Extraction of activity patterns on large video recordings

Jose Luis Patino Vilchis, Hamid Benhadda, Etienne Corvee, François Bremond, Monique Thonnat

Extraction of Activity Patterns on large Video Recordings

[1]Luis Patino, [2] Hamid Benhadda, [1]Etienne Corvee, [1]Francois Bremond and [1]Monique Thonnat

[1] INRIA, 2004 route des Lucioles, 06902 Sophia Antipolis (FRANCE)
[2] Thales Comunication, 160 Boulevard de Valmy, 92704 Colombes (FRANCE)

Abstract

Extracting the hidden and useful knowledge embedded within video sequences and thereby discovering relations between the various elements for helping the efficient decision making process is a challenging task. The task of knowledge discovery and information analysis is possible because of recent advancement in object detection and tracking. In this paper we present how video information is processed with the ultimate aim to achieve knowledge discovery of people activity and also extract the relationship between the people and contextual objects in the scene. First, the object of interest and its semantic characteristics are derived in real-time. The semantic information related to the objects is represented in a suitable format for knowledge discovery. Next, two clustering processes are applied to derive the knowledge from the video data. Agglomerative hierarchical clustering is used to find the main trajectory patterns of people and relational analysis clustering is employed to extract the relationship between people, contextual objects, and events. Finally, we evaluate the proposed activity extraction model using real video sequences from underground metro networks (CARETAKER) and a building hall (CAVIAR).

Key words: Behaviour recognition, video surveillance, data mining, temporal event pattern, trajectory clustering

Introduction

Nowadays, more than ever, the technical and scientific progress requires human operators to handle large quantities of data. To treat this huge amount of records, the data-mining field can provide adequate solutions to synthesize, analyze and extract valuable information, which is generally hidden in the raw data. Applying data mining techniques in large video data is now possible mainly because of the advance made in the field of object detection and tracking [1]. Data mining on video data has mainly been employed for annotation/retrieval processes [2-5]. The task consists in mining multiple visual features into categories associated with meaningful semantic keywords that will allow the retrieval of the video. Usually low level features such as colour, texture, shape, and motion information are employed. A recent review in video retrieval can be found in [6]. The structured representation issued from the mining procedure gives a domain-dependent association that tries to solve the well-known gap between low-level features and high-level concepts. Recently, particular attention has been turned to the trajectory information associated to mobile objects observed in the video. On one hand, because trajectory descriptors have been shown to be very useful on their own for video indexing and retrieval [7]. On the other hand because the application of data mining and machine learning techniques to the study of trajectories has started to show its importance for activity understanding. This kind of analysis come as a complement to current video monitoring/surveillance systems such as PRISMATICA [8], VISOR-BASE [9], or ADVISOR [10], which were

rather oriented towards the real-time recognition of events of interest (fighting between persons, vandalism, person jumping above a barrier, group of people blocking an exit and overcrowding situation). Although these systems begin to recognise robustly predefined events in the video, data mining/knowledge discovery on the activities contained in the video has not been addressed.

In this paper we show the procedure to achieve knowledge discovery to obtain meaningful trajectory patterns from video-data. A hierarchical agglomerative clustering algorithm is employed for this purpose. Moreover, we show how semantic meaning can be obtained from these patterns. For instance we can study the dynamics of the people characterised by a given trajectory type. We have proposed an effective knowledge modelling format which allows us to combine the motion pattern of a person or a group of persons with other features describing their interactions with the environment. We show as well in our contribution how we can further apply data-mining techniques on this information. Specifically, we employ the relational analysis [11] for this purpose and show how behavioural patterns of interaction can be extracted. The techniques employed have been evaluated on annotated videos from the CAVIAR project [12] and on large underground video recordings (GTT metro, Torino, Italy and ATAC metro, Roma, Italy) from the CARETAKER project (www.ist-caretaker.org). These videos are associated with manually generated ground-truth.

This research has been done in the framework of the CARETAKER project, which is an European initiative to provide an efficient tool for the management of large multimedia collections. Such system could be used in applications such as

surveillance and safety issues, in urban planning, resource optimization, elderly person monitoring. This work was partially presented in [13]. In this paper we employ an analysis on more features extracted from mobile objects detected in the videos, we present a new evaluation for our trajectory clustering algorithm and we present results from both, Torino and Roma underground sites.

The rest of the paper is structured in the following way. In section 2 we present the overall architecture of the proposed approach. While the video analysis system to track objects and detect events of interest in the video is explained in section 3, the clustering of trajectories from tracked objects is detailed in section 4. The proposed knowledge representation format is presented in section 5. The relational analysis applied on the trajectory and the contextual information is explained in section 6. Results are presented in section 7. The proposed method is assessed in section 8.

2 General structure of the proposed approach

The monitoring system is mainly composed of two different processing components (shown in Figure 1). The first one is an on-line analysis subsystem for the real time detection of objects and events previously defined in an ontology. This is a processing that goes on a frame-by-frame basis. At this level, detected events already contain semantic information describing people behaviour and interactions with the contextual objects of the scene. The second subsystem works off-line and achieves the extraction of activity patterns from the video. This subsystem is composed of three modules: The trajectory analysis module where we perform the clustering of trajectories, the object statistical analysis module where we compute meaninful measures on the object dynamics, and the relational analysis module where we obtain

behavioural patterns of interaction. For the storage of video streams and the metadata obtained after both, on-line video processing and off-line analysis, three different database (DB) exist: raw database (Audio/video streams), on-line analysis database (tracked objects and real-time detected events), off-line analysis database (data mining-based events). In this work we only consider the video analysis although audio data has also been acquired in the project.

Streams of video are acquired at a speed of 5 to 25 frames per second, then directly stored in the raw database. The on-line analysis subsystem takes its input directly from the data acquisition component. Objects and events of interest are detected in real time and tracking results are written in the on-line database at a speed of 5 frames per second. The on-line system trigger alarms for the security operators to take immediate actions.

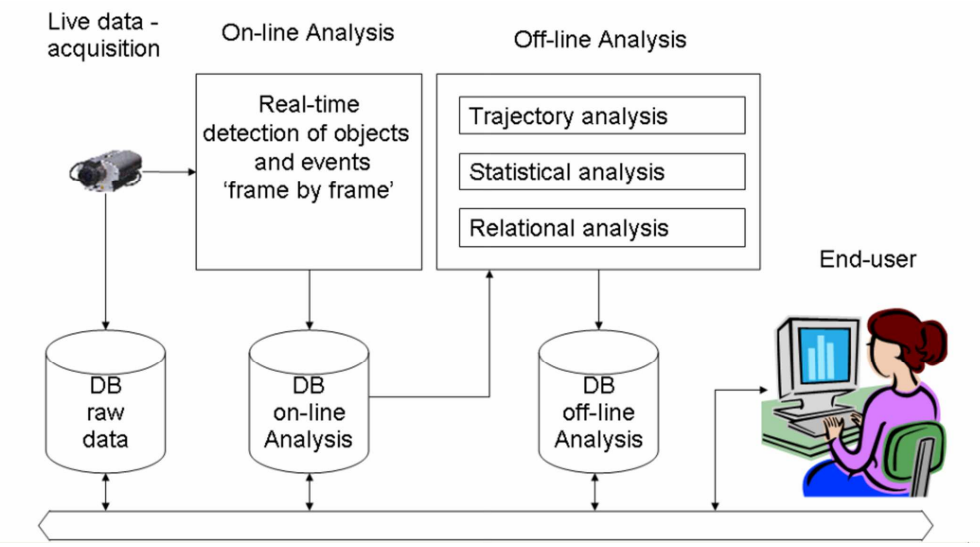


Figure 1. General architecture.

The off-line analysis subsystem takes its input from the on-line database. This 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 databases to visualize specific events, streams of video and off-line information.

3 Real-time Object/Event detection

The first task of our data-mining approach is to detect in real time objects and events of interest. A brief description on how both, objects and events of interest are detected in this work, is given next. It must be noted that our data-mining approach can however be applied with any other detection and tracking technique.

3.1 Object detection

Assuming objects in a scene are individually detected in the image plane; tracking these objects would be a straightforward task to perform and reliable trajectories would be obtained. However, the efficiency of a tracking algorithm directly depends on the quality of the detected objects which is very sensitive to many factors such as the quality of the image (level of compression, accumulated noise throughout the optical device), dynamic occlusions (when several mobile projections onto the image plane overlap) [14], and the complexity of interaction of the objects evolving in a scene as we described below.

Detecting objects in an image is a difficult and challenging task. Several algorithms have been proposed., two recent surveys[1, 15], include the research from the last ten years in this regard. One solution widely employed consists of performing a thresholding operation between the pixel intensities of this frame with the pixel intensities of the background frame. The background image can be a captured image of the same scene having no foreground objects, or no moving objects in front of the camera.

The result of the thresholding operation is a binary mask of foreground pixels. The neighbouring foreground pixels are grouped together to form regions often referred to as 'blobs' which corresponds to the moving regions in the image. If the 'moving objects' projection in the image plane do not overlap with each other, i.e. no 'dynamic occlusion', then each detected moving blob corresponds to a single moving object. However, as soon as occlusion occurs between objects, their moving regions fuse and separating them is not an easy task to do. Adding more informational cues about the detected objects, such as their colour content, shape information or multiple view tracking increases the likelihood in separating the occluded regions. The detailed description of the background subtraction algorithm, which also estimates when the background reference image needs to be updated, can be found in [16].

Having 3D information about the scene under view enables the calibration of the camera. Point correspondences between selected 3D points in the scene and their corresponding point in the 2D image plane allow us to generate the 3D location of any points belonging to moving objects. Thus, the 3D (i.e. width and height) of each

detected moving blob can be measured as well as their 3D location on the ground plane in the scene with respect to a chosen coordinate system. The 3D object information is then compared against several 3D models defined by the user. From this comparison, a detected object is linked to a semantic class. For example, we have chosen a human being to be of average size: 170 cm in height and 60 cm in width. Smaller sizes are chosen for luggages, and bigger sizes for group of persons and very large sizes for crowds. The description and use of these kind of 3D models can be found in [17]

The noisy detected objects, associated with noisy 3D model, are classified according to the closest model they belong to.

3.2 Object Tracking

Detected and classified 3D objects evolving in a scene can be tracked within the scope of the camera using the 3D information of their locations on the ground as well as their 3D dimensions. Tracking a few objects in a scene can be easy as far as they do not interact heavily in front of the camera: i.e. occlusion is rare and short. However, the complexity of tracking several mobile objects becomes a non-trivial and very difficult task to achieve when several objects' projected images overlap with each other on the image plane. Occluded objects have missing 3D locations, which create incoherency in the temporal evolution of their 3D locations.

Our tracking algorithm [18-19] builds a temporal graph of connected objects over time to cope with the problems encountered during tracking. The detected objects are connected between each pair of successive frames by a frame to frame (F2F) tracker. Links between objects are associated with a weight (i.e. a matching likelihood)

computed from three criteria: the similitude between their semantic classes, 3Ddimensions, and differences on 3D distance in the ground plane.

The graph of linked objects provided by the F2F tracker is then analysed by the tracking algorithm, also referred to as the Long Term tracker, which builds paths of each mobiles according to the link features. The best path is then taken out as the trajectory of the related mobiles.

3.3 Event detection

Events of interest for the user are created by the user himself according to a specific semantic language introduced by Vu et al. [20]. This language allows the user to use a designed ontology to detect from simple to complex events. The ontology is the set of all concepts relative to video events and of all the relations between concepts. There are two main types of concepts to be represented: physical objects (including physical objects of interest to be observed in a scene, also called as 'mobile objects' and 'contextual objects', which are defined by the user) and video events occurring in this scene related to the objects of interest.

A physical object of interest '*o*', is a physical object evolving in the scene whose semantic class (i.e. person, group, crowd and luggage) is predefined by end-users and whose motion cannot be foreseen using a priori information. The tracked object is characterised by 2D and 3D features (e.g. a 3D location, width and height), a trajectory and an identifier. Using the 2D and 3D features the object classification algorithm compares the object attributes with the predefined semantic classes (i.e. person, group, crowd and luggage) and assigns the corresponding semantic label to the tracked object. A contextual object is a physical object attached to the scene,

usually an equipment '*eq*', or a zone of interest '*z*'. The contextual object is usually static and whenever in motion, its motion can be foreseen using a priori information. For instance, the movements induced by a door or a chair can be foreseen.

A video event describes any event, action or activity happening in the scene and visually observable by cameras. Video events are characterised by the involved objects of interest (as described above), and their starting - ending times. Examples of events are "detection of a person inside a zone", "detection of an abandoned bag", and "a meeting between two people". For instance "abbandonned bag" consists in the detection of a bag with nobody around for a certain time. For event detection we have our self inspired from methods published in PETS2006 workshop [21], and the specific implementation we have in our system can be found in [10].

We distinguish four types of video events i.e. 'primitive state', 'composite state', 'primitive event' and 'composite event' which are classified into two categories i.e. 'state' and 'event' defined as follows:

+ A 'state' is a spatio-temporal property of a physical object valid at a given instant or constant on a time interval. A state characterises one or several physical objects of interest (e.g. person, crowd or vehicle) with or without respect to other physical objects.

+ A 'primitive state' is a state which is directly inferred from visual attributes of physical objects computed by perceptual components. Usually, visual attributes have a numerical value and can correspond to general physical object properties for most of video understanding applications. For example: "*A* is inside a zone".

+ A ‘composite state’ is a combination of states. We call ‘state components’ all the sub-states composing the state and we call ‘constraints’ all the relations involving its components and its physical objects. For example: “Person p_1 is close to machine m and person p_2 stays inside zone z”.

+ An ‘event’ is one or several change(s) of state values at two successive time instants or on a time interval.

+ A ‘primitive event’ is a change of primitive state values. Primitive events are more abstract than states but they represent the finest granularity of events. For example: “Person p moves from zone z_1 to zone z_2”.

+ A ‘composite event’ is a combination of states and events. Usually, most abstract composite events have a symbolical/Boolean value and are directly linked to the goals of the given application. We call ‘event components’ all the sub-states/events composing the event and we call ‘constraints’ all the relations involving its components and its physical objects.

In the applications of the work presented in this paper, the following events have been defined:

- inside_zone(*o*, *z*)’: when an object ‘*o*’ is in the zone ‘*z*’.
- ‘stays_inside_zone(*o*, *z*, T_1)’: when the event ‘inside_ zone(*o*,*z*)’ is being detected successively for at least T1 seconds
- ‘close_to(*o*, *eq*, *D*)’: when the 3D distance of an object location on the ground plane is less than the maximum distance allowed, D, from an equipment object ‘eq’

- 'stays_at(*o*, *eq*, *D*, T_2)': when the event 'close_to(*o*, *eq*, *D*)' is being consecutively detected for at least T2 seconds.

- 'crowding_in_zone(crowd, *z*)': when the event 'stays_inside_zone (crowd, *z*, T_3)' is detected for at least T_3 seconds.

Also, we have employed the following variables:

For mobile objects:

- object *o*={p, g, c, l, t, u} with p=person, g=group, c=crowd l=luggage, t=train, and u=unknown.

For contextual objects:

- zone *z*={platform, validating_zone, vending_zone}

- equipment *eq*={g1, …, g10, vm1, vm2 } where 'g_i' is the ith gate and vm_i is the ith vending machine.

Event thresholds:

- T_1=60 s, D=1.50 m, T_2=5 s, T_3=120 s.

D corresponds to the Euclidean distance between 3D points of people position, given by the contact point of the person with the ground floor, and the 3D equipment localization.

4 Trajectory analysis

The second layer of analysis in our approach is related to the knowledge discovery of higher semantic events from off-line analysis of activity recorded over a period of time that can span, for instance, from minutes to a whole day. Patterns of activity are first extracted from the analysis of trajectories. Then, the knowledge representation format we propose coupled with the statistical analysis provide a rich overview of the activities in the scene. For the analysis of more complex relationships between the objects observed in the scene, we employ the relational analysis clustering technique.

4.1 Trajectory analysis: Background and related work

Data mining/knowledge discovery techniques applied to trajectory data extract patterns hidden on the raw video that are critical to find out relevant information about the motion behaviour of a person (or set of persons) and their interactions with contextual objects of the scene in the video. In this regard, probably the most active research area has been normal/abnormal behaviour detection. Piciardelli et al. [22] employ a splitting algorithm applied on very structured scenes (such as roads) represented as a zone hierarchy. The drawback of this approach is that it is difficult to generalize on other domains where trajectories have less structure inherited by the scene. Anjum et al. [23] employ PCA to reduce the dimensionality of trajectories . They analyse the PCA first two components of each trajectory together with their associated average velocity vector. Mean-shift, with these features, is employed to seek the local modes and generate the clusters. The modes associated to too few data points are considered as outliers. The outlier condition is set as the 5% of the maximum peak in the dataset, but again the drawback of the approach is that the analysis is adapted to highly structured scenes. Similarly, Naftel et al. [24] 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. [25] transform the trajectory data employing Independent Component Analysis (ICA), while the final clusters are found employing an agglomerative hierarchical algorithm. In these approaches it is however delicate to select the number of coefficients that will represent the data after dimensionality reduction. Calderara et al. [26] employ a k-medoids clustering algorithms on a transformed space modelling different possible trajectory directions to find groups of normal behaviour. Abnormal behaviour is detected as a trajectory that does not fit into the established groups however the approach is validated with acted abnormal trajectory.

Data mining of trajectories has also been applied with statistical methods. Gaffney et al. [27] employed mixtures of regression models to cluster hand movements, although the trajectories were constrained to have the same length. Hidden Markov Models (HMM) have also been employed [28-30]. However, it has been observed that the structures and probability distributions of this kind of approaches are highly domain dependent and require a tedious stage of parameter tuning [23].

All these techniques are interesting, but little has been said about the adequacy of the trajectory clusters and end user expectations.

More than trajectory clusters, in this paper we are interested with extracting meaningful activity clusters, which differs also from normal abnormal behaviour extraction and where clustering techniques have also been employed, not only on trajectory data but also on event data [31-34]. Thus we show first how it is possible to obtain meaningful trajectory patterns from video-data by selecting a large set of features from the trajectory and then employing an agglomerative clustering algorithm. Second this trajectory clustering stage is coupled with statistical analysis

to infer meaningful activities occurring in the scene. Moreover, we complement the trajectory analysis with relational analysis to find out complex activity patterns corresponding to higher semantic relations between variables.

4.2 Trajectory analysis: Proposed method

For the trajectory pattern characterisation of the object, we have selected a comprehensive, compact, and flexible representation. It is suitable also for further analysis as opposed to many video systems. They actually store the sequence of object locations for each frame of the video, which is a cumbersome representation with no semantic information.

If the dataset is made up of N objects, the trajectory for object j in this dataset is defined as the set of points $[x_j(t), y_j(t)]$ corresponding to the points with main direction changes; 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 $[cos(\theta), sin(\theta)]$, where θ is the angle which defines the vector joining $[x_j(1), y_j(1)]$ and $[x_j(end), y_j(end)]$.

We feed the feature vector formed by these elements to a hierarchical clustering algorithm. For a data set made of N trajectories there are $N*(N-1)/2$ pairs in the dataset. We employ the Euclidean distance as a measure of similarity to calculate the distance between all trajectory features. To avoid one feature to prime over the

others, particularly because distances in x and y are much bigger than distances on direction, input data related to the beginning and end of a trajectory is first normalised. The mean for each feature in x and y is first removed, then each record has its value divided by the standard deviation calculated from each feature. Because [$cos(\theta)$, $sin(\theta)$] are already bounded with values [-1, 1], these directional features are not transformed. Other distances have been envisaged such as the weighted sum of the features. Object trajectories with the minimum distance are clustered together. When two or more trajectories are set together the mean of the trajectory features is taken into account for further clustering. The successive merging of clusters is listed by a dendrogram. The evaluation of the dendrogram is typically subjective by adjudging which distance threshold appears to create the most natural grouping of the data. The final number of clusters is set by maximizing the evaluation criteria, which is defined in the next section. As the acquisition is performed in a multi-camera environment the clusters obtained can be generalised to different camera views with a 3D calibration matrix stage carried out for the on-line analysis system.

4.3 Trajectory analysis: Evaluation

In order to evaluate our trajectory analysis approach, we have defined a Ground-truth data set containing over 300 trajectories. The ground-truth trajectories were manually drawn on an image illustrating the empty scene. Figure 2 shows the empty scene for the Torino metro with some drawn trajectories. Semantic descriptions such as 'From North Entry to Vending Machine1' were generated. There are one hundred of such annotated semantic descriptions, which are called in the following trajectory types. Each trajectory type is associated with a main trajectory that best matches that description. Besides, two complementary trajectories define the confidence limits

within which we can still associate that semantic description. In figure 2 the main trajectory of each trajectory type is represented by a thick line while thin lines represent the complementary trajectories. Thus, each trajectory type is associated to triplets of trajectories.

We compute two performance measures to validate the quality of the proposed clustering approach, namely, Confusion and Dispersion. The former gives an indication of how many trajectory types of the ground-truth (how many main trajectories) are merged together in a single cluster resulting from the agglomerative procedure. Ideally the clustering algorithm should be able to dissociate all main trajectories in order to separate all trajectory types. In this case we would have Confusion = 1 (only one trajectory type is included in one cluster). If two main trajectories are included in the same cluster, then we say that Confusion = 2 as two different trajectory types are merged in the same cluster. In general terms the Confusion value for a cluster equals the number of main trajectories included.

The latter performance measure (Dispersion) indicates the number of erroneous clustered trajectories. This refers to the case where the main trajectory and any of its complementary trajectories (trajectories defining the confidence limits of the main trajectory) have been splitted into different clusters. Each ‘left-apart’ complementary trajectory increments the Dispersion measure by one unit. Figure 3 depicts the evolution of these two factors depending on the number of clusters which is chosen when running the clustering algorithm. For instance, for 21 clusters we have in total 15 complementary trajectories badly clustered (Dispersion) and 5 main trajectories per cluster (mean Confusion)

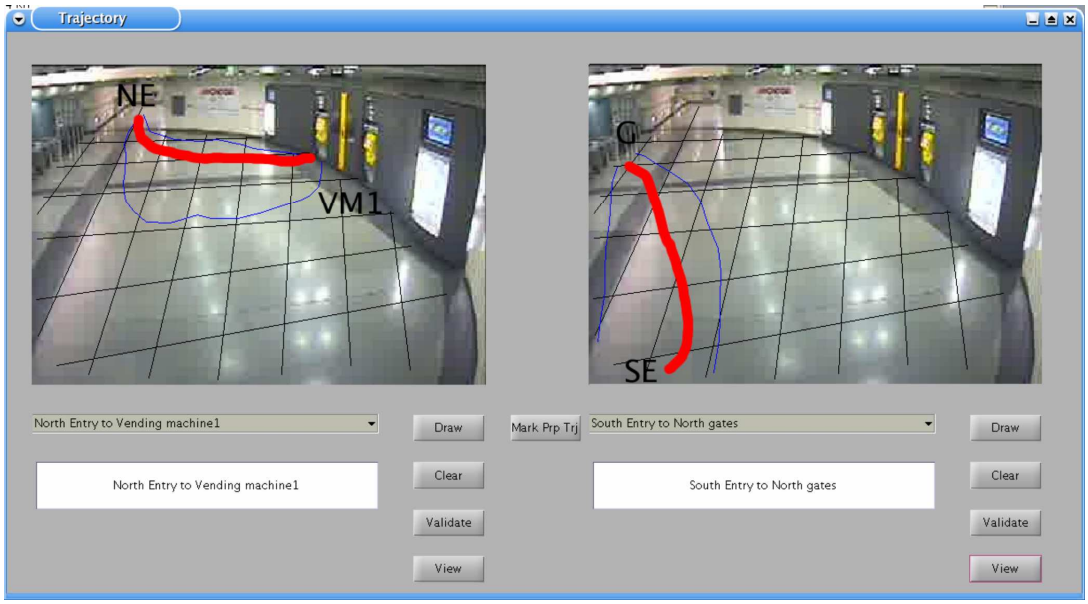


Figure 2. Ground-truth for two different semantic clusters. Left panel shows trajectories associated with the trajectory type 'From North Entry (NE) to Vending Machine1 (VM1)'. Right panel shows trajectories associated with the trajectory type 'From South Entry (SE) to Gates (G)'.

Because all trajectories can not be equally observed by the camera (for instance distinguishing all turnstiles in the upper left corner would require a larger spatial resolution) it is actually very difficult to achieve a bijection between the semantic labels and the resulting clusters. However, we aim at having the lowest possible confusion level together with the lowest percentage of dispersion. From Figure 3 it can be observed that a good compromise is achieved for a number of clusters between 15 to 35.

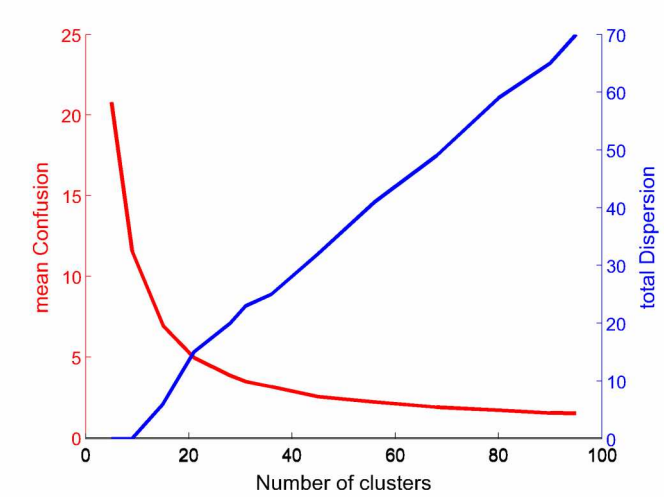


Figure 3. Evolution of the clustering quality measures Confusion (fusion of ground-truth semantic labels) and Dispersion (erroneous clustered trajectories) as a function of the number of clusters.

We also compared the agglomerative clustering algorithm in our application to other well-known clustering techniques such as k-means [35] and Kohonen self-organising maps (SOM) [36] employing the same ground-truth data. We observed that in the range of 1 to 35 clusters the Agglomerative algorithm has the less dispersion while the confusion remains almost identical for the three techniques. When the number of clusters increases, the SOM algorithm has a smaller dispersion than the other two techniques. However the amount of erroneously clustered data goes beyond 10%. The choice of the agglomerative algorithm remains thus valid for the number of clusters chosen before (between 15 to 35). Figure 4 depicts the evolution of the confusion and dispersion measures according to the number of clusters.

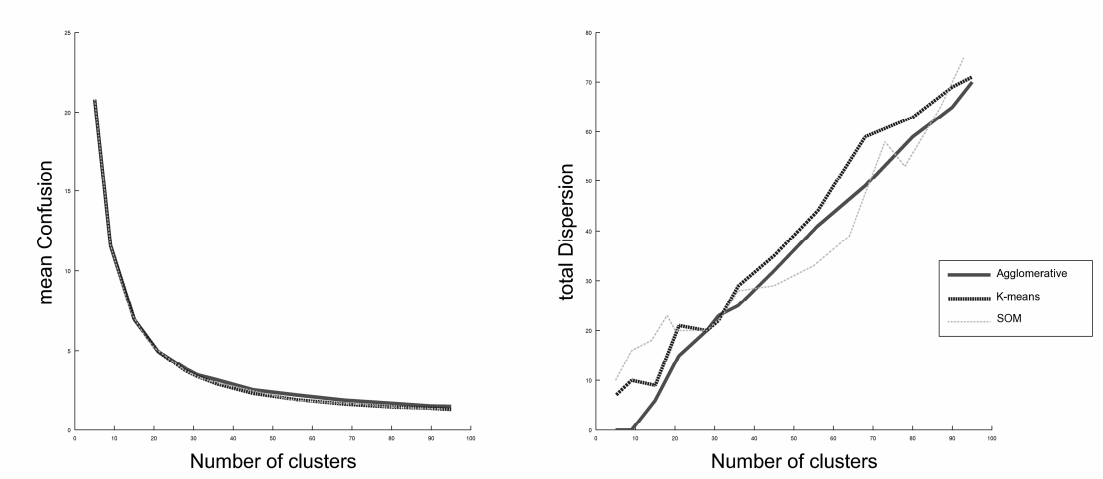


Figure 4. Evolution of the clustering quality measures Confusion (fusion of ground-truth semantic labels) and Dispersion (erroneous clustered trajectories) as a function of the number of clusters for the agglomerative clustering algorithm, k-means and Kohonen self-organising maps (SOM).

We further evaluated the three techniques with typical clustering validity indexes such as Silhouette, Dunn and Davis Bouldin indexes, which are described next:

Silhouette index

The Silhouette index [37-38] is defined as follows: Consider a data object v_j $j \in \{1,2,...,N\}$ belonging to cluster cl_i $i \in \{1,...,c\}$. This means that object v_j is closer to the prototype of cluster cl_i than to any other prototype. Let the average distance of this object to all objects belonging to cluster cl_i be denoted by a_{ij} Also, let the average distance of this object to all objects belonging to another cluster i' $i \sim = i'$ be called $d_{i'j}$ Finally let b_{ij} be the minimum $d_{i'j}$ computed over $i' = 1,...,c$ which represents the dissimilarity of object j to its closest neighbouring cluster. The Silhouette index is then

$$S = \frac{1}{N}\sum_{j=1}^{N} s_j \text{ where } s_j = \frac{b_{ij} - a_{ij}}{\max\{a_{ij}, b_{ij}\}}$$

This way, the best partition is achieved when S is maximized, which implies minimizing the intra-cluster distance (a_{ij}) while maximizing the inter-cluster distance (b_{ij})

Dunn index

The Dunn index [39] is defined as follows: Let cl_i and $cl_{i'}$ be two different clusters of the input dataset. Then, the diameter Δ of cl_i is defined as $\Delta(cl_i) = \max_{v_j, v_{j'} \in cl_i} \{d(v_{j'}, v_j)\}$

Let δ be the distance between cl_i and $cl_{i'}$ Then δ is defined as

$\delta(cl_i, cl_{i'}) = \min_{v_j \in cl_i, v_{j'} \in cl_{i'}} \{d(v_{j'}, v_j)\}$, and, $d(x, y)$ indicates the distance between points x and y.

For any partition, the Dunn index is

$$v_D = \frac{\min}{i}\left\{\frac{\min}{i'}\left\{\frac{\delta(cl_i, cl_{i'})}{\max_i\{cl_i\}}\right\}\right\} \text{ and } i, i' \in \{1, \ldots, N\},\ i' \neq i$$

Larger values of v_D correspond to a good clustering partition.

Davis Bouldin index

The Davis Bouldin index [40] is defined as follows: This index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The scatter within cluster, cl_i, is computed as

$$S_i = \frac{1}{|cl_i|} \sum_{v_j \in cl_i} \left\{ \left\| v_j - m_i \right\| \right\}$$

m_i is the prototype for cluster cl_i . The distance δ betweenclusters cl_i and $cl_{i'}$ is defined as

$$\delta(cl_i, cl_{i'}) = \left\| m_{i'} - m_i \right\|$$

The Davies-Bouldin (DB) index is then defined as

$$DB = \frac{1}{N} \sum_{i=1}^{N} R_i \text{ with } R_i = \max_{i,i'} R_{ii'} \quad i,i' \in \{1,...,N\},\ i' \neq i \text{ and}$$

$$R_{ii'} = \frac{S_i + S_{i'}}{\delta(cl_i, cl_{i'})}$$

Low values of the DB index are associated with a proper clustering.

The results from applying these measures are summarised in table 1 for a number of clusters equal to 31, chosen as best choice from the confusion dispersion curves. It can be observed that the agglomerative algorithm is the best choice according to Silhouette and Dunn indexes. According to Davis Bouldin index, the SOM technique would be the best choice. However, for this index and the Silhouette, all methods are close from each other. For the Dunn index, the agglomerative method is clearly the best choice.

		Clustring Technique			
		Agglomerative	K-means	SOM	
Validity index	Silhouette	0.32116	0.29501	0.3036	Higher is better
	Dunn	0.10098	0.064249	0.05031	Higher is better
	Davies Bouldin	0.58685	0.49352	0.37311	Smaller is better

Table 1. Evaluation of the clustering quality indexes (Silhouette, Dunn and Davis Bouldin) for the agglomerative clustering algorithm, k-means and Kohonen self-organising maps (SOM).

5 Knowledge representation and statistical analysis

There are two main types of concepts to be represented from the video: physical objects of the observed scene and video events occurring in the scene. The former category can still be further subdivided into two types of physical objects of interest: mobile and contextual objects. Mobile objects of interest are the source of action occurring in the scene. Contextual objects are 3D objects of the empty scene model corresponding to the static environment of the scene. For the off-line analysis of both types of concepts, and with the aim of setting the data in a suitable format to achieve knowledge discovery, we separate the information corresponding to the activities occurring over a period of time on three different semantic tables, namely mobile objects, contextual objects and video events. This information is characterized by a set of specific features, which is currently being enriched along the CARETAKER project. We are reporting in the following the features that have been used in our experimentations.

5.1 Mobile objects

The mobile object, m, can be represented as a ten-tuple:

$$m = \left\{ m^{id}, m^{type}, m^{start}, m^{end}, m^{duration}, m^{dist\_org\_dest}, m^{shape}, m^{involved\_events}, m^{significant\_event}, m^{trajectory} \right\}$$

where m_j with $j \in \{1,2,\ldots,N\}$ is then,

- m_j^{id} : the identifier label of the object. $m_j^{id} \in \mathbb{Z}^+$

- m_j^{type} : the class the object belongs to:

  $m_j^{type} \in \{'Person','Group','Crowd','Train','Luggage'\}$.

- m_j^{start} : the time when the object is first seen.

- m_j^{end} : the time when the object is last seen.

- $m_j^{duration}$: the total time the object has been observed. $m_j^{duration} = m_j^{end} - m_j^{start}$

- $m_j^{dist\_org\_dest}$: the total distance walked from origin to destination. $m_j^{dist\_org\_dest} = \sum_{t=1}^{T_j - 1} \sqrt{\left(x_j(t+1) - x_j(t)\right)^2 + \left(y_j(t+1) - y_j(t)\right)^2}$ and $t \in \{1,\ldots,T_j\}$ where T_j is the number trajectory sampled points for the object m_j.

- m_j^{shape} : the label describing the object's shape depending on the object's ratio height/width. $m_j^{shape} = \{'Small','Medium','Large'\}$

- $m_j^{involved\_events}$: all occurring Events related to the identified object.

- $m_j^{significant\_event}$: the most significant event among all events. This is calculated as the most frequent event related to the mobile object.

- $m_j^{trajectory}$: the trajectory cluster identifier characterising the object. $m_j^{trajectory} = i \Leftrightarrow m_j \in cl_i$

5.2 Contextual objects

The activities involving a contextual object, c, can be characterised by a 12-tuple:

$$c = \left\{ \begin{matrix} c^{id}, c^{type}, c^{start}, c^{end}, c^{involved\_events}, c^{significant\_event}, c^{rare\_event}, c^{event\_histogram}, c^{involved\_mobiles}, \\ c^{mobiles\_histogram}, c^{use\_duration}, c^{mean\_time\_of\_use} \end{matrix} \right\}$$

where c_k and $k = \{1,...,O\}$ is then,

$c_k^{id}, c_k^{involved\_events}, c_k^{significant\_event}$ are defined in the same way as for the mobile objects but referring to contextual objects and $c_k^{type} \in z \cup eq$, as defined above $z = \{'platform', 'validating\_zone', 'vending\_zone'\}$, $eq = \{g_1, ...g_{10}, vm_1, vm_2\}$ where g_i is the ith gate and vm_i is the ith vending machine.

$c_k^{start}, c_k^{end}, c_k^{involved\_mobiles}$ are defined in the same way as for the mobile objects interacting with the contextual object.

The remaining fields indicate

- $c_k^{rare\_event}$: this is the rarest event.

- $c_k^{event\_histogram}$: gives the number of occurrence of all involved events per type of event.

- $c_k^{mobiles\_histogram}$: gives the number of appearance for all involved mobile objects per object type.

- $c_k^{use\_duration}$: percentage of occupancy (or use of a contextual object). For instance, the Ticket Machine has a 10% of use over the observation time.

- $c_k^{mean\_time\_of\_use}$: average time for a mobile object to interact with the contextual object.

The contextual objects to be monitored are predefined by the end-users in the model of the scene environment. This modelling phase is a quick process and enables to acquire the end-user expertise on the objects of interest. For the video sequences analysed in this work, the contextual objects of interest are: 'Platform hall', 'Gates', and 'VendingMachines'.

5.3 Video events

A video event, e , can be represented as a 6-tuple:

$$e = \left\{ e^{id}, e^{type}, e^{start}, e^{end}, e^{involved\_mobiles}, e^{involved\_context\_obj} \right\}.$$

where e_l and $l = \{1,...,M\}$ is then,

- e_l^{id} : the identifier label for the detected Event. $e_j^{id} \in \mathbb{Z}^+$

- e_l^{type} : the class where the Event belongs to. $e_l^{type} = \{'close\_to','stays\_at',...\}$

- e_l^{start} : the first time when the Event is detected.

- e_l^{end} : the last time when the Event is seen.

- $e_l^{involved\_mobiles}$: the identifier label of the mobile objects involved in that event.

- $e_l^{involved\_context\_obj}$: the contextual object involved in that event.

5.4 Statistical analysis

Statistical information can be obtained from the mobile objects and the contextual objects as well as their interactions. This is a major information source for the end-user. For instance, on large metro video recordings, there is spatial and temporal information on the use of contextual objects. In this work we calculate the following statistical indexes.

- number_of_users: the total number of people interacting with a contextual object. The number_of_users is calculated as follows:

  Let $E = \{e_{l'}\} | e_{l'}^{involved\_context\_obj} = c_k^{type}$, and $l' \subseteq l$, $l = \{1,...,M\}$. M is the total number of events observed in the video, thus $E \subseteq \{e_l\}$.

  $number\_of\_users = cardinal\left(c_k^{involved\_mobiles}\right)$ with $c_k^{involved\_mobiles} = \left\{e_{l'}^{involved\_mobiles}\right\}$.

  $c_k^{involved\_mobiles}$ contains no repeated elements.

- percentage_of_use: the ratio between the total period of time a contextual object is in use to the total observation period

  $percentage\_of\_use = \dfrac{e_{l'(M')}^{end} - e_{l'(1)}^{start}}{observation\_period}$ if $cardinal(E) = M'$ and $M' \leq M$.

  M' is the total number of events observed in the video where the contextual object c_k appears.

- interaction_duration: the mean time a user spends when interacting with a contextual object.

  Let $E' = \{e_{l''}\} \mid e_{l''}^{involved\_context\_obj} = c_k^{type}, e_{l''}^{involved\_mobiles} \subseteq c_k^{involved\_mobiles}$

  $Interaction\_duration = mean\left(e_{l''(M'')}^{end} - e_{l'(1)}^{start}\right)$ if $cardinal(E') = M''$ and $M'' \leq M'$.

  M'' is the total number of events observed in the video where one of the objects listed in $c_k^{involved\_mobiles}$ interacts with the contextual object c_k.

The statistics can be visualised with an interactive user interface enabling to study a given variable such as zone of interest, equipment, etc.

6 Relational analysis

Once all statistical measures of the activities in the scene have been computed and the corresponding information is put into the proposed model format, we aim at discovering complex relationships that may exist between mobile objects themselves, and between mobile objects and contextual objects in the scene. For this task, the clustering methodology we decided to use RARES (Relational Analysis And Regularized Similarity). This methodology gathers two different technologies: relational analysis theory and regularized similarity [41-43]. Relational analysis has been initiated and developed at the European Centre of Applied Mathematics (ECAM) at IBM France By F. Marcotorchino and P. Michaud [11, 44]. The principle of relational analysis consists in transforming the data usually represented as a $N \times M$ rectangular matrix where N is the number of objects to be clustered and M is the number of variables measured on these objects to two new $N \times N$ matrices

representing respectively a global similarity and a global dissimilarity measures for each pair of objects. The relational analysis algorithm will compare, for any two objects, their similarity and their dissimilarity. If the similarity is greater than the dissimilarity, then this two objects will be put in the same cluster of the final obtained partition. The output of the relational analysis algorithm is a set of groups of objects, where inside each group the objects are more similar to each other than to the objects belonging to another group.

To define the global similarity matrix, relational analysis transforms each variable V^k $(k = 1,2,\ldots,M)$ to a $N \times N$ matrix S^k where the term $s^k_{ii'}$ is the similarity measure between the two objects i and i' w.r.t. variable V^k. A dissimilarity measure $\bar{s}^k_{ii'}$ is then computed as the complement to the maximum similarity measure possible between these two objects. As the similarity between two different objects is less or equal to their self-similarities: that is $s^k_{ii'} \le \min\left(s^k_{ii}, s^k_{i'i'}\right)$, the dissimilarity is $\bar{s}^k_{ii'} = \min\left(s^k_{ii}, s^k_{i'i'}\right) - s^k_{ii'}$, then we define that the global similarity measure between objects i and i' over the M variable is $s_{ii'} = \sum_{k=1}^{M} s^k_{ii'}$ and their global dissimilarity is also $\bar{s}_{ii'} = \sum_{k=1}^{M} \bar{s}^k_{ii'}$. To cluster a population of N objects described by M variables, the relational analysis theory is based on the maximisation of the Condorcet criterion $C(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(s_{ii'}x_{ii'} + \bar{s}_{ii'}\bar{x}_{ii'}\right)$ where X is a binary $N \times N$ matrix representing the partition to discover in the data. The general term $x_{ii'}$ of matrix X is defined as follows:

$$x_{ii'} = \begin{cases} 1 & \textit{if } i \textit{ and } i' \textit{ are in the same cluster} \\ 0 & \textit{otherwise} \end{cases}$$

and $\bar{x}_{ii'} = 1 - x_{ii'}$

The mathematical formulation of the criterion to be maximised is:

$$\max(C(X)) \; w.r.t.$$
$$\begin{cases} x_{ii} = 1 & \textit{reflexivity} \\ x_{ii'} = x_{i'i} & \textit{symmetry} \\ x_{ii'} + x_{i'i''} - x_{ii''} \leq 1 & \textit{transitivity} \end{cases}$$

It seems evident that variables having only two modalities, for example, will tend to generate more similarities than those variables having bigger number of modalities. For example, it is less likely, for two persons chosen at random in Paris, to live in the same district (20 symbolic values) than to have the same gender (only 2 symbolic values). Regularized similarity, developed by H. Benhadda and F. Marcotorchino [42-43], is a theory taking into account these internal structures, during the computation of the individuals' similarities; to rebalance the influences (too strong or too weak) induced, in an implicit way, by these structures. This will favour certain variables compared to the others and will create, thus, some biases. Regularized similarity did not bring noticeable clustering improvements in our analysis. Indeed, the number of symbolic values between variables does not change dramatically. Simple similarity is reported in this paper. Relational analysis is employed here for activity clustering as it is able to characterise heterogeneous data (including symbolic and numerical attributes) contrary to hierarchical clustering, which works with numerical-only data.

In the present work, the input of the system is then the whole set of detected mobile objects, $m_j \quad j \in \{1,2,...,N\}$, with the features described in section 5.1, namely:

$m_j = \left\{ m_j^{type}, m_j^{duration}, m_j^{dist\_org\_dest}, m_j^{shape}, m_j^{significant\_event}, m_j^{trajectory} \right\}$, thus in our analysis M=6. The output of the analysis is the final partition, named Ω, which indicates the elements m_j and $m_{j'}$ that should be set together in the same cluster Π.

Several indicators are computed during the clustering process, to build up the final partition, named Ω and to measure the quality of the obtained results. For a couple of objects i and i' belonging to the population P and two clusters Π and Π' belonging to Ω, we define:

- The maximum similarity possible $\Lambda_{ii'}$ between any two objects i and i' is: $\Lambda_{ii'} = Min(c_{ii}, c_{i'i'})$
- the link $L_{ii'}$ between two objects by: $L_{ii'} = c_{ii'} - \alpha \times \Lambda_{ii'}$, where α is a parameter such that $0 < \alpha \leq 0.5$.
- the link $L_{i\Pi}$ between object i and cluster Π, by : $L_{i\Pi} = \sum_{i' \in \Pi} L_{ii'}$
- the agreement $A_{\Pi\Pi'}$ between the two clusters, by: $A_{\Pi\Pi'} = \sum_{i \in \Pi} \sum_{i' \in \Pi'} c_{ii'}$
- the disagreement $\overline{A}_{\Pi\Pi'}$ between two clusters, by: $\overline{A}_{\Pi\Pi'} = \sum_{i \in \Pi} \sum_{i' \in \Pi'} \overline{c}_{ii'}$
- the maximal own similarity $\Lambda_{\Pi\Pi'}$ between two clusters, by: $\Lambda_{\Pi\Pi'} = \sum_{i \in \Pi} \sum_{i' \in \Pi'} \Lambda_{ii'}$
- and the link $L_{\Pi\Pi'}$ between two clusters, by: $L_{\Pi\Pi'} = A_{\Pi\Pi'} - \alpha \times \Lambda_{\Pi\Pi'}$

The quality Q_Π of a particular cluster Π is defined by:

$$Q_\Pi = \frac{A_{\Pi\Pi} + 2 \times \sum_{\Pi' \neq \Pi} \overline{A}_{\Pi\Pi'}}{\Lambda_{\Pi\Pi} + 2 \times \sum_{\Pi' \neq \Pi} \Lambda_{\Pi\Pi'}}$$

This measure takes into account at the same time the inner homogeneity and the external heterogeneity of cluster Π. The more the objects belonging to Π are similar

to each other and in the same time the more they are dissimilar from the objects belonging to the other clusters, the better is the quality of the cluster.

The quality Q of the final partition Ω is an indicator that measures the total coherence of Ω, and is given by the formula:

$$Q = \frac{\sum_{\Pi \in \Omega} \left(A_{\Pi\Pi} + \sum_{\Pi' \neq \Pi} \overline{A}_{\Pi\Pi'} \right)}{\sum_{\Pi \in \Omega} \sum_{\Pi' \in \Omega} \Lambda_{\Pi\Pi'}}$$

7 Results

7.1 Results with annotated data

We first tested the validity of our clustering algorithms (Hierarchical and Relational clustering) on labelled video data. Caviar is an EC founded project that has made available a dataset of video clips with hand-labelled ground truth [12]. We focused our attention on the first part of the dataset, which contains people observed at the lobby entrance of a building. The annotated ground-truth include for each person its bounding box (id, centre coordinates, width, height, main axis orientation) with a description of his/her movement type (inactive, active, walking, running) for a given situation (moving, inactive, browsing) and with a given scenario context (browsing, immobile, left object, walking, drop down).

We have applied first the hierarchical clustering algorithm to a dataset containing 164 persons or objects. We have extracted their trajectories as the centre coordinates of the bounding box over time. We have tuned the algorithm with the user interface to obtain 31 meaningful clusters. Figure 5 shows in panel 5a all trajectories from this dataset. The remaining plots in the figure are the three most common paths undertaken. As it can be observed clear trajectory patterns can be extracted from the clusters. For instance, the three paths clusters from figure 4 can be labelled: cluster 8: 'entering right and up / exiting left' (panel 5b), cluster 11: 'entering right bottom / exiting left' (panel 5c), and cluster 15: 'entering right-middle / exiting right-bottom' (panel 5d).

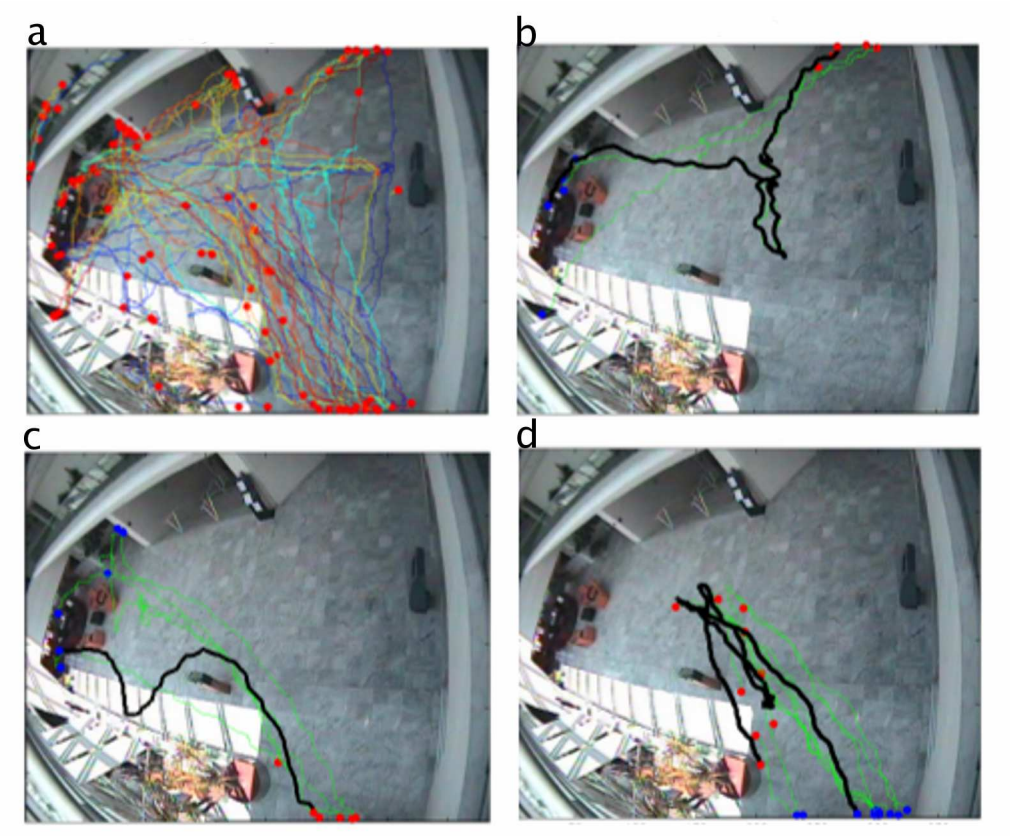


Figure 5. (a) Original set of CAVIAR trajectories (upper-left panel) and three

clusters showing most common undertaken paths. (b) Cluster 8; (c) Cluster 11; (d) Cluster 15. Beginning of the trajectories are indicated by red points. End of trajectories are indicated by blue points.

We have then applied relational analysis in order to obtain higher relations between objects. For this purpose we have employed the object representation format described before. To perform the evaluation and because the annotated ground-truth was available with a situation and context description, we have generated a set of events by concatenating the three pieces of available information: movement's type (T), context (C) and situation (S). The different symbolic values that the 3-tuple TCS can take are presented in the table 2.

Type	Context	Situation
(i)nactive	(b)rowsing	(m)oving
(a)ctive	(i)mmobile	(i)nactive
(w)alking	(l)eft object	(b)rowsing
(r)unning	(w)alking	
	(d)rop down	

Table 2. Semantic event information in CAVIAR.

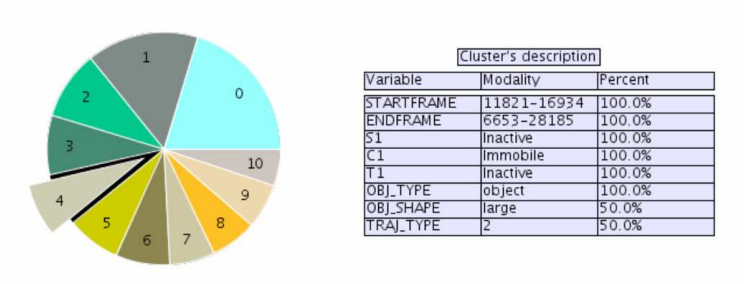
For example, an Event having the value 'awm' is related to an object with movement's type = '(a)ctive', within a context = '(w)alking' and situation = '(m)oving'. In the analysed data, an object is usually involved in between 1 and 12 such events during the observation time. To give account of the temporal succession of events, the 3-

tuple TCS is sequentially numbered as the events appear. For instance if 3 events are in succession associated to a mobile object, these events will be designated as [T1 C1 S1], [T2 C2 S2] and [T3 C3 S3]. A portion of the input data is shown below in Table 3.

obj_id	obj_type	startframe	endframe	trajectory_type	obj_shape	Involved_Event1	Involved_Event2	Involved_Event3
84	persor	´3785	14312	19	big	wwm		
85	persor	´4338	14611	6	small	iii	aii	wir
86	persor	´4459	14523	16	tall	wwm		
88	persor	´4463	14684	7	big	rwm	wwm	
87	persor	´4491	14684	10	big	wm	aii	
89	persor	´4685	15097	15	tall	wm	iii	aii
90	persor	´4757	14938	16	tall	wwm		
93	persor	´5054	15542	16	big	wm	aii	wir
92	object	´5256	15547	3	small	iii		
94	persor	´5548	15670	15	large	wm	aii	iii
95	persor	´5695	15750	14	small	iii	aii	iii
96	persor	´5817	16699	14	big	abb	wbm	abb
98	object	´6051	16563	19	small	iii		
99	persor	´6920	17082	14	small	wm	aii	iii
103	persor	´6942	17513	4	small	wm	aii	iii
104	persor	´6966	17513	4	small	wm	aii	iii

Table 3. Input matrix used for the relational analysis clustering method. Remark that only some objects from the total detected set are represented in the table.

One of the clusters that RARES has discovered in the CAVIAR dataset is presented in Figure 6. This cluster contains 4 detected objects. All objects are involved in only one Event corresponding to inactive objects, in an inactive situation and in an immobile context. This cluster actually corresponds to the objects ‘bag’ that were annotated in the CAVIAR database as abandoned.



Cluster's description

Variable	Modality	Percent
STARTFRAME	11821-16934	100.0%
ENDFRAME	6653-28185	100.0%
S1	Inactive	100.0%
C1	Immobile	100.0%
T1	Inactive	100.0%
OBJ_TYPE	object	100.0%
OBJ_SHAPE	large	50.0%
TRAJ_TYPE	2	50.0%

Figure 6. Resulting partition of the CAVIAR data after running the relational analysis algorithm. Properties of cluster 4 are given.

Another cluster is presented in Figure 7. In this case all items are people that are involved in at least three events. For all people, the first event is of type walking, in a moving situation and within a context where something drops_down (75 % of the cases). Then the situation and the context will evolve and all people will be involved in an event (3rd event), described with an inactive situation, within a context where something drops_down, and with an inactive movement type (75 % of the cases). This cluster matches actually the objects that were annotated in the CAVIAR database as 'falls down' and included in the fighting (one man down) scenarios.

Cluster's description

Variable	Modality	Percent
ENDFRAME	6653-28185	100.0%
C4	Drop_Down	100.0%
S3	Inactive	100.0%
C3	Drop_Down	100.0%
S1	Moving	100.0%
T1	Walking	100.0%
OBJ_TYPE	person	100.0%
STARTFRAME	6654-11821	75.0%
S5	Moving	75.0%
T4	Active	75.0%
T3	Inactive	75.0%
S2	Inactive	75.0%
C2	Drop_Down	75.0%
T2	Active	75.0%
C1	Drop_Down	75.0%
OBJ_SHAPE	large	75.0%
TRAJ_TYPE	25	75.0%
C5	Drop_Down	50.0%
T5	Walking	50.0%
S4	Moving	50.0%

Figure 7. Properties for cluster 5 in the CAVIAR data partition.

The activities described in the CAVIAR ground-truth are thus correctly retrieved with our algorithm. The table 4 gives a quantitative evaluation on the correspondence between the events annotated in the CAVIAR ground-truth and the clusters obtained by the relational analysis algorithm. As it can be observed, all instances from the left luggage event are well recognised in cluster 4. Fighting situations are generally recognised in cluster 0. The missing fighting cases are those corresponding to the case where one of the individuals involved in the fight falls down and lies on the floor. This kind of situation is matching cluster 5. Most cases of the browsing event are matched in cluster 2.

Event	Number of cases	Corresponding Cluster	Detection performance			
			True positive	False positive	True negative	Cluster Quality
Left baggage	4	4	100%	0	0	92%
Fight	12	0	67%	0	33%	72%
Fight and fall down	4	5	75%	25%	25%	73%
Browsing	12	2	86%	0	14%	74%

Table 4. Performance evaluation on the correspondence between the events annotated in the CAVIAR ground-truth and the relational analysis activity clusters.

7.2 Results with large video recordings

We have processed 73000 frames of video from the Torino underground (GTT, Italy), with an acquisition rate of 25 frames/s equals to about fifty minutes of video. We have analysed off-line this period of time. We have applied the hierarchical clustering algorithm on the trajectories of mobile objects to obtain the common behavioural paths undertaken by the people in the hall. Figure 8 presents the whole dataset of trajectories that we have analysed. Using our interactive user interface to maximize the evaluation criteria (explained in section 4.3), we have applied the hierarchical clustering selecting 31 clusters. The most common paths that people are taking are shown in Figure 9.

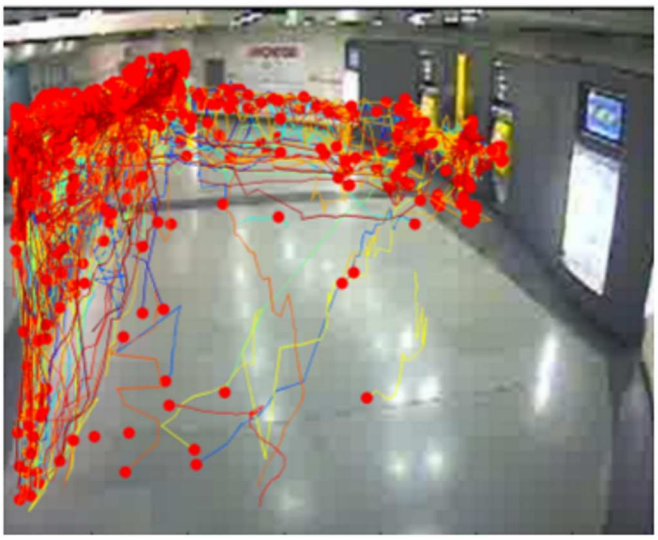


Figure 8. Trajectories detected in one station of the Torino underground (2025 trajectories during 50 minutes). Beginning of the trajectories are indicated by a red point.

The trajectory clusters give an useful semantic information on the behaviours undertaken by the metro users. For instance, Cluster 14 indicates that most users enter the station by the north doors and go rather straight to the gates. Cluster 1 indicates that most people take also the north doors to exit the station. Cluster 25 represents fewer users exiting the gates and leaving the station to go through the south doors. Cluster 2 shows people with stationary activity near the gates. Cluster 26 indicates that few users buy a ticket before going trough the gates. Cluster 5 indicates that after buying a ticket, users go straight to the gates to take the metro.

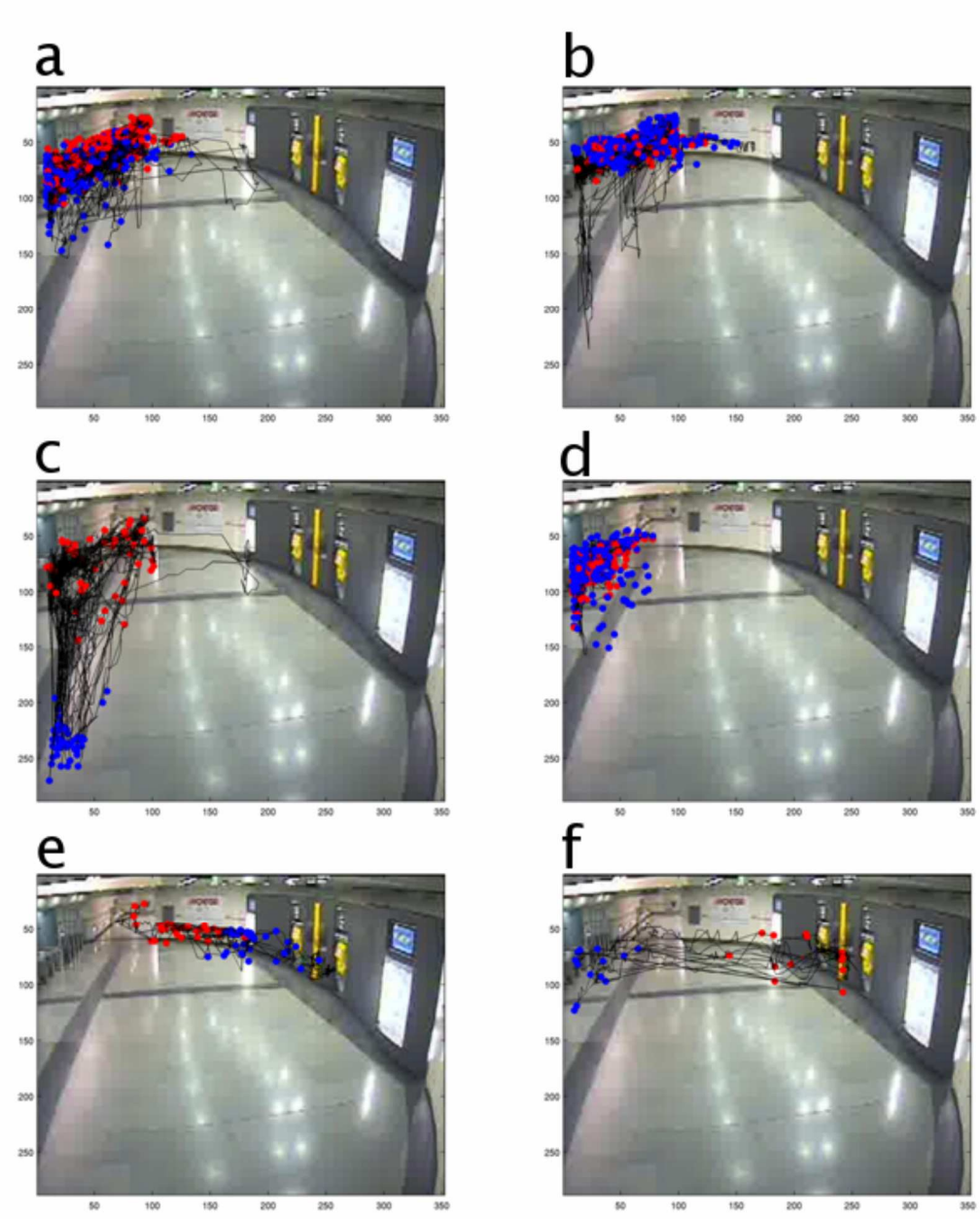


Figure 9. Clusters showing the most common paths obtained from the dataset shown in figure 7. (a) Cluster 14 with 443 trajectories; (b) Cluster 1 with 744 trajectories; (c) Cluster 25 with 50 trajectories; (d) Cluster 2 with 338 trajectories; (d) Cluster 26 with 41 trajectories; (e) Cluster 5 with 13 trajectories. Beginning of the trajectories are indicated by red points. End of trajectories are indicated by blue points.

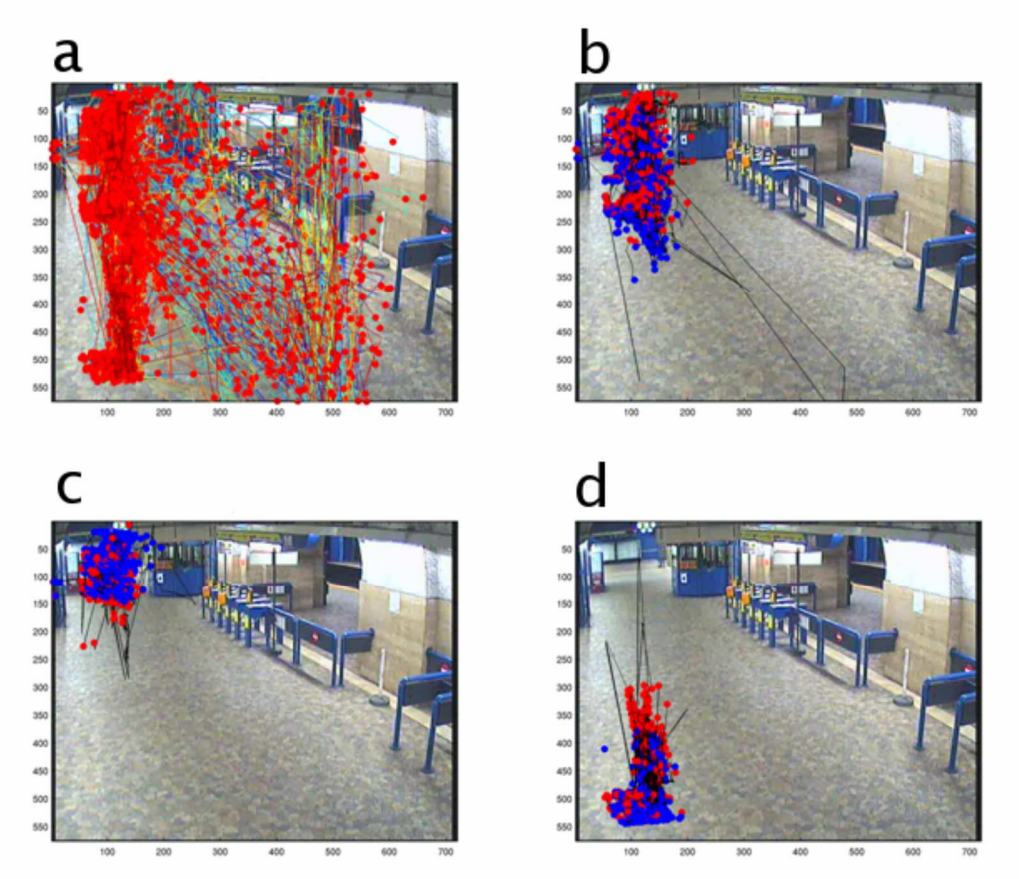


Figure 10. Clusters showing the most common paths obtained from the Roma dataset. (a) Whole dataset; (b) Cluster 20 with 1451 trajectories; (c) Cluster 19 with 809 trajectories; (d) Cluster 7 with 790 trajectories. Extremity points of the trajectories are represented as in figure 9 by red and blue points.

The data processing of the Roma underground (ATAC, Italy) corresponds to five hours and half of video, which is acquired at a rate of 5 frames/s. The same unsupervised clustering algorithm was applied to the data, which contains over 14000 tracked objects. Figure 10 presents the whole dataset of trajectories that we have analysed together with some of the clusters representative of the most common paths. For instance, trajectory cluster 20 in the upper right panel of the figure

presents people entering the hall from the north doors. This is actually the biggest trajectory cluster followed by trajectory cluster 19 (lower left panel) showing people leaving the hall through north doors. Trajectory cluster 7 shows people exiting the hall trough south doors, however, the number of elements of this cluster is smaller than the two previous ones. This means that south doors are less employed to enter/exit the hall.

We have further performed statistical analysis to measure the interactions between users and contextual objects. As mentioned in section 3.2 there are two types of contextual objects of interest in the scene, the zones (mainly the hall), and the equipement (the Gates, and the VendingMachines). Over the whole observation time, concerning the Torino underground, people were practically constantly present in the hall as we obtained a percentage of use of the hall of 91% of the observation time. The gates had a percentage of use of 36% indicating that the flow of people trough the gates was not constant over the observation time. The two vending machines of the Torino station had a percentage of use of only 8% and 7% respectively indicating that most people did not stop long-time or did not need to buy a ticket while in the station. This is further confirmed by the fact that most users buying tickets were detected as single individuals and came more rarely to the machines as groups. No crowd (*i.e.* no queue) was detected at the vending machines neither at the gates, whereas crowding was detected in the hall but at a low degree.

Regarding the interactions of people with contextual objects in the Roma station, people were observed in the hall 95% of the observation time and the gates were employed 91% of the time indicating that the flow of people trough the gates is much more constant than, for instance, in the Torino underground. The vending machines

are often used, 77% of the observation time. Compared to the Torino station, this utilization time indicates that the Roma station is in general much busier. This is further confirmed by the fact that person groups are frequently detected at the vending machines (632 in the last two hours), at the hall (1441) and at the gates (813). Crowding situations were detected only in the hall (227 crowd groups).

We have translated the information related to detected objects and statistical information in the format presented in section 4. As an example, a portion of the Torino semantic tables obtained are presented next (Tables 5 to 7).

ctx_obj_id	ctx_obj_type	startframe	endframe	sig_evnt_type	rare_evnt_type
1	Platform	20050	92815	inside_zone(14486)	group_stays_inside_zone(30)
2	Gates	20055	92745	close_to(5103)	group_stays_at(40)
3	VendingMachine2	26560	90725	close_to(1020)	group_stays_at(200)
4	VendingMachine1	30680	90650	close_to(834)	group_stays_at(160)

ctx_obj_id	evnt_hist
1	inside_zone(14486) group_inside_zone(5523) crowd_inside_zone(1750) stays_inside_zone(1276) crowding_in_zone(109)
2	close_to(5103) stays_at(3489) group_close_to(329) group_stays_at(40)
3	close_to(1020) stays_at(490) group_close_to(341) group_stays_at(200)
4	close_to(834) stays_at(482) group_close_to(307) group_stays_at(160)

ctx_obj_id	mob_obj_hist	percnt_of_use	mean_time_of_use
1	Person(491) Unknown(102) PersonGroup(136) Luggage(34) Crowd(4)	91.1296%	35500.854
2	Person(135) Unknown(28) Luggage(12) PersonGroup(17)	36.8284%	35682.2396
3	Unknown(19) Person(15) Luggage(10) PersonGroup(2)	8.8704%	7401.087
4	Unknown(12) Person(11) Luggage(8)	7.7504%	16931.2903

Table 5. Contextual Objects semantic table.

evnt_id	evnt_type	startframe	endframe	inv_objs_id	ctx_type
10161	inside_zone	39085	39085	1573	Platform
10162	close_to	39085	39085	1444	Gates
10163	inside_zone	39090	39090	1444	Platform
10164	inside_zone	39090	39090	1573	Platform
10165	close_to	39090	39090	1444	Gates
10166	inside_zone	39095	39095	1444	Platform
10167	inside_zone	39095	39095	1573	Platform
10168	close_to	39095	39095	1444	Gates
10169	inside_zone	39100	39100	1444	Platform
10170	inside_zone	39100	39100	1573	Platform
10171	close_to	39100	39100	1444	Gates
10172	group_inside	39105	39105	1573	Platform

Table 6. Events semantic table.

mob_obj_id	mob_obj_type	duration_in_sec	traj_type	dist_org_dest	shape_type	sig_evnt_id
1	Person	[48.24 - 52.04]	1	[334.9776 - 485.3803]	small Person	inside_zone_Platform
2	Person	[364.84 - 367.24]	1	[976.9468 - 1113.7527]	small Person	stays_at_Gates
3	Unknown	[364.84 - 367.24]	1	[3158.1909 - 3230.0034]	small Unknown	void
4	Luggage	[295.24 - 298.84]	2	[671.8608 - 841.0493]	small Luggage	void
5	PersonGroup	[41.24 - 47.44]	14	[60.075 - 205.4118]	small PersonGroup	group_inside_zone_Platform
6	Person	[0.04 - 5.64]	1	[334.9776 - 485.3803]	small Person	inside_zone_Platform
7	Unknown	[33.84 - 40.24]	1	[60.075 - 205.4118]	small Unknown	void
8	Person	[0.04 - 5.64]	1	[0 - 59.0762]	small Person	inside_zone_Platform
9	PersonGroup	[336.84 - 336.84]	1	[3630.5716 - 3723.4675]	small PersonGroup	group_stays_at_Gates

Table 7. Mobile Objects semantic table.

From the contextual object table, we have been able to follow the evolution of the interactions with contextual objects. For Torino, Figure 11 shows a graphic on the evolution of the number of people present in the Torino hall during the observation time. For this observation period, the peak hour is detected at 6 h 45 min with 200 people in five minutes with an average of 180 people. Figure 12 shows the evolution on the usage of a vending machine. Interestingly, a user spends more time (about 40% increase of time) with the vending machine when the hall is not crowded.

For Roma, the number of people travelling trough the hall is bigger than in Torino with an average of 355 people constantly in the hall.

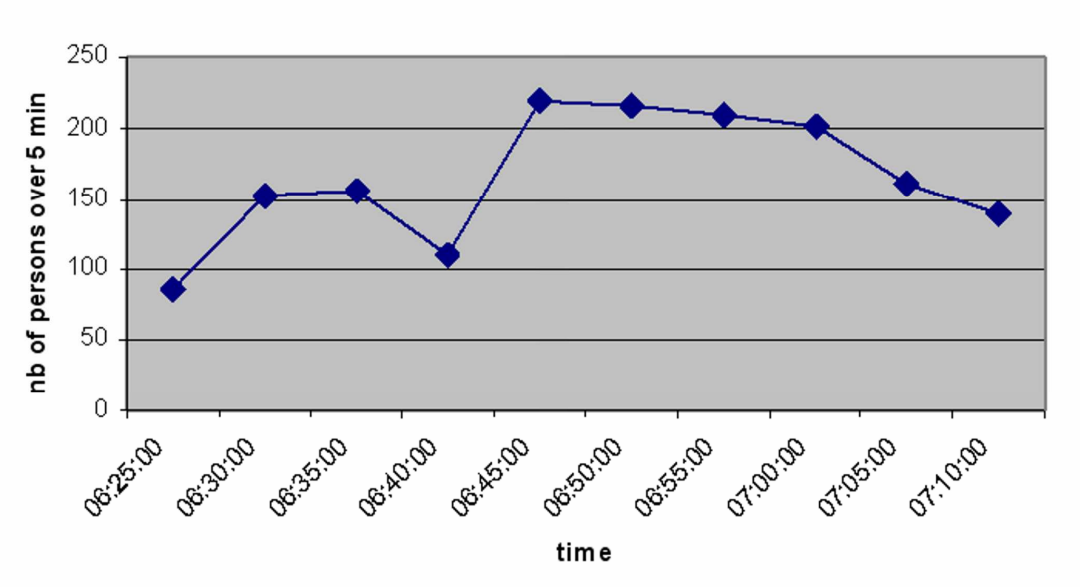


Figure 11. Temporal evolution in the number of people occupying the station hall in the Torino station.

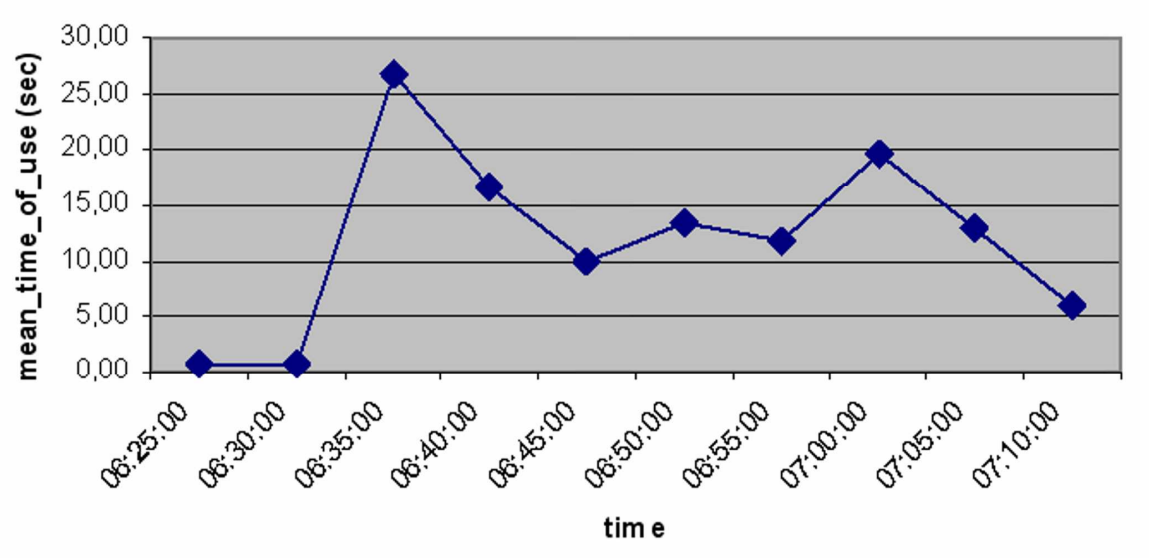


Figure 12. Temporal evolution of the mean time a user spends on a vending machine in the Torino station hall.

In the last step towards knowledge discovery we have applied the relational analysis process explained before in section 5. The input was the mobile object semantic table described above and containing in total 2025 detected objects in the case of the Torino data and 14757 for the Roma data. Some of the clusters returned are detailed next.

For the Torino dataset, the biggest cluster, labelled 0, contained 661 elements with the following variable properties: shape type : small person (100% of the elements); mobile object type : person (100% of the elements); duration : [0.04 s – 5.64 s] (82% of the elements); significant event : inside_zone_hall (67% of the elements); distance origin to destination : [0 cm – 59 cm] (47% of the elements); trajectory type : 1 (44% of the elements). In other words this activity cluster is made only of persons detected for a very short period of time (less than 6 seconds) and also moving in a short range

of about 60 cm. They are associated to Trajectory type 1 ‘from gates to north doors’ (shown in fugure 9). Similar descriptions occur with the next biggest clusters but formed respectively of Unknown (524) and small PersonGroup (325) elements. Meaning that people take north doors as main exit; sometimes groups of persons are formed in the flow of people but principally individuals are detected. This region (between gates and north doors) is the most visited of the scene.

Cluster labelled 4 is only made of persons (56 elements in the cluster) with the associated event stays at Gates. They have been detected from about 6 to 15 seconds and by walking they are covering a distance between 60 cm and 2 m. The associated trajectory type is 14 (From north doors to gates). Thus, people entering the gates from north doors rarely have to wait at the gates to enter (relative small number of elements included in the cluster). This, however, may happen as ‘gates – north doors’ and vice versa is the main people path.

Other example of cluster activity is given by cluster labelled 3, made only of Luggage elements (272 items) detected for a very short period of time ([0.04 s – 5 s]) and mainly associated with trajectory type 2 ‘stationary at the gates’ (shown in figure 9) meaning also that it is at the gates that most frequently people leave their luggage but only for a very short period of time. One last example of the Torino data is, for instance, Cluster 7, made up only of persons with associated significant event ‘inside_zone_ hall’ and mainly trajectory type 25 ‘from gates to south doors’ (see figure 9). People following this path indeed walk longer to go through the south doors and thus are longer inside the hall.

Because no ground-truth is available for the Torino data, it is not possible to perform the same evaluation of the clustering that we did for the CAVIAR dataset based on the calculation of True Positives (TP) and False Positives (FP). However, to have a quantitative measure that indicates the number of activities correctly detected by our clustering algorithm, we have considered True Positive detection, a cluster whose data correspond mainly after visual inspection by end-user to a clear and meaningful description as for the clusters just presented above. False Positive is then defined as a cluster whose description cannot be associated to any coherent activity. In this sense, we had 20/27 clusters as True Positives (activities corresponding to 90% of the detected objects) and 7/27 clusters as False Positives.

For the Roma dataset, the biggest cluster, contained 1672 elements. It is made only of small Person Groups detected inside the zone hall. They cover a distance of 50 cm to 1 m in 1 to 3 seconds. This means that although groups of persons appear very often from north doors towards the hall (trajectory type 20; also shown in figure 10), they are able to walk at a normal speed. Another example for the Roma station is given by cluster labelled 6. This cluster (694 elements) is made up of only persons with associated significant event ‘outside zone hall’. Indeed trajectory type 7 (see figure 10) indicates that people are leaving the hall through the south doors. They cover a distance 1.67 m to 2.30 m in about 3 to 4 seconds, which is a relative slow speed. South doors are not as busy as North doors where numerous Groups were detected also leaving the hall (Cluster 5 (not shown) with 971 small groups detected).

For the Roma dataset, we have again considered True Positive detection as a cluster whose data correspond to a clear and meaningful description, and False Positive as a cluster whose description cannot be associated to any coherent activity. In this sense,

we had 72/109 clusters (activities) as True Positives and 37/109 clusters as False Positives.

Thus, this way the relational analysis can help us to group together people having similar behaviour. This is of particular interest to the end-users because the activities in the metro station can be better quantified.

In order to asses the impact of the object detection and tracking on the trajectory clustering step and how much the trajectory clustering process would then affect the results of the relational analysis, we evaluated our system on four different sets of tracked objects built from three hours of observation of the Roma station. The first dataset (Dataset1: 2983 trajectories) contains all objects detected and tracked on this observation period including noisy data and fragmented trajectories. The next dataset (Dataset2: 2605 trajectories) contains all objects minus trajectories of short duration most likely to represent noise in the data. The third dataset (Dataset3: 1713 trajectories) is Dataset2 minus all trajectories where the track for the last part of the trajectory was lost. Dataset4 (1713 trajectories) is Dataset 3 minus all trajectories where the track was lost at some point and then continued. The lost part of the trajectory is then inferred. We evaluated the impact of the different trajectory datasets on the trajectory clustering tool by measuring the Silhouette, Dunn and Davis Bouldin indexes. To evaluate how the relational analysis process is affected we calculated the resulting clustering quality as explained in section 6. These results are shown in table 8. As it can be observed, the quality of the trajectories extracted by the object detection and tracking module has an influence on the clusters obtained by the

trajectory clustering module. The less noise and lost tracks in the original dataset, the better the partition of the trajectory clusters (the Dunn index monotonally increases with the trajectory quality. The Silhouette index generally increases and the Davis Bouldin index generally decreases). This influence is however less propagated into the relational analysis as the quality of the final activity clusters remains rather constant.

Regarding our implementation, one hour of video takes approximately one minute to be processed off-line by the trajectory clustering, statistical and relational analysis. This is a reasonable processing time in adequacy with end-user requirements.

Input	Number of trajectories	Trajectory performance indexes			Relational analysis performance indexes	
		Silhouette	Dunn	Davis Bouldin	Clustering quality	Resulting nuber of clusters
Dataset1	2983	0.314	0.049	19.571	70%	35
Dataset2	2605	0.254	0.056	13.79	69%	46
Dataset3	1713	0.325	0.071	17.632	69%	42
Dataset4	1476	0.343	0.091	17.854	69%	39

Table 8. Evaluation of the impact of the object detection and tracking on the trajectory clustering and relational analysis processes.

8 Discussion and conclusion

In this paper we have presented how knowledge discovery can be achieved on large recordings of video using an efficient knowledge representation format. The richness

of the representation comes from the fact that both, moving objects and the contextual objects from the scene are studied together with their interaction. Yet, the proposed representation provides a useful support and enables all activity knowled ge to be structured into three different appropriate tables, namely mobile objects, contextual objects and video events. The proposed representation supports a rich set of spatial topological and temporal relations and captures not only quantitative properties but also higher semantic concepts. Furthermore, a first layer of meaningful knowledge is directly extracted from the video streams by detecting events corresponding to the interactions between moving objects and contextual objects. A second layer of knowledge is extracted by the off-line long term analysis of these interactions. First, statistical information is obtained from the mobile objects (in particular their trajectory) and the contextual objects as well as their interactions (*i.e.* events). Statistical information is a major information source for the end-user. For instance, on large metro video recordings the average number of people localized in specific zones of interest or interacting with particular equipment and its evolution along the time provide an operational information on how to manage the metro station on a day by day basis. We are currently analysing sequentially chunks of video of about two hours and then will further analyse the temporal evolution for durations such as one day or one week. On a second step, we perform the trajectory characterisation by employing a hierarchical algorithm. It must be remarked that there is an scalability issue when employing standard hierarchical algorithms, and these may not work efficiently for very large datasets. An alternative solution is to implement, for instance, parallel hierarchical clustering algorithms [45]. However, to solve the main problem concerning accesing and processing a larger number of stored data, we are currently working on a new version where we will be able to update on-line the clusters for continous processing. In this new version also other

features such as duration, mean speed and distance walked are currently being studied in the clustering of trajectories.

The semantic knowledge gained from trajectory characterisation and statistical analysis is then used for the discovery of complex relationships. The relational analysis proposed in this paper shows how to highlight hidden relations between people, their trajectories (behavioural information) and the significant interactions between themselves and contextual objects. Thus, relational analysis can define the typical activities in the subway represented as activity clusters.

The performance evaluation has been performed at the knowledge discovery level by providing manually ground-truth trajectories in two sites. In a building hall for the CAVIAR project and in the Torino metro hall for the CARETAKER project. Performance evaluation is a sensitive point in the knowledge discovery due to the large number of parameters to employ and the small description on expected results to be provided by the end-users. In opposition to what is usually done in trajectory and activity clustering, we have proposed a new framework for evaluation. With this framework we have been able to asses the system. These results are however to be taken with care as the size of the data analysed is small. This unfortunately has been an endemic problem in the area of computer vision, where annotated datasets (with an evaluation ground-truth) are difficult to find. More work is still to be done, not only for evaluating more hours of video, but also, for taking into account other types of parameters such as spatial and temporal granularity of the scene.

Figures

Figure 7. Properties for cluster 5 in the CAVIAR data partition.

Figure 8. Trajectories detected in one station of the Torino underground (2025 trajectories during 50 minutes). Beginning of the trajectories are indicated by a red point.

Figure 9. Clusters showing the most common paths obtained from the dataset shown in figure 7. (a) Cluster 14 with 443 trajectories; (b) Cluster 1 with 744 trajectories; (c) Cluster 25 with 50 trajectories; (d) Cluster 2 with 338 trajectories; (d) Cluster 26 with 41 trajectories; (e) Cluster 5 with 13 trajectories. Beginning of the trajectories are indicated by red points. End of trajectories are indicated by blue points.

Figure 10. Clusters showing the most common paths obtained from the Roma dataset. (a) Whole dataset; (b) Cluster 20 with 1451 trajectories; (c) Cluster 19 with 809 trajectories; (d) Cluster 7 with 790 trajectories. Extremity points of the trajectories are represented as in figure 9 by red and blue points.

Figure 11. Temporal evolution in the number of people occupying the station hall in the Torino station.

Figure 12. Temporal evolution of the mean time a user spends on a vending machine in the Torino station hall.

Tables

[1] Moeslund, T., Hilton, A., Krüger, V.: 'A survey of advances in vision-based human motion capture and analysis', Computer Vision and Image Understanding, 2006, 104, (2-3), pp. 90-126

[2] Zhang, H., Kantankanhalli, A., Smoliar, S.: 'Automatic Partitioning of Full-Motion Video', ACM Multimedia Systems, 1993, 1, (1), pp. 10-28

[3] Tong, S., Chang., E.: 'Support Vector Machine Active Learning for Image Retrieval', Proc. ACM Multimedia, Ottawa, Canada, 2001, pp. 107–118

[4] Chong-Wah, N., Ting-Chuen, P., Hong-Jiang Z.: 'Video Processing: On clustering and retrieval of video shots', Proc. 9th ACM int. conf. on Multimedia MULTIMEDIA '01, 2001

[5] Ewerth, R., Freisleben, B.: 'Semi-supervised learning for semantic video retrieval', Proc. 6th ACM int. conf. on Image and video retrieval CIVR '07, 2007

[6] Smeaton, A. F.: 'Techniques used and open challenges to the analysis, indexing and retrieval of digital video', Information Systems, 2007, 32, (4), pp. 545-559

[7] Le, T.-L., Boucher, A., Thonnat, M. : 'Subtrajectory-based video indexing and retrieval', Proc. 13th Int. Conf. on Advances of Multimedia Modelling, 2007, 1, pp. 418-427

[8] Velastin, S.A., Boghossian, B.A., Lai Lo, B. P., Sun, J., Vicencio-Silva M.A.: 'PRISMATICA: Toward Ambient Intelligence in Public Transport Environments', IEEE Trans. Syst. Man Cy. A, 2005, 35, pp. 164-182

[9] Piater, J., Richetto, S., Crowley, J.: 'Event based activity analysis in live video using a generic object tracker'. In: Freyman, J. (Ed.) Proc. 3rd IEEE Workshop on performance evaluation of tracking and surveillance (PETS), Copenhagen, Denmark, 2002

[10] Cupillard, F., Brémond, F., Thonnat, M.: 'Automatic Visual Recognition for Metro Surveillance'. Proc. Int. Conf. Measuring Behavior, Wageningen The Netherlands, 2005

[11] Marcotorchino, F., Michaud, P.: Optimisation en analyse ordinale des données, Masson, 1978

[12] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

[13] Patino, J.L, Benhadda, H., Corvee, E., Bremond, F., Thonnat, M.: 'Video-Data modelling and Discovery', IET Int. Conf. on Visual Information Engineering VIE'07, 2007, 9 pages

[14] Georis, B., Bremond, F., Thonnat, M., Macq, B.: 'Use of an Evaluation and Diagnosis Method to Improve Tracking Performances', Proc. 3rd IASTED Int. Conf. on Visualization, Imaging and Image Proceeding (VIIP), 2003, 2

[15] Wang, L., Hu, W., Tan, T.: 'Recent developments in human motion analysis', Pattern Recognition, 2003, 36, pp. 585-601

[16] Fusier, F., Valentin, V., Brémond, F., Thonnat, M., Borg, M., Thirde, D., Ferryman, J.: 'Video Understanding for Complex Activity Recognition', In Machine Vision and Applications Journal , 2007, 18, pp. 167-188

[17] Cupillard, F., Brémond, F., Thonnat, M.: 'Tracking Group of People for Video Surveillance', In IEEE Proc. of the 2nd European Workshop on Advanced Video-Based Surveillance System AVBS'02, 2002

[18] Avanzi, A., Bremond, F., Tornieri, C., Thonnat, M.: 'Design and assessment of an intelligent activity monitoring platform', EURASIP, 2005, pp. 2359-2374

[19] Avanzi, A., Brémond, F., Thonnat, M.: 'Tracking Multiple Individuals for Video Communication', In IEEE Proc. of International Conference on Image Processing, 2001

[20] Vu, V.T., Bremond, F., Thonnat, M.: 'Automatic video interpretation: A novel algorithm for temporal scenario recognition', Proc. 18th Int. Joint Conf. on Artificial Intelligence. IJCAI'03, 2003, pp. 1295-1302

[21] http://www.cvg.rdg.ac.uk/PETS2006/index.html

[22] Piciarelli, C., Foresti, G. L., Snidaro, L.: 'Trajectory Clustering and its Applications for Video Surveillance', IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, AVSS'05, 2005

[23] Anjum, N., Cavallaro, A. : 'Single camera calibration for trajectory-based behavior analysis', IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, AVSS'07, 2007

[24] Naftel, A., Khalid, S.: 'Classifying spatiotemporal object trajectories using unsupervised learning in the coefficient feature space', Trans. of Multimedia Systems, 2006, 12, (3), pp 45 – 52

[25] Antonini, G., Thiran, J. P.: 'Counting pedestrians in video sequences using trajectory clustering', IEEE Trans. on Circuits and Systems for Video Technology, 2006, 16, (8), pp 1008 – 1020

[26] Calderara, S., Cucchiara, R., Prati, A. : 'Detection of Abnormal Behaviors using a Mixture of Von Mises Distributions', IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, AVSS'07, 2007

[27] Gaffney, S., Smyth, P.: 'Trajectory clustering with mixtures of regression models', Proc. Int. Conf. on Knowledge Discovery and Data Mining, 1999, California, USA

[28] Porikli, F.: ‘Learning object trajectory patterns by spectral clustering’, Proc. IEEE Int. Conf. on Multimedia and Expo ICME '04, 2004, 2, pp. 1171 – 1174

[29] Bashir, F. I., Khokhar, A. A., Schonfeld, D.: ‘Object Trajectory-Based Activity Classification and Recognition Using Hidden Markov Models’, IEEE Transactions on Image Processing, 2007, 16, (7), pp. 1912 – 1919

[30] Oliver, N.M., Rosario, B., Pentland, A.P.: ‘A Bayesian computer vision system for modeling human interactions’, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22, (8), pp. 831 – 843

[31] Jian, F., Wu, Y., Katsaggelos, A., ‘Abnormal Event Detection from Surveillance Video by Dynamic Hierarchical Clustering’, ICIP 2007

[32] Xiang, T., and Gong, S.: ‘Video behaviour profiling and abnormality detection without manual labelling’, In Proc. IEEE International Conference on Computer Vision, 2005, pp. 1238-1245

[33] Xiang, T., Gong, S.: ‘Incremental and adaptive abnormal behaviour detection’, Computer Vision and Image Understanding, in press, 2008

[34] Toshev, A., Brémond, F., Thonnat, M.: ‘An A priori-based Method for Frequent Composite Event Discovery in Videos’,. In Proceedings of 2006 IEEE International Conference on Computer Vision Systems, 2006

[35] Hartigan, J.A.: ‘Clustering Algorithms’, New York, John Wiley & Sons, 1975

[36] Kohonen, T.: ‘Self-Organizing Maps’, New York, Springer-Verlag, 2001

[37] Kaufman, L., Rousseeuw, P.J., ‘Finding Groups in Data. An Introduction to Cluster Analysis’, New York, Wiley, 1990

[38] Campello, R.J.G.B., Hruschka, E.R., ‘A fuzzy extension of the silhouette width criterion for cluster analysis’, Fuzzy Sets and Systems, 2006, 157, pp. 2858-2875

[39] Dunn, J., ‘Well separated clusters and optimal fuzzy partitions’, Journal of Cybernetics, 1974, 4, pp. 95-104.

[40] Davies, D.L., Bouldin, D.W., ‘A cluster separation measure’, IEEE Transactions on Pattern Recognition and Machine Intelligence, 1979, 1, pp. 224-227

[41] Benhadda, H., Marcotorchino, F. : ‘L'analyse relationnelle pour la fouille de grandes bases de données’, Revue des Nouvelles Technologies de l'Information, 2007, RNTI-A-2, pp. 149-167

[42] Benhadda, H. : ‘La similarité régularisée et ses applications en classification automatique’, PhD thesis, PARIS VI University, 1998

[43] Benhadda, H., Marcotorchino, F. : ‘Introduction à la similarité régularisée en analyse relationnelle’, Revue de Statistique Appliquée, 1998, 46, (1), pp. 45-69

[44] Marcotorchino, F.: ‘Seriation Prolem : An overview’, AppliedStochastic models and data analysis, 1991, 7, pp. 139-151

[45] Rajasekaran, S.: ‘Efficient parallel hierarchical clustering algorithms’ IEEE Transactions on Parallel and Distributed Systems, 2005, 16, pp. 497-502