Data mining for activity extraction in video data

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

HAL Id: inria-00497576
https://inria.hal.science/inria-00497576

Submitted on 5 Jul 2010

Data mining for activity extraction
in video data

JoseLuis PATINO, Etienne CORVEE
François BREMOND, Monique THONNAT

INRIA, 2004 route des Lucioles, 06902 Sophia Antipolis (FRANCE)
{jlpatino, Etienne.Corvee, Francois.Bremond, Monique.Thonnat}@sophia.inria.fr
http://www-sop.inria.fr/orion/

Summary. The exploration of large video data is a task which is now possible because of the advances made on object detection and tracking. Data mining techniques such as clustering are typically employed. Such techniques have mainly been applied for segmentation/indexation of video but knowledge extraction of the activity contained in the video has been only partially addressed. In this paper we present how video information is processed with the ultimate aim to achieve knowledge discovery of people activity in the video. First, objects of interest are detected in real time. Then, in an off-line process, we aim to perform knowledge discovery at two stages: 1) finding the main trajectory patterns of people in the video. 2) finding patterns of interaction between people and contextual objects in the scene. An agglomerative hierarchical clustering is employed at each stage. We present results obtained on real videos of the Torino metro (Italy).

1 Introduction

Nowadays, more than ever, the technical and scientific progress requires human operators to handle more and more quantities of data. To treat this huge amount of data, most of the work can now be performed in the data-mining field to synthesize, analyze and extract valuable information, which is generally hidden in the raw data. Clustering is one of the most commonly used techniques in data mining to perform knowledge discovery tasks on large amount of data with no prior knowledge of what could be hidden in the data. There exists many clustering techniques in the literature, and the main goal of all these techniques is to obtain a partition of the data by organizing it automatically into separate groups where the objects inside a specific group are more similar to each other (with regards to their extracted and measured attributes, or variables) than to the objects of the other groups. Mining of text documents (Blatak 2005; Lemoine et al., 2005; Xing et Ah-Hwee 2005) and web-related

information (Chia-hui et Kayed 2006; Facca et Lanzi 2005; Mccurley et Tomkins 2004) are two well-known application fields of data mining. Applying data mining techniques in large video data is now possible also because of the advances made on object detection and tracking (Fusier et al., 2007, Vu et al., 2003). Previous research has focused on semantic video classification for indexing and retrieval (Oh et Bandi 2002; Ewerth et al. 2007) or creation of video summary (Benini et al., 2006) but knowledge extraction on the activity contained in the video has been only partially addressed. Recently it has been shown that the analysis of motion from mobile objects detected in the video can give information about the normal and abnormal trajectory (Porikli, 2004; Naftel et Khalid 2006; Anjum et Cavallaro 2007). However little investigation has been done to find the patterns of interaction between the mobile objects detected and the contextual objects in the scene. A first work was presented by (Patino et al. 2007) where clustering techniques were employed to find patterns of trajectories and patterns of activity. In this paper we present a richer set of features and define new distances between symbolic features. We apply the agglomerative hierarchical clustering 1) to find the main trajectory patterns of people in the video. 2) to extract complex relations between people and contextual objects in the scene.

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. Currently it is being tested on large underground video recordings (GTT metro, Torino, Italy and ATAC metro, Roma, Italy). In this work we present results obtained on real videos from three cameras of the Torino metro (Italy). The general structure of our system is presented in section 2. Object and event detection is explained in section 3. Trajectory analysis of mobile objects is explained in section 4 while extraction of meaningful interactions between people and contextual objects in the scene is described in section five. In section six we present the obtained results. Our conclusion is given is section 7.

2 General structure of the proposed approach

There are three main components which define our approach: the data acquisition, the on-line analysis of video streams, and the off-line analysis. An overview of the system is shown in Figure 1. Video streams are directly fed into our on-line analysis system for real time detection of objects and events in the scene. This procedure goes on a frame-by-frame basis and the results are stored into a specific on-line database. At this level, the detected events already contain semantic information describing the interaction between objects and the contextual information of the scene. This is a first layer of semantic information in our system. The analysis of the detected objects and events retrieved from the on-line database will deliver new information difficult to see directly on the video streams. This constitutes a second layer of semantic information. In this layer the trajectories undertaken by the users are characterised. This information is set up in a suitable knowledge representation model from which complex relationships can be discovered between mobile objects and contextual objects in the scene. Statistical measures such as the most frequent paths, the time spent by the users to interact with the contextual objects of the scene can be inferred.
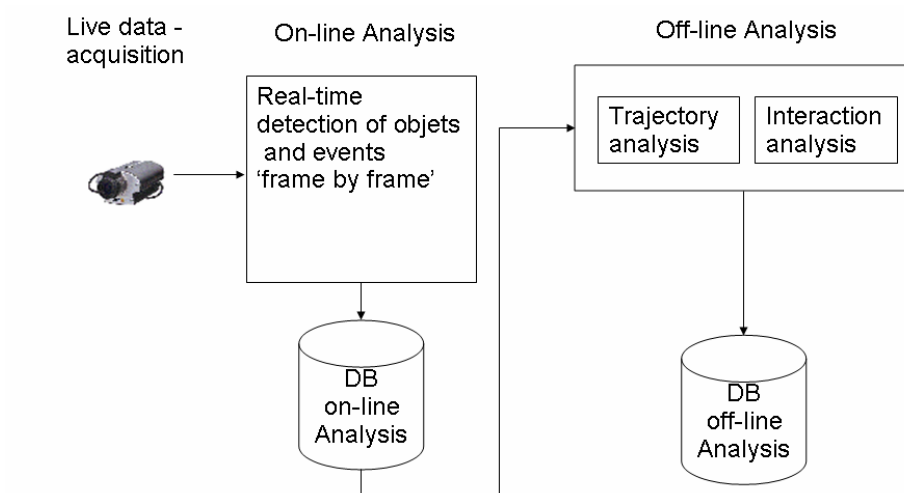
F<small>IG</small>. 1 – *Overview of the proposed approach.*

3 Object and Event detection

The first task of our data-mining approach is to detect in real time objects present in video and events of interest.

3.1 Object tracking

Tracking several mobile objects evolving in a scene is a difficult task to perform. Motion detectors often fails in detecting accurately moving objects referred to as 'mobiles' which induces mistracks of the mobiles. Such errors can be caused by shadows or more importantly by static (when a mobile object is hidden by a background object) or by dynamic (when several mobiles projections onto the image plane overlap) occlusion (Georis et al., 2003).

Briefly speaking a motion detector algorithm allows the detection of objects before being classified and tracked throughout time. The motion detector segments, from the background reference image, the foreground pixels which belong to moving objects by a simple thresholding operation. The foreground pixels are then spatially grouped into moving regions represented by the bounding boxes. These regions are then classified into semantic object classes according to their 3D sizes. The tracking algorithm 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 (Avanzi et al., 2005). The graph of linked objects is analysed by the tracking algorithm, also referred to as the Long Term Tracker, which builds potential paths for each mobile according to the links established by the F2F tracker. The best path is then recognised and the detected objects linked by this path are labelled with the same identifier.

3.2 Event detection

Events of interest are defined according to the semantic language introduced by Vu et al., (2003). This presupposes an ontology where objects of interest 'o'; zone of interest 'z' and contextual objects of interest 'eq' (Contextual objects are part of the empty scene model corresponding to the static environment) are defined. Spatio-temporal relations are then built to form the events of interest:

- inside_zone(o, z)': when an object 'o' is in the zone 'z'.
- 'stays_inside_zone(o,z,T1)': when the event 'inside_ zone(o,z)' is being detected suc-cessively 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': when the event 'close(o, eq, Dmax, T2)' is being consecutively detected for at least T2 seconds.
- 'crowding_in_zone': when the event 'stays_inside_zone (crowd, z, T3)' is detected for at least T3 seconds.

In our particular application we have employ the following variables:

- object o={p, g, c, l, t, u} with p=person, g=group, c=crowd l=luggage, t=train, and u=unknown.
- zone z={platform, validating_zone, vending_zone}
- equipment eq={g1, …, g10, vm1, vm2 } where 'gi' is the ith gate and vmi is the ith vending machine.
- T1=60 s, D=1m50, T2=5 s, T3=120 s.

4 Trajectory Analysis

For the trajectory pattern characterisation of the object, we have selected a comprehensive, compact, and flexible representation suitable also for further analysis as opposed to many video systems which 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 m objects, the trajectory for object i in this dataset is defined as the set of points $[x_i(t),y_i(t)]$; 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, $[x_i(1),y_i(1)]$ and $[x_i(end),y_i(end)]$ 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 include the directional information given as $[\cos(\theta), \sin(\theta)]$, where θ is the angle which defines the vector joining $[x_i(1),y_i(1)]$ and $[x_i(end),y_i(end)]$.

We feed the feature vector formed by these six elements to an agglomerative hierarchical clustering algorithm (Kaufman et Rousseeuw, 1990). For a data set made of m objects there are m*(m-1)/2 pairs in the dataset. We employ the Euclidean distance as a measure of similarity to calculate the distance between all trajectory features. Object trajectories with the minimum distance are clustered together. When two or more trajectories are set together their global centroid is taken into account for further clustering. The successive merging of

clusters is listed by the dendrogram. The evaluation of the dendrogram is typically subjective by adjudging which distance threshold appears to create the most natural grouping of the data. For this reason we have created an interface that allows the user to explore the dendrogram. The final number of clusters is set manually and typical values are between 12 to 25 for a data set of 1000 to 2500 mobile objects. As the acquisition performs in a multi-camera environment the clusters obtained can be generalised to different camera views thanks to a 3D calibration matrix applied during the on-line analysis system.

In order to evaluate our trajectory analysis approach, we have defined a Ground-truth data set containing over 300 trajectories. The data set was manually annotated. Semantic attributes such as 'From south Doors to Vending Machines' were registered into the database. There are one hundred of such annotated semantic attributes. In general, each semantic description is associated with a trajectory that best matches that description. Besides, two more trajectories define the confidence limits within which we can still associate that semantic description. Figure 2 shows some examples of the database trajectories and their associated semantic meaning.

We compute two performance measures to validate the quality of the proposed clustering approach, namely, Merge and Recall. The former gives an indication of how many semantic labels of the ground-truth (or classes) are put together in a single cluster resulting from the agglomerative procedure. These ground-truth semantic labels (containing three trajectories per label) are called 'ground-truth associated to the cluster'. Ideally all the ground-truth trajectories associated to the same semantic label should be included in the same cluster. The latter performance measure (Recall) indicates the number of ground-truth trajectories matching a given cluster relative to the number of 'ground-truth associated to the cluster'. Figure 3 depicts the evolution of these two factors depending on the number of clusters which is chosen when running the clustering algorithm.
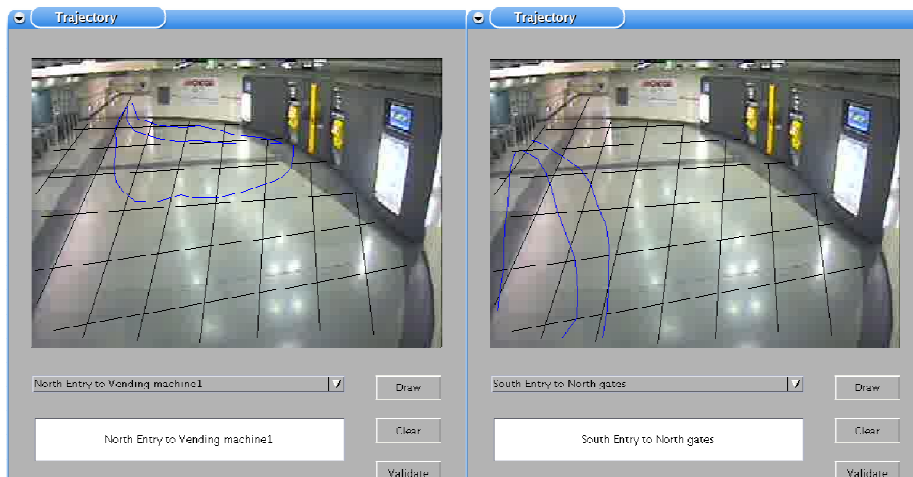

FIG. 2 – *Ground-truth for two different semantic clusters.*

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 merge level and the highest percentage of recall. From Figure 3 it can be observed that a good compromise is achieved for a number of clusters of about 21.
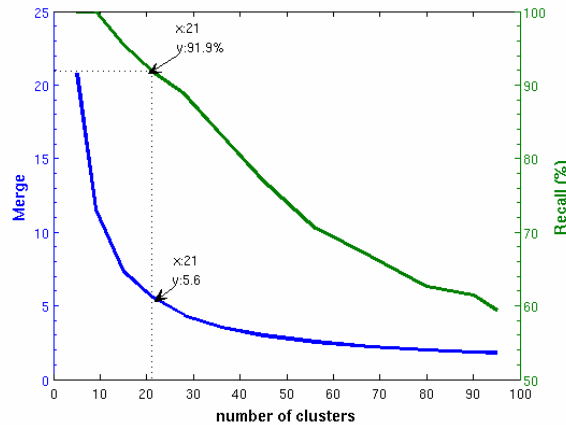

FIG. 3 – *Evolution of the clustering quality measures Merge (fusion of ground-truth semantic labels) and Recall (retrieval of ground-truth trajectories with same semantic label) as a function of the number of clusters.*

5   Interaction Analysis

5.1   Clustering of mobile object table

In a second step of the off-line analysis, we analyse the trajectory of detected mobile objects together with other meaningful features that give information about the interaction between mobiles and contextual elements of the scene. The following mobile object features are employed.

- m_id: the identifier label for the object.
- m_type: the class the object belongs to: Person, Group, Crowd or Luggage.
- m_start: time the object is first seen.
- m_duration: time in which the object is observed.
- m_significant_event: the most significant event among all events. This is calculated as the most frequent event related to the mobile object.
- m_trajectory_type: the trajectory pattern characterising the object.

5.2 Clustering of mobile object table

Once all statistical measures of the activities in the scene have been computed and the corresponding information has been 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 we run a new agglomerative clustering procedure where the data set is the entire mobile object table it-self. Each record of the table is thus defined with five features as the identifier tag is not taken into account for the clustering algorithm. It must be remarked that for this clustering process, the set of features contains numeric (for instance the start time and duration time of an object) and symbolic values (for instance, the object type and the significant event) opposed to the clustering of trajectories where all features are numeric. In order to apply the agglomerative clustering algorithm, we have defined a specific metric for the symbolic values:

For the Object type

$$\left\|o_i - o_j\right\| = \begin{cases} 0 \Leftrightarrow o_i(obj\_type) = o_j(obj\_type) \\ 0.5 \Leftrightarrow o_i(obj\_type) \supset 'PersonGroup', o_j(obj\_type) \supset 'Person' \\ 0.5 \Leftrightarrow o_i(obj\_type) \supset 'PersonGroup', o_j(obj\_type) \supset 'Crowd' \\ 1 \quad otherwise \end{cases}$$

For the Trajectory type

$$\left\|o_i - o_j\right\| = \left\|tc_i - tc_j\right\|$$

Where tc_i, tc_j are the centres or prototypes of the trajectory clusters respectively associated to o_i and o_j and resulting from the last step of trajectory clustering.

For the Significant event we have a logical comparison

$$\left\|o_i - o_j\right\| = \begin{cases} 0 \Leftrightarrow o_i(sig\_event) = o_j(sig\_event) \\ 1 \quad otherwise \end{cases}$$

6 Results

We present the result of our approach on one video sequence of the Torino metro lasting 48 minutes. In total our algorithm detected 2052 mobile objects. Figure 4 shows for instance a tracked person labelled 1 and a tracked crowd of people labelled 534. Figure 5 shows also a tracked person, labelled 58 with two new objects: a group of persons labelled 24 and an

unclassified tracked object labelled 68. Due to the poor contrasted lower part of the group of persons, this group has been segmented into two tracked objects (instead of one) labelled 24 and 68.

FIG. 4 – *A person with label '1' and a crowd with label '534' are tracked. The event 'person' stays at gates has been detected.*

Figure 5 shows the tracked person labelled 1 which remained long enough in front of the validating ticket machines labelled 'Gates' so that the event 'stays_at' is detected. The tracked group of persons labelled 24 in Figure 3 is interacting with the vending machine number 2 long enough for the event 'stays_at' to be detected. In both these figures, the primitive event 'inside_zone' was not shown but was also detected for the remaining objects present in the hall.

FIG. 5 – *A person with label '1' and a crowd with label '534' are tracked. The event 'group stays at vending machine' has been detected.*

We then clustered the trajectories of the 2052 detected mobile objects into 21 clusters employing the hierarchical clustering algorithm. Each cluster represents then a trajectory type. The characterisation of trajectories gives important information on behaviour and flows of people. For instance, trajectory cluster 21 shows people that have just used the Vending machines and move then directly to the gates. Trajectory cluster 13 shows people coming from north doors/gates and exiting through south doors (see Figure 6).
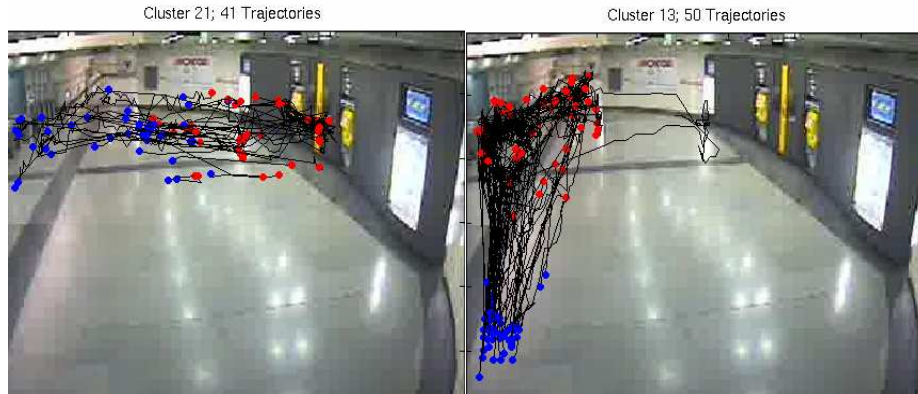
FIG. 6 – *Trajectory cluster 21. People move from the vending machines to the gates.(left panel); Trajectory cluster 13. People coming from the gates going to south doors (right panel)*

Some knowledge that can be inferred from the clustering of trajectories is the following:
- 64% of people are going directly to the gates without stopping at the ticket machine
- 70% of people are coming from north entrance
- At rush hours people are 40% quicker to buy a ticket
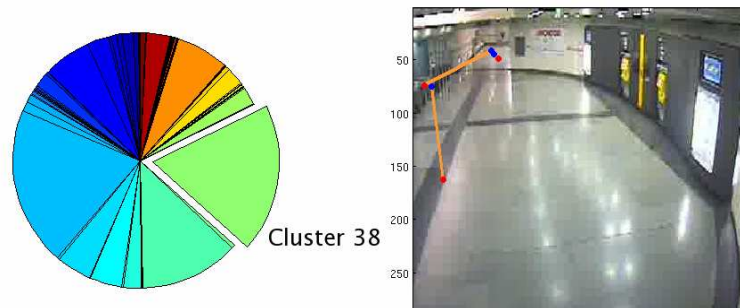- Most people spend 10 sec in the hall ;

FIG. 7 – *Cluster 38 resulting from the clustering of the mobile object table. The right panel indicates the prototype trajectories involved in that cluster.*

Once the trajectories of mobile objects has been characterised, all information is formatted according to the semantic table given in section 4.2. It is then possible to run again the agglomerative clustering algorithm this time on the mobile object table. Some of the clusters found are now detailed. Figure 7 (Cluster 38) represents the biggest cluster found; its detailed description is given in table 1. Cluster 38 is made of 'Unknown objects' for which also no event could have been detected during the tracking phase (section 2). These objects are mainly associated with trajectories of type 4 'exiting gates – going to the north doors (thickest trajectory line). Indeed it is at the gates that object recognition is the hardest to achieve as most activities takes place there. Actually the following two biggest clusters (not shown)

have similar description but involve respectively 'Person' and 'Person Group'. We have thus the knowledge that north doors are the most employed by users.

	Cluster 38	Cluster 6
Number of objects	385	15
Object types	types: {'Unknown'}	types: {'Person'}
	freq: 385	freq: 15
Start time (min)	[0.1533, 48.4633]	[28.09, 46.79]
Duration (sec)	[0.04, 128.24]	[2.04, 75.24]
Trajectory types	types: {'4' '3' '7'}	types: {'13' '12' '19'}
	freq: [381 1 3]	freq: [13 1 1]
Significant event	types: {'void '}	types: {'inside_zone_Platform '}
	freq: 385	freq: 15

TAB. 1 – *Properties for cluster 38 and cluster 6 after clusterisation of the mobile object table.*

Figure 8 presents another cluster (6) including only 15 mobiles but they all have in common the type of object being person and all were detected as being inside the platform. Interestingly this cluster shows us that the trajectories of type '12' and '19' can be related to trajectory '13' (shown in Figure 6). In all three prototype trajectories the exit point are south doors.
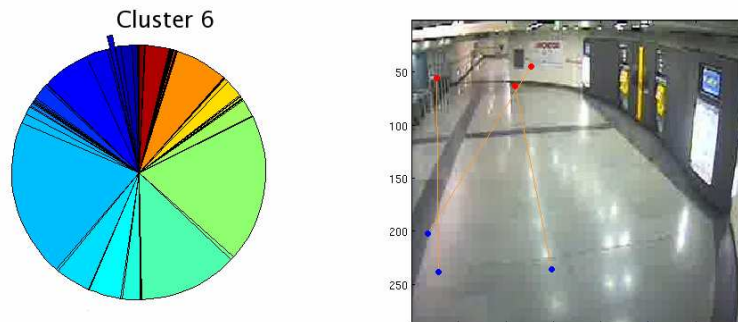

FIG. 8 – *Cluster 6 resulting from the clustering of the mobile object table. The right panel indicates the prototype trajectories involved in that cluster.*

7 Conclusion

In this paper it has been shown how clustering techniques can be applied on video data for the extraction of meaningful information. First hierarchical clustering was applied in order to obtain the prototype trajectories that characterise flows of people in the underground. Then, we apply in a second step again the hierarchical clustering with the aim to achieve knowledge discovery taking into account other meaningful information besides motion such as the type of the detected object and its significant event. For this purpose we

created a specific knowledge modelling format that gathers all information from tracked objects of interest in the scene. This kind of representation allows the end-user to explore the interactions between people and contextual objects of the scene. In this way it is actually possible to obtain statistics on the underground activity and thus optimise available resources. We have defined some semantic distances that allow us to relate different kinds of objects and different types of trajectories. By doing so we directly work on all features characterising mobile objects and analyse heterogeneous variables. This let us find relationships between people, their trajectories and their occurrences. In our future work we will include a learning stage to better define the object models of the scene and diminish the number of 'unknown' detected objects. We will look to add more meaningful features that may characterise a trajectory. We will also work to improve the semantic distances we have implemented such that better relations can be extracted.

References

Anjum N., Cavallaro A. (2007). Single camera calibration for trajectory-based behavior analysis, *Proceedings of IEEE conference on advanced video and signal based surveillance, AVSS'07*, pp 6.

Avanzi A., Bremond F., Tornieri C., Thonnat M. (2005). Design and assessment of an intelligent activity monitoring platform, *EURASIP*, 2359-2374.

Benini, S.; Bianchetti, A.; Leonardi, R.; Migliorati, P. (2006). Extraction of Significant Video Summaries by Dendrogram Analysis, *IEEE International Conference on Image Processing*, 133–136.

Blatak J. (2005). First-order Frequent Patterns in Text Mining, *Proceedings of EPIA*, 344 – 350.

Chia-Hui CH., Kayed, M., Giris, M.R., Shaalan, K.F. (2006). A Survey of Web Information Extraction Systems, *IEEE Transactions on Knowledge and Data Engineering*, 18:1411-1428.

Ewerth R., Freisleben B. (2007). Semi-supervised learning for semantic video retrieval, *Proceedings of the 6th ACM international conference on Image and video retrieval CIVR '07*.

Facca FM., Lanzi PL. (2005). Mining interesting knowledge from weblogs : a survey, DataKnowledge Engineering, 53:225-241.

Fusier F., Valentin V., Brémond F., Thonnat M., Borg D., Thirde M., Ferryman J. (2007) Video Understanding for Complex Activity Recognition. *Machine Vision and Applications Journal*, 18:167-188.

Georis B., Bremond F., Thonnat M., Macq B. (2003). Use of an Evaluation and Diagnosis Method to Improve Tracking Performances, *Proceedings of the 3rd IASTED International Conference on Visualization, Imaging and Image Processing (VIIP)*, 2.

Kaufman L. et Rousseeuw P. J. (1990). *Finding Groups in Data. An Introduction to Cluster Analysis*. New York: Wiley.

Data mining for activity extraction in video data

Lemoine J., Benhadda H., Ah-Pine J. (2006). Classification non supervisee de documents heterogenes : Application au corpus '20 Newsgroups', *The 11th IPMU International Conference*, 3:2538-2544.

Marcotorchino F., Michaud P. (1981). Agrégation des similarités en classification automatique, *Revue de Statistique Appliquée*, 30.

Mccurley K.S., Tomkins, A. (2004). Mining and knowledge discovery from the Web, *Proceedings of the 7th International Symposium on Parallel Architectures, Algorithms and Networks*, 4-9.

Naftel A., Khalid S. (2006). Classifying spatiotemporal object trajectories using unsupervised learning in the coefficient feature space, *Tranactions. of Multimedia Systems*, 12:45–52.

Oh J., Bandi B. (2002). Multimedia data mining framework for raw video sequences, *Multimedia data mining MDM/KDD*, 1-10.

Patino J.L, Benhadda H., Corvee E., Bremond F., Thonnat M. (2007) Video-Data modelling and Discovery, *International Conference on Visual Information Engineering VIE 2007*, pp 9.

Porikli, F. (2004). Learning object trajectory patterns by spectral clustering, *IEEE International Conference on Multimedia and Expo, ICME '04*, 2:1171-1174.

Vu VT., Bremond F., Thonnat M. (2003). Automatic video interpretation: A novel algorithm for temporal scenario recognition, *Proceedings of the IJCAI'03*, 1295-1302.

Xing J., Ah-Hwee T. (2005). Mining ontological knowledge from domain-specific text documents, *Fifth IEEE International Conference on Data Mining*, 4 pp.

Résumé

L'exploration de larges bases de données vidéo est une tâche qui devient possible grâce aux avancées techniques dans la détection et le suivi d'objets. Les méthodes de fouille d'information comme le clustering sont typiquement employées. Celles-ci ont été principalement appliquées pour la segmentation/indexation vidéo mais l'extraction de connaissances sur l'activité présente dans la vidéo a été seulement partiellement adressée. Dans cet article nous présentons comment ces techniques peuvent être utilisées pour traiter de l'information vidéo pour l'extraction de connaissances. Tout d'abord, les objets d'intérêt sont détectés en temps réel. Ensuite, dans un traitement supplémentaire, nous recherchons à extraire des nouvelles connaissances en deux etapes : 1) extraction des motifs caractéristiques des trajectoires des personnes dans la vidéo. 2) extraction des motifs d'interaction entre les personnes et les objets contextuels dans la scène. Dans les deux cas, nous appliquons un clustering hiérarchique agglomératif. Nous présentons des résultats obtenus sur des vidéos du metro de Turin (Italie)