R. Polikar, L. Upda, S. Upda, and V. Honavar, "Learn++: an incremental learning algorithm for supervised neural networks," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 31, pp. 497–508, 2001.

[27] G. A. Carpenter, S. Grossberg, N. Markuzon, J. Reynolds, and D. B. Rosen, "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps." *IEEE transactions on neural networks*, vol. 3, pp. 698–713, Januar 1992.

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

[29] M. Livny, T. Zhang, and R. Ramakrishnan, "BIRCH: an efficient data clustering method for very large databases," in *ACM SIGMOD international Conference on Management of Data*, vol. 1, Montreal, 1996, pp. 103–114.

[30] A. Avanzi, F. Bremond, C. Tornieri, and M. Thonnat, "Design and Assessment of an Intelligent Activity Monitoring Platform," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, pp. 2359–2374, 2005.

[31] J. A. Hartigan, *Clustering algorithms*. New York: John Wiley & Sons, Inc., 1975.