Semantic Classification of Human Behaviors in Video Surveillance Systems

Alberto Amato, Vincenzo Di Lecce
DIASS
Politecnico di Bari
Via Alcide De Gasperi
Italy
{a.amato, v.dilecce}@aeflab.net http://www.aeflab.net

Abstract: The semantic analysis of the human behavior in video streaming is still an open issue for the computer vision research community, especially when real-time analysis of complex scenes is concerned. The researchers' community has achieved many progresses in this field. A popular class of approaches has been devised to enhance the quality of the semantic analysis by exploiting some background knowledge about scene and/or the human behavior, thus narrowing the huge variety of possible behavioral patterns by focusing on a specific narrow domain. Aim of this paper is to present an innovative method for semantic analysis of human behavior in video surveillance systems. Typically, this kind of systems are composed of a set of fixed cameras each one monitoring a fixed area. In the proposed methodology, the actions performed by the human beings are described by means of symbol strings. For each camera a grammar is defined to classify the strings of symbols describing the various behaviors. This system proposes a generative approach to human behavior description so it does not require a learning stage. Another advantage of this approach consists in the simplicity of the scene and motion descriptions so that the behavior analysis will have limited computational complexity due to the intrinsic nature both of the representations and the related operations used to manipulate them. This methodology has been used to implement a system to classify human behaviors in a scene. The results are discussed in this paper and they seem to be encouraging.

Key-Words: - human behavior analysis, grammar based approach, semantic analysis of video streaming, video surveillance systems, generative human behavior description.

1 Introduction

In the latest years, information and communication technology (ICT) has had a strong improvement having significant influence on everyday life. The effects of the technological improvements in the fields of video sensor manufacturing and communication networks are particularly relevant. Indeed, they have also enabled and booted the development of complex video surveillance systems using network technologies that are able to monitor wide areas where human beings perform their activities. The main drawback of this kind of systems is that they produce a huge quantity of not structured data (video streaming). Nowadays, these data are stored into various storage systems and they are used only to see the record of a scene where some relevant actions have occurred. Furthermore, there are video surveillance systems using human operators for real time analysis of recorded scenes. The main drawbacks of the latter approach are that it is not cost effective and the low human efficiency in this task (namely, monitoring simultaneously different the video sampled by a large number of cameras).

On the other hands, the interest of international scientific community is moving towards video surveillance systems implementing real-time automatic human behavior analysis. This application area has a relevant interest also due to the facts characterizing the history of the last decade. In particular, the aim of the applications in this area is monitoring and understanding human behavior in public and crowded areas (such as streets, bus and train stations, airports, shopping malls, sport arenas, and museums). These applications can be used both for management (for example: evaluation of crowding in the monitored areas, human flow analysis, and detection of the congestion points) and security (for example: human behavior analysis, activity recognition of individuals and groups, and detection of suspicious persons).

The economic and social relevance of potential applications (especially the security, entertainment, and medical ones), the scientific complexity, the

speed and price of current hardware intensified the effort within the scientific community towards automatic capture and analysis of human motion.
Automatic understanding of the human behavior from video sequences is a very challenging problem since it implies understanding, identifying, and either mimicking the neuro-physiological and psychological processes, which are naturally performed in humans or creating similar outcomes by means of appropriate information and knowledge processing.
Currently, there is not a comprehensive and universally valid method to implement systems for automatic understanding of human behavior from video sequences. To solve this problem or at least to reduce the effects of the semantic gap, researchers have been working on exploiting some knowledge about scene and/or the human behavior, thus narrowing the huge variety of possible behavioral patterns by focusing on a specific narrow domain.
In this framework, this work proposes a new approach to analyzing and understanding of human behavior by using a regular grammar to describe the recorded actions. This allows for high simplicity in the scene and motion descriptions so that the behavior analysis will have limited computational complexity, thanks to the intrinsic nature both of the representations and the related operations used to manipulate them. On the other hand, this approach has a great flexibility because it uses a generative approach to scene description and it does not require a learning stage.
To show the effectiveness of the proposed approach, a demonstrative system has been implemented and applied to a publically available video database: the "Edinburgh Informatics Forum Pedestrian Database" [1]. On this database a set of experiments were carried out to verify the effectiveness of the proposed system. In particular, the system was used to classify the various behaviors recorded into the database. The obtained results seem to be encouraging.
The remaining part of this paper is articulated in the following way: section 2 reports a brief literature overview, section 3 presents the proposed methodology while in section 4 some experiments carried out to evaluate the effectiveness of the proposed methodology are described. Conclusions and final remarks are reported in section 5.

2 Related works

Even though in literature there are many works on human behavior analysis and recognition, this is still an open research field. This is due to the inherent complexity of such task. Indeed, human behavior recognition can be seen as the vertex of a computational pyramid as shown in Fig. 1.

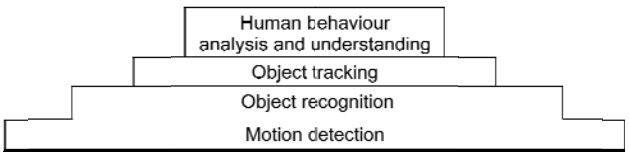

Fig. 1 - A hierarchical overview of the computational chain for human behaviour recognition

Each level of this processing pyramid has its own characteristics and difficulties often due the partial completeness of the used data (for example: the tentative to extract 3D data about moving objects working on 2D images).

Since this work focuses on human behavior analysis and recognition, this section overviews the approaches about this problem proposed in literature.

A common accepted way to classify these works is the following one:

- **scene interpretation**: these works try to interpret the whole image without identifying particular objects or humans; typically, these systems have a learning stage (where the observed trajectories of the moving objects are classified using various methods to build the system knowledge base) and an operating stage (where the knowledge base is used to classify the runtime observed object trajectories).

  These systems are used in frameworks where a well defined set of situations are allowed while others are forbidden or in applications where the goal is to classify the scenes/behaviors as usual or unusual ones. At this level of visual abstraction, the moving objects are considered in their completeness. This allows for building systems that analyze at a high level the objects trajectories and their interactions. In literature there are many works using this approach. For example, in [2] an approach is proposed to classify scenes in usual and unusual ones by starting from the analysis of a single frame or object rather than from a sequence of frames. This method is tested in various environments both indoor and outdoor. It is designed to work on huge datasets, indeed, the results presented in the paper are obtained on a database composed of 4 years of continuous video acquisition (that give millions of records). Despite the fact that this work is quite old, it still remains a key work

because it represents clearly the potentiality of the methods using scene interpretation.

A second class of approaches for detecting anomalies in video sequences and/or in single images has been proposed in [3]. This method tries to decompose the normal (or the allowed) video sequences and/or images in small multi-scale portions defining the knowledge base of the system. When the system analyzes a query video and/or image, it tries to rebuild the query using the portions stored in the knowledge base. The regions of the query that can be rebuilt using large portions from the database are considered as "normal" while those that can not be rebuilt or that can be rebuilt using small and fragmented portions are regarded as "suspicious".

A third class of approaches is based on the study of the trajectories of the moving objects and focuses the attention both on the geometric characteristics of the trajectories and on their cinematic aspects. The idea standing at the base of these works is that in a given place, similar trajectories can be associated to similar activities. In this way, analysing and classifying the trajectories described by a moving object is equivalent to analyze and classify its activities. For example, [4] proposes a method to distinguish between objects traversing spatially dissimilar paths, or objects traversing spatially proximal paths but having different spatial-temporal characteristics.

- **human recognition**: the works falling in this class try to infer information about the human beings activities analyzing the dynamic of their movements. Some of these works try to recognize and study the motion of the individual body parts while others consider the whole body as a unique element. An example of this kind of system is [5]. Here the authors propose a hierarchical approach to implement a system for human behavior analysis in the narrow domain (tennis match). Here, a given human behavior is considered as composed of a stochastic sequence of actions. Actions are described by a feature vector comprising both trajectory information (position and velocity), and a set of local motion descriptors. Action recognition is achieved by means of a probabilistic search method applied to the database representing previously seen actions. The possible actions are modelled by means of Hidden Markov Models (HMM), high-level behavior recognition is achieved by computing the likelihood that a set of predefined Hidden Markov Models explains the current action sequence.

Other approaches are based on the concept of "*temporal templates*" (a static vector-image where the vector value at each point is a function of the motion properties at the corresponding spatial location in an image sequence). This idea was proposed in [6] where the authors used a two-component version of the templates: the first value is a binary value indicating the presence of motion and the second value is a function of the recency of motion in a sequence. These components are called MEI (Motion-Energy Image) and MHI (Motion-History Image) respectively. MEIs are cumulative binary motion images, namely, binary images where the value of each pixel is set to 0 if its value does not change for each frame of a given sequence (namely, no moving objects have passed over it) while it is set to 1 otherwise. MHIs are grey scale images where pixel intensity is a function of the temporal history of motion at that point. In these images, the more recently moving pixels are brighter. From a certain point of view, considering a given scene, the MEI image describes *where* the motion occurs while the MHI describes *how* it occurs. Matching temporal templates is based on Hu moments [7].

Another approach is that of ''*Actions Sketches*'' or ''*Space-Time Shapes*'' in the 3D XYT volume. In [8] the authors propose to model an action based on both the shape and the motion of the object performing the action. When the object performs an action in 3D, the points on the outer boundary of the object are projected as 2D (x, y) contour in the image plane. A sequence of such 2D contours with respect to time generates a spatiotemporal volume (STV) in (x, y, t), which can be treated as 3D object in the (x, y, t) space. The differential geometric surface properties of this 3D object (such as peaks, pits, valleys and ridges) are considered specific action descriptors capturing both spatial and temporal properties. For example, a pit surface is generated when the contour first moves in the direction that is normal to the contour, then stops and moves in the opposite direction.

A critical element for this kind of systems is the fact that they are substantially view dependent also if the dynamic regularities features that they use are not view dependent.

An attempt to create a view invariant system has been done in [9]. Here the authors, starting from

the findings in [10] (where the authors developed relationships between six-tuple 3D points and their corresponding image coordinates that are satisfied for all views of the 3D points), propose a 3D approach aiming at viewpoint invariance.

- **action primitives and grammars**: works falling into this class attempt to decouple actions into action primitives and interpret actions as a composition on the alphabet of these action primitives. A method employing techniques from the dynamical systems framework is presented in [11] where the authors propose to decompose a human activity into a set of elementary actions. These elementary actions can be seen as symbols of an "alphabet" and so, they can be used to describe human motions similar to the way phonemes are used in speech. They call these primitives of motion "movemes". By using system identification techniques and pattern recognition techniques they develop an on-line joint segmentation and classification algorithm and provide analytical error analysis. Once that the primitives are detected, an iterative approach is used to find the sequence of primitives for a novel action.
  A vision based approach is proposed in [12] where the authors propose to describe human actions in terms of *action units* called "*dynamic instants*" and "*intervals*" which can be computed studying the spatio-temporal curvature of a 2-D trajectory. The *dynamic instants* are due to changes in the forces applied to the object during the activity. They are perceived as a change in the direction and/or speed and can be reliably detected by identifying maxima in the spatio-temporal curvature of the action trajectory. An *interval* is the period of time between two dynamic instants during which the motion characteristics do not change. The authors formally show that the *dynamic instants* are view-invariant, except for the limited cases of accidental alignment.
  In [13] the authors propose a system performing a hierarchical analysis of a video stream. The lowest level analyzes the poses of individual body parts including head, torso, arms and legs are recognized using individual Bayesian networks (BNs), which are then integrated to obtain an overall body pose. The middle level models the activity of a single person using a dynamic Bayesian network (DBN). The higher level of the hierarchy works on the results of the mid-level layer. Here, the descriptions for each person are juxtaposed along a common time line to identify an interaction between two persons.
  Working on the same concept of multi level analysis, where at the lower levels the action primitives are recognized and sent at the higher levels to perform a more complex analysis, in [14] the authors propose to use a Stochastic Context Free Grammar (SCFG) to obtain a high semantic level analysis of human behavior. In this work, the authors propose a probabilistic approach to the analysis of temporally extended actions encompassing also the problem of interactions among moving objects. The system is composed of two levels. The first level detects action primitives using standard independent probabilistic event detectors to propose candidate detections of low-level features. The outputs of these detectors are sent as input stream to the second level. Here a stochastic context-free grammar parsing mechanism is used to analyze the stream and perform a higher semantic level analysis. The main advantages of this approach are that it provides longer range temporal constraints, disambiguates uncertain low-level detections, and allows the inclusion of a priori knowledge about the structure of temporal events in a given domain.
  An interesting aspect of this method is the use of the grammar as a convenient means for encoding the external knowledge about the problem domain, expressing the expected structure of the activity.
  The main limit of this approach is the fact that it uses low-level features detectors that are able to model and recognize only a fixed number of action primitives.
  A more comprehensive literature analysis of this research field is presented in [15] and [16].

3 Proposed system

The kernel of the proposed system is based on the idea at the base of the scene interpretation systems that try to interpret the scene studying the trajectories of some relevant points of the moving objects (their barycentres). The idea behind this kind of approaches is the mapping between trajectories and behaviours. The ratio can be synthesized into the observation that in order to accomplish a given task one must follow a prefixed series of movements.
In this work, the trajectories are represented by means of string of symbols (the alphabet of the domain specific grammar) each one having a semantic value into the specific domain.

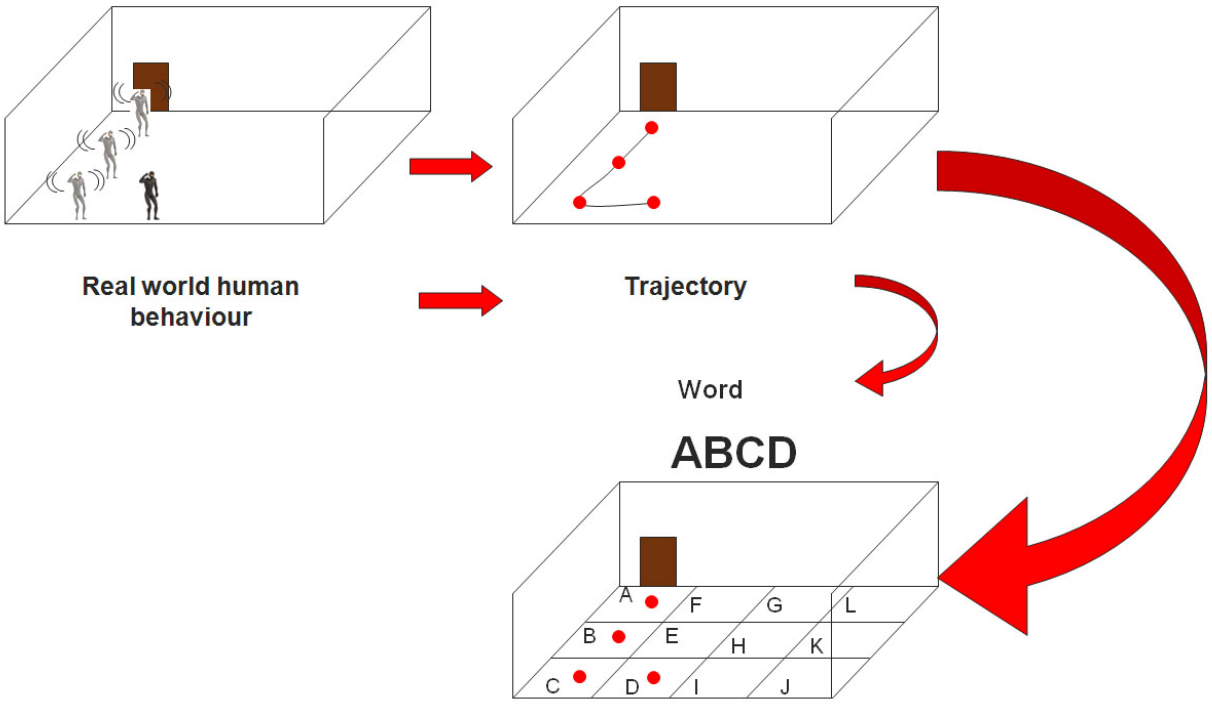

Fig. 2 - An example of domain conversion. The real world action (a) is "translated" into a curve in the domain of the trajectories (b). The trajectory is "translated" into a word into the linguistic domain (c)

The mapping among symbols and the semantic meaning of areas of the scene is done manually at design time. For each domain of application, a grammar is defined in order to specify the recognizable sentences of that domain. In this way it is possible to define a correspondence between the

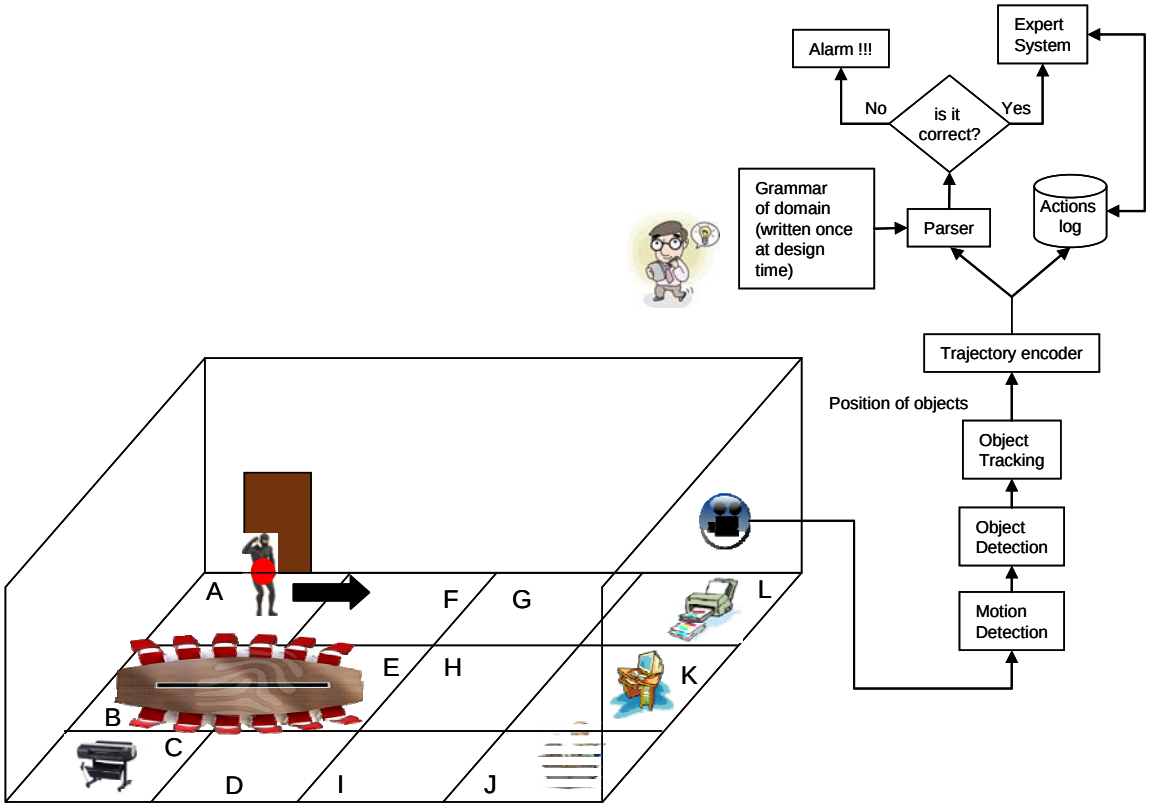

Fig. 3 - A schematic overview of the proposed system

set of the sentences writable with this grammar and the set of allowed actions in the scene.
Since the mapping among portions of scene and symbols/semantic meanings and the grammar are specific for each application, this system belongs to the class of narrow domain systems.
This approach introduces a domain switching for the problem of trajectories analysis. Indeed, by labelling the environment, it is possible to "translate" the geometric data about the trajectories into words (Fig. 2). In this way, studying the characteristic of a word means to study the geometric characteristic of a trajectory. So the geometric analysis becomes a linguistic problem.
From this perspective, the problem changes its appearance. The issue of understanding which behaviours (and so which trajectories) are allowed in a given environment becomes the issue of understanding which words one can write using the symbols (labels) defined for that environment.
This problem can be faced defining a specific grammar for each environment. This approach gives a strong flexibility and reliability to the proposed methodology. Indeed, in this context, defining a grammar means to define the utilizable rules to write the words describing the behaviours. In this way, this methodology inherits one of the most interesting characteristics of the language theory: the possibility of defining infinite set of words (behaviours) starting from a finite set of symbols (the labels used to describe the domain of interest).
Fig. 3 shows a schematic overview of the proposed approach for human behaviour analysis. In the bottom left part of the figure, a schematic example of an indoor environment is presented. At design time, the environment is virtually divided into a certain number of areas and each area is labelled with a symbol (in this work, letters of the English alphabet have been used). It is possible to define areas of whatever shapes and dimensions. This fact allows for a more accurate definition of the areas around some particularly relevant points, namely the areas where it is possible to attribute specific semantic meanings (i.e., see the area around the printer in Fig. 3). The dimension of the areas should be coherent with the scale of observation of the environment and hence with the resolution of the video streaming. In other words, the partition should be defined in order to obtain a good resolution in the successive stage of string generation. Each trajectory should be represented by a string of symbols with at least a symbol for each semantic area that it crosses.
The stream sampled by the camera follows the chain:
motion detection → object detection → trajectory encoder → high level analysis
The tracking algorithm uses a string for each detected moving object (see the module "trajectory encoder" in Fig. 3). The process of strings generation and update is handled by the "trajectory encoder" module. It evaluates the coordinates of a given moving object and appends the symbol with which the relative area is labelled to the string.
In order to consider the time, the *trajectory encoder* has a time driven behaviour. It generates a symbol ever T seconds. The value of T is a constant for the system and it must be chosen at design time according to the dynamic of the analyzed environment. On the other hand, this parameter is not too critical from a computational point of view, because, thanks to the kind of grammar used to generate the language describing the behaviours, the string analysis process has a linear time complexity.
From a theoretical point of view, it should be noticed that, using this procedure, it is possible to generate any kind of string. On the other hand, the real world phenomenon under analysis, (i.e., the motion of a man into a room) has its physical constraints. For example, looking at Fig. 3, for the principle of continuity of motion, when the man is in the area "*D*", he can not go directly into the area "*J*". He should follow a path through the area "*I*" or a longer path through the other neighbourhood areas. Furthermore, it is possible that in a real world case there are other constraints. For example, in the case shown in Fig. 3 there is a desk in the areas "B" and "E". In this case, no one can walk on these areas. Other possible constraints can be introduced according to the analyzed scene.
Using the proposed methodology, the respect of these constraints has a straightforward implementation. Indeed, thanks to the double context switching from real world motion to trajectory and from trajectory to string, this methodology allows to face this problem as a linguistic one.
In this context, by defining a grammar it is possible to evaluate if a given *string* belongs to a given language (and so it is correct) or not.
The "expert system" (Fig. 3) analyzes the strings generated by the trajectory encoder to infer various information about each action, for example:

1) if a word belongs to the grammar (namely, it is correct) it is describing an allowed behaviour. Furthermore, it is possible to say "what" behaviour and thus to obtain an high semantic

level description of that behaviour by means of successive labelling operations.

2) if a word does not belong to the grammar (namely, it is not correct) it is describing a forbidden behaviour. Analysing this kind of strings it is possible to infer other information about the human behaviour. For example, if a string contains a symbol that is not present in no one rule of the grammar it means that it is describing a forbidden behaviour. But, if a string is describing a not-continuous path, it means that the object tracking algorithm lost the moving object. From this point of view, this grammar can be used as a system to recover the trajectories of moving objects in crowding scenes.

4 Experiments and results

The proposed system was tested using a publically available video database: the "Edinburgh Informatics Forum Pedestrian Database" [1]. This database is composed of a set of detected targets of people walking through the Informatics Forum, the main building of the School of Informatics at the University of Edinburgh. The data were sampled for several months and there are about 1000 observed trajectories each working day. By July 4, 2010, there were 27+ million target detections, of which an estimated 7.9 million were real targets, resulting in 92,000+ observed trajectories. From a functional point of view, this database can be seen as the output of a video surveillance system with a fixed camera.

Fig. 4 shows a view of the scene and image data from which the detected targets are found. The main entry/exit points are marked with arrows. They are placed at the bottom left (front door), top left (cafe), top center (stairs), top right (elevator and night exit), bottom right (labs). In the database there are also some false detections due to noise, shadows, reflections, etc. Normally, only about 30% of the captured frames contain a target and normally there are only a few targets in each frame (1 target in 46% of active frames, 2:25%, 3:14%, 4:8%, 5:4% 6-14:3% of time). The videos are recorded by a fixed camera overhead approximately 23m above the floor. The distance between the 9 white dots on the floor is 297 cm vertically and 485 cm horizontally. The images are 640x480, where each pixel (horizontally and vertically) corresponds to 24.7 mm on the ground.

The database does not contain the raw image data but a set of pre-computed files describing the trajectories of the detected moving objects. In the database there is a file for each monitoring day.

Fig. 4 - A view of the scene and image data from which the detected targets are found

In order to test the proposed system, the trajectories detected on 24 August have been used. On this day, 664 moving objects (people) have been detected and their trajectories are stored in a single file [17].

On these data, the proposed system has been used to classify the detected behaviors.

The scene has been partitioned and labeled as shown in Fig. 5. It is possible to notice that the scene has been partitioned using areas of different sizes. Each area has an homogeneous semantic meaning, for example, the area labeled with the letter "B" is a transit area while the area "C" represents the gateway to the lifts. Each recorded trajectory has been translated into a word using the trajectories encoder. In this framework, for example, when the trajectories encoder produces a string like "NBBBBC", the expert system infers that a person is walking from the atrium area to the lifts.

In order to classify all the observed trajectories, a clustering algorithm has been applied on the dataset composed of all the words encoded by the trajectories encoder.

Clustering is a well-known technique used to discover inherent structure inside a set of objects. Clustering algorithms attempt to organize unlabeled pattern vectors into clusters or "natural groups" so that points within a cluster are more similar to each other than to points belonging to different clusters.

For each cluster, the centroid is computed and then it is used as "templates" to represent the entire cluster.

In literature, many clustering methods have been proposed [18], [19].

In this work the k-medoid algorithm has been used.

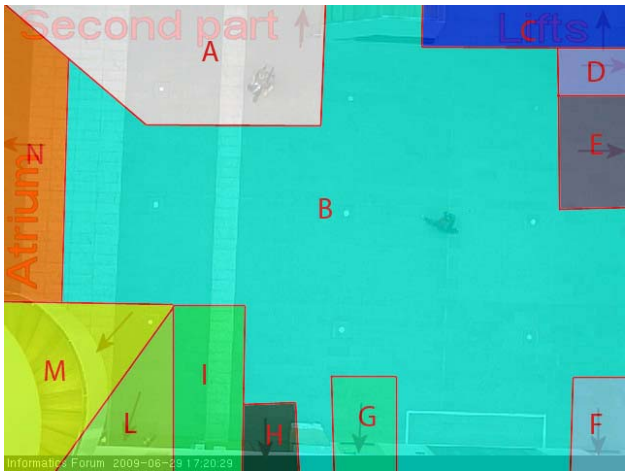

Fig. 5 – This figure shows the partition used to label the scene under analysis

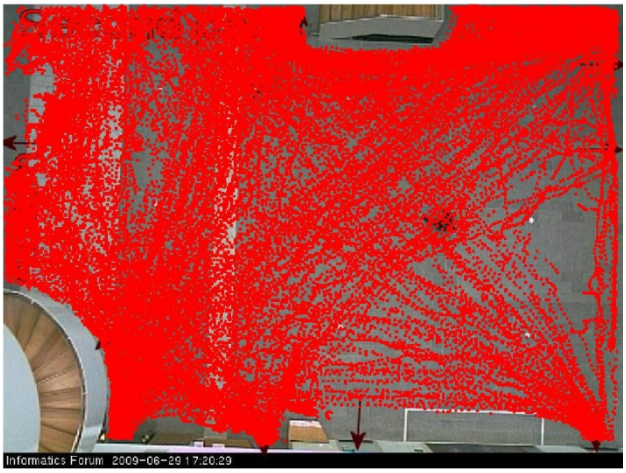
Fig. 6 – this figure shows a plot of all the 664 trajectories stored into the used dataset

K-medoid is a partitive clustering algorithm. As the well known k-means algorithm, the target of the k-medoid algorithm is to minimize the distance between all the points of a given cluster and the point designated to be its center. For both these algorithm the number of desired clusters is an input parameter. The main difference between k-means and k-medoid is the method used to compute the coordinates of the center of each cluster. Indeed, k-medoid uses as center a point belonging to the dataset while the k-means uses a virtual point given by the barycenter of the dataset points belonging to a given cluster.

In this work various experiments were carried out in order to find the optimal number of clusters to classify all the relevant trajectories/behavior recorded in this scene.

Fig. 6 shows a plot of all the detected trajectories stored into the dataset under analysis. These trajectories represent all the behaviors detected in a day of monitoring.

In order to find the most representative behaviors the k-medoid algorithm has been applied to this dataset many times changing the number of clusters. Fig. 7, 8 and 9 show the results obtained respectively using 8, 12 and 20 clusters.

These figures highlight the ability of the proposed method to represent in an effective way many different human behaviors. In particular, Fig. 9 shows that the proposed system is able to describe all the observed behaviors using only 20 words.

These experiments highlight the feasibility of using the proposed system as a video retrieval system working at high semantic level. In particular, it is suitable for all the videos recorded by the modern video surveillance systems.

Fig. 7 – a plot of the most frequent trajectories detected using the k-medoid algorithm to obtain 8 clusters

Fig. 8 - a plot of the most frequent trajectories detected using the k-medoid algorithm to obtain 12 clusters

Fig. 9 - a plot of the most frequent trajectories detected using the k-medoid algorithm to obtain 20 clusters

4 Conclusion

In this work an innovative system for high semantic level analysis of videos in the narrow domain has been presented.

The system operates a double context switch. The first one is from human motion to barycentre trajectory and it is a well known and accepted method in literature. The second one is from trajectory to word and is an original contribute of this work.

The external knowledge is introduced into the system by labelling with a set of symbols the various areas in which the scene is partitioned. Since in the narrow domain it is possible to attribute a semantic value at each area and thus at each symbol, using this methodology it is possible to achieve a high semantic level description of the recorded scenes.

This methodology uses a robust approach to the problem of strings/words recognition. Once that the environment has been labelled, a grammar on the set of symbols used to label the scene is defined. The production rules set of this grammar contains the rules to describe all the allowed behaviours in a given scenario. Using this grammar it is possible to define a language composed of the set of recognizable words (and thus the allowed behaviours).

It should be highlighted the fact that this methodology proposes a generative approach to human behaviour recognition. Defining a specific grammar *G* for a given domain, the system uses its rules to describe a human behaviour writing a word. The system is able to recognize all the behaviours that can be described using a word belonging to the language defined on the grammar *G*. This is a strong improvement in comparison to many works in literature that are able to recognize only a finite set of actions learnt in a training stage.

The proposed experiments and results show that this system is suitable also for high semantic video retrieval process. Indeed, in this framework, searching for a given word means searching for a given behaviour.

References:

[1] B. Majecka, "Statistical models of pedestrian behaviour in the Forum", *MSc Dissertation, School of Informatics*, University of Edinburgh, 2009

[2] C. Stauffer, W.E.L. Grimson, Learning patterns of activity using real-time tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (8) (2000) 747–757.

[3] O. Boiman, M. Irani, Detecting irregularities in images and in video, in: *Internatinal Conference on Computer Vision*, Beijing, China, Oct. 15–21, 2005.

[4] I.N. Junejo, O. Javed, M. Shah, Multi feature path modeling for video surveillance, in: *International Conference on Pattern Recognition*, Cambridge, UK, Aug 2004.

[5] N. Robertson, I. Reid, Behavior understanding in video: a combined method, in: *Internatinal Conference on Computer Vision*, Beijing, China, Oct 15–21, 2005.

[6] A. Bobick, J. Davis, The recognition of human movement using temporal templates, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (3) (2001) 257–267.

[7] M. Hu, Visual Pattern Recognition by Moment Invariants, *IRE Trans. Information Theory*, vol. 8, no. 2, pp. 179-187, 1962.

[8] A. Yilmaz, M. Shah, Actions sketch: a novel action representation, in: *Computer Vision and Pattern Recognition*, San Diego, California, USA, June 20–25, 2005.

[9] V. Parameswaran, R. Chellappa, View invariance for human action recognition, *International Journal of Computer Vision* 66 (1) (2006) 83–101.

[10] I. Weiss and M. Ray. Model-based recognition of 3d objects from single images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23, February 2001.

[11] D.D. Vecchio, R.M. Murray, P. Perona, Decomposition of human motion into dynamics-based primitives with application to drawing tasks, *Automatica 39* (12) (2003) 2085–2098.

[12] C. Rao, A. Yilmaz, M. Shah, View-invariant representation and recognition of actions, *Journal of Computer Vision* 50 (2) (2002) 203–226.

[13] S. Park, J.K. Aggarwal, Semantic-level understanding of human actions and interactions using event hierarchy, in: *CVPR Workshop on Articulated and Non-Rigid Motion*, Washington DC, USA, June 2004.

[14] Y. Ivanov, A. Bobick, Recognition of visual activities and interactions by stochastic parsing, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (8) (2000) 852–872.

[15] Ji, X. and Liu, H. 2010. Advances in view-invariant human motion analysis: a review. *Trans. Sys. Man Cyber Part C* 40, 1 (Jan. 2010), 13-24. DOI= http://dx.doi.org/10.1109/TSMCC.2009.2027608

[16] Moeslund, T. B., Hilton, A., and Krüger, V. 2006. A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. 104, 2 (Nov. 2006), 90-126. DOI= http://dx.doi.org/10.1016/j.cviu.2006.08.002

[17]http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING/SINGH/tracks.Aug24.zip

[18] M. Ester, H. Kriegel et Al., A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *Proc. Second Int'l Conf. KDD*, pp. 226-231, 1996.

[19] E.J. Pauwels, P. Fiddelaers and L. Van Gool, DOG-Based Unsupervized Clustering for CBIR, *Proc. Second Int'l Conf. Visual Information Systems*, pp. 13-20, Dec. 1997.