Surveillance video retrieval: what we have already done?

Thi-Lan Le*, Alain Boucher†‡, Monique Thonnat§, and Francois Bremond§
* MICA, Hanoi university of Technology, Hanoi, VietNam
Email: Thi-Lan.Le@mica.edu.vn
† IRD, UMI 209, UMMISCO, IRD France Nord, Bondy, F-93143, France
‡ Institut de la Francophonie pour l'Informatique, MSI, UMI 209, Hanoi, Vietnam
Email: Alain.Boucher@auf.org
§PULSAR team, INRIA Sophia Antipolis, France
Email: {Monique.Thonnat, Francois.Bremond}@sophia.inria.fr

***Abstract*—While many overview papers have been published for information retrieval in general and image retrieval in particular, there is a lack of paper in the literature focusing on retrieval for surveillance video. The aim of this paper is to provide an analysis on what we have ready done for surveillance video retrieval and therefore to point out what are still challenges in this domain. By supposing that there are two main types of information in surveillance video named object and event, we divide the existing approaches in the literature into two sub categories: approaches at object level and approaches at both object and event levels. A quantitative comparison of three approaches of the former category in the same dataset is also given.**

I. Introduction

While many oveview papers have been published for information retrieval in general and image retrieval in particular [1], [2], [3], there is a lack of paper in the litterature focusing on retrieval for surveillance video. The date of birth for surveillance video retrieval is still unknown. To the best of our knowledge, this domain was created in 2000 with the publication of Stringa et al. [4]. Since then, few papers have been published on this domain in comparison with other domains. The aim of this paper is to provide an analysis on what we have ready done for surveillance video retrieval and therefore to point out what are still challenges in this domain.

Surveillance video retrieval is different from that of news videos and movies. In [1], the authors have introduced two new terms: scripted and unscripted content. Scripted content video is a video that is ”carefully produced according to a script or plan that is later edited, compiled and distributed for consumption” and video content that is not scripted is then referred to as unscripted. Based on these definitions, news video is scripted content while surveilance video is unscripted content. The authors have also shown that the representation for unscripted video content is bottom-up: from play and break detection to audiovisual marker detection and highlight identification. Readers are suggested to read the paper [1] to have more information. Object and event are two main important markers for surveillance video content. In general, users want to retrieve objects with some characteristics (e.g. a person wearing a red jacket) and particular events (e.g. abandonned luggage).

Figure 1 shows the surveillance video indexing and retrieval architecture. Videos coming from camera will be interpreted by the video analysis module. There are two modes for using analyzed results. In the first mode, the corresponding alarms will be sent to security staffs to inform them about the situation. In the second mode, analyzed results are stored in order to be used in the future. The surveillance video indexing and retrieval is based more or less on the video analysis module. In general, video analysis module contains object detection, object tracking, object classification and event recognition. Figure 2 presents an example of video analysis and interpretation process [5]. It is worth noting that most of the works in the litterature do not make a clear separtion between video analysis and video indexing and retrieval. In order to analyze the effect of video analysis on indexing and retrieval techniques for surveillance video, in this paper we seperate clearly video analysis from video indexing and retrieval.

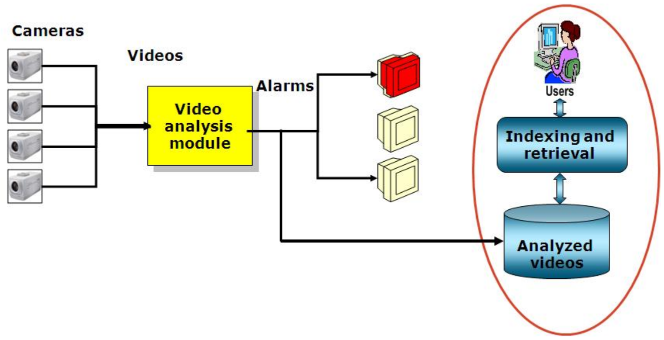


Fig. 1. Surveillance video indexing and retrieval architecture. Videos coming from cameras will be interpreted by the video analysis module. There are two modes for using analyzed results: (1) the corresponding alarms will be sent to security staffs to inform them about the situation; (2) the analyzed results are stored in order to be used in the future.

The remaining paper is organized as follows. In section 2, we analyze existing approaches for surveillance video indexing and retrieval at the object level. In this section, a quantitative

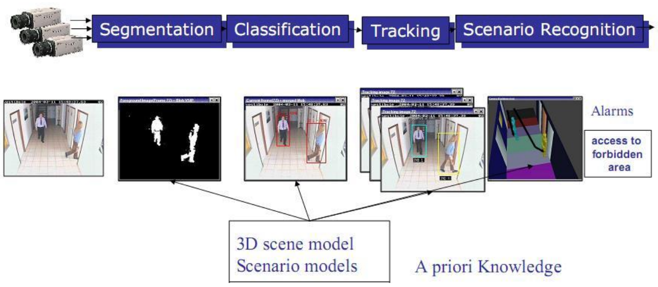


Fig. 2. An example of video analysis and interpretation process [5]. The mobile region detection (object detection), mobile region tracking (object tracking) and scenario recognition and interpretation (event recognition) are performed automatically with help of the a priori knowledge.

evaluation of three state of the art approaches on the same dataset is carried out. Section 3 aims at discussing about surveillance video indexing and retrieval approaches at both object and event level. Based on the analysis, we attempt to point out several challenges of this domain. It is worth noting that in this paper, we focus on semantic information extraction, object signature representation and matching rather than indexing techniques such as Kd-tree.

II. Surveillance video retrieval at the object level

Approaches for surveillance video retrieval at the object level aim at retrieving/searching a particular object or a class of objects with specific characteristics.

A. Analysis

Objects in video surveillance are physical objects (e.g. people, vehicles) that are present in the scene at a certain time. In general, they are detected and tracked in a large number of frames. Consequently, an object is represented by a set of blobs. ”An object blob is a region determined by a minimal bounding box in a frame where object is detected” [6]. The minimal bounding box is determined by object detection algorithms.

Before analyzing the state of the art approaches, we mention three problems for surveillance video retrieval at the object level. As object retrieval utilizes output of the video analysis, the first problem is the quality of object retrieval which depends on the quality of video analysis that is, however, not always perfect. We have analyzed three metrics presented by Nghiem et al. [7] for evaluating object detection and object tracking algorithms: object area, object ID persistence and object ID confusion. For the object detection, the object area metric evaluates the number of pixels in ground-truth object that have been detected. This metric can be expressed in percentage. A good object detection algorithm has a high value for this metric. Figure 3 illustrates three cases: (a) object is not present in the blob (object area = 0%); (b) object is partially present in the blob (object area ¡ 50%), (c) and (d) object is totally present in the blob (object area = 100%) For the object tracking, the object ID persistence metric helps to evaluate the ID persistence. It computes over the time how many tracked objects are associated to one ground-truth object (ID persistence). On the contrary, the object ID confusion metric computes the number of objects per detected object (having the same ID). Figure 4 shows two examples in which the object ID confusion is 3 (Fig. 4.a) and the object ID persistence is 2 (Fig. 4.b). A good object tracking algorithm obtains a small value for these two metrics (minimum is 1). Using directly results of object tracking algorithms with poor performance (evaluation metrics are greater than 1) can lead to irrelevant results for object retrieval. An effective object indexing and retrieval approach should be able to work with video analysis having different qualities.

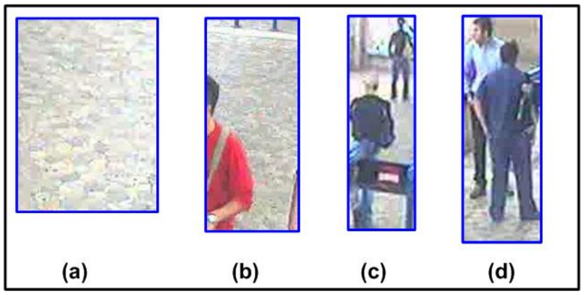


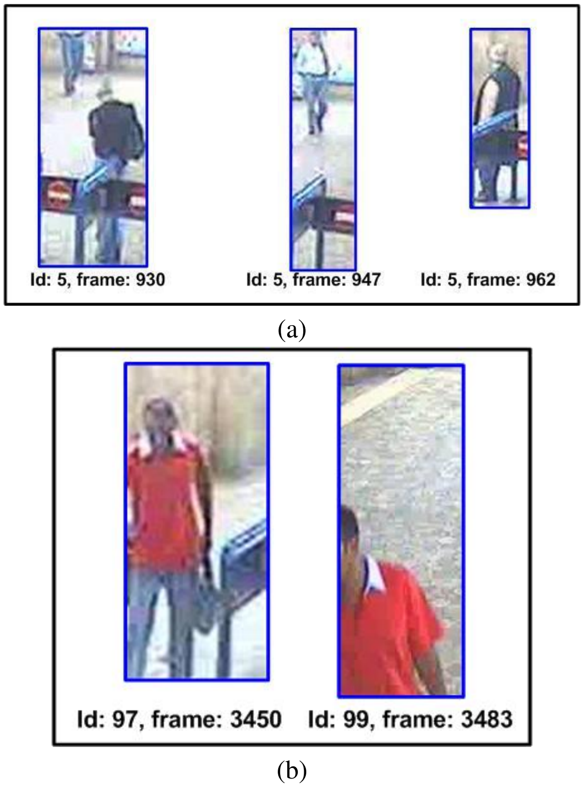
Fig. 3. (a) object is not present in the blob (object area = 0%); (b) object is partially present in the blob (object area ¡ 50%); (c) and (d) object is totally present in the blob (object area = 100%).

The second problem concerns mobile object visual signature. Objects in video surveillance are physical objects (e.g. people, vehicles) that are present in the scene at a certain time. In general, they are detected and tracked in a large number of frames. Consequently, an object is represented by a set of blobs. Due to errors in object detection, using all these blobs for object indexing and retrieval is irrelevant. Moreover, it is redundant because of the similar content between blobs.

The third problem is a large variety of object appearance: just to name a few, people appear in different poses, they are often partially occluded, the lighting is different. The object indexing and retrieval approach has to take into account these variations.

The state of the art surveillance video retrieval approaches at the object level attempt to solve these problems by using different object descriptors, defining efficient object signature and proposing new object matching method. The robust and relevant object descriptor can help to solve the first and the third problems because it allows to match effective object blobs while the way to define object signature attempt to face with the second problem. Because an object is represented by a set of blobs, in order to take into account all object appearance aspects, average descriptor over a set of blobs is computed or certain representative blobs are detected. The relevant object matching method aims at addressing the first problem.

Because the employed object descriptors for surveillance video indexing are similar to descriptors for image retrieval, in the remaining of this section, we analyze two aspects of the existing approaches: object signature and object matching



(a)

(b)

Fig. 4. (a) three ground-truth objects IDs associated to one sole detected object (object ID confusion = 3); (b) two tracked objects created for one sole ground-truth object (object ID persistence = 2).

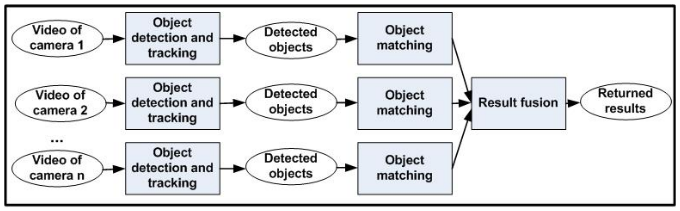


Fig. 5. Late fusion object retrieval approaches: the object detection and tracking is performed on video stream of each camera. Then, the object matching compares query and the detected objects of each camera. The matching result will be fused to form retrieval results.

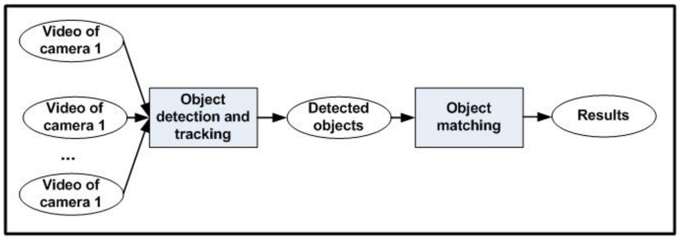


Fig. 6. Early fusion object retrieval approaches.

method.

The existing approaches in this level require that the video analysis module has at least object detection and object tracking modules. In the surveillance application, one object can be seen at the same time by several cameras. Therefore, object detection and matching can be carried out in two modes: late fusion and early fusion. In the late fusion mode (cf. Fig. 5), the object detection and tracking is performed on the video stream of each camera. Then, the object matching compares query and the detected objects of each camera. The matching result will be fused to form retrieval results. In the early fusion mode (cf. Fig. 6), the data fusion is done in object detection and tracking module. We can see that the object retrieval method in this early fusion mode has more opportunities to obtain a good result because if an object is not totally observed by a camera, it may be well captured by other cameras. Most of the state of the art works we analyze below belong to the early fusion mode. However, fusion strategy is not explicitly discussed in these works except the work of Calderara et al. [9].

In [8], objects are firstly detected and tracked by using the Kalman filter. Then, the MPEG-7 descriptors such as dominant colors, edge histograms are computed over the object's life time. This method is not effective because average descriptors cannot characterize reliably the objects when object detection and tracking are not perfect.

The approach presented by Ma et al. [10] consists of two methods: a representative blob detection method and an object matching method. The representative blob detection is based on the agglomerative hierarchical clustering and the covariance matrix extracted on object blobs. After performing the agglomerative clustering on all blobs of an object, clusters containing a small number of elements (outliers) are removed. For the other clusters, one representative blob is defined for each cluster. Therefore, the representative blob detection method dominate errors of the object detection if they occur in a small number of frames. Concerning object matching, the Hausdorff distance is then used to compute the distance between two sets of representative blobs. However, the Hausdorff distance is not relevant when working with object tracking algorithms having a high value of object ID confusion because this distance is extremely sensitive to outliers. If two sets of points A and B are similar, all the points are perfectly superimposed except only one single point in A which is far from any point in B, then the Hausdorff distance is determined by this point.

The work of Le et al. presented in [6] consists of a representative blob detection method and an object matching method. The representative blob detection method outperfom the method introduced by Ma et al. [10] because it removes the blobs without computed objects before performing agglomerative hierarchical clustering. For the object matching, the authors proposed to use EMD (Earth Movers Distance) [11] to compute object distance based on distance of each pair of blobs. The experimental results show that the method proposed by Le et al. is more robust than that of Ma et al. [10] while working with imprecise object detection and tracking module.

The work of Calderara et al. [9] focuses on searching blobs of an object over a network of overlapping cameras. A mixture of Gaussians is used to summarize the appearance of the object observed by a set of cameras. At each instant, the mixture of Gaussians is updated by dominant colors computed over object blob. Query is also represented by a mixture of Gaussians. Objects are retrieved based on distance between mixtures of Gaussians. In this work, objects are successfully retrieved if the object detection and tracking are reliable.

B. *Experimental analysis*

In order to compare the three existing approaches, surveillance video sequences coming from the CARETAKER (Content Analysis and REtrieval Technologies to Apply Extraction to massive Recording) project [1] and CAVIAR project [2] are employed. Videos coming from the CARETAKER and the CAVIAR projects depict human activity in a metro station and a hallway in a shopping centre in Lisbon respectively. These videos of CARETAKER project were analyzed by the VSIP platform of the PULSAR team at INRIA [12]. Video information and analyzed results are presented in Tab. I. In this section, we present two evaluations: visual descriptor and object matching method evaluation.

TABLE I
VIDEOS COMING FROM THE CARETAKER PROJECT AND RESULTS OF VIDEO ANALYSIS (OBJECT DETECTION AND TRACKING)

Name	Duration (min)	Frames	Detected objectss
Video1	$\simeq$20	51450	810
Video2	$\simeq$20	51580	777

1) Visual descriptor evaluation: In this evaluation, we want to answer the question "Which descriptor is the most relevant descriptor for object matching in surveillance videos?". Among an important number of proposed descriptors, we use the dominant color (DC) [13], [14], the edge histogram (EH) [15], the covariance matrix (CM) [10] and the SIFT descriptor [16] because these descriptors are widely used for image retrieval. In this experiment, in order to avoid the effect of object signature and object matching methods with retrieval result, we take only one blob per object. The obtained results are shown in Fig. 7. The results show that if the objects are detected while the background and context objects are not present in the blob, the used descriptors allow to retrieve objects with relatively good results. For other cases, the covariance matrix is more efficient than the other descriptors. Table II shows an example of retrieval result, where the query blob is shown in the left. For this query blob, there are 5 blobs of the same person that are detected at different times or captured by different cameras. We can see that with the covariance matrix, the rank of the relevant blobs is small (the smaller the rank the best the result is). It is interesting to see that when the covariance matrix represents information of all pixels in a blob, the points of interest uses only few pixels. The dominant color and edge histogram use the approximate information of pixel color and edge. A pair of descriptors (covariance matrix and dominant color) or (covariance matrix and edge histogram) or (covariance matrix and SIFT descriptors) may be chosen as descriptors used by default.

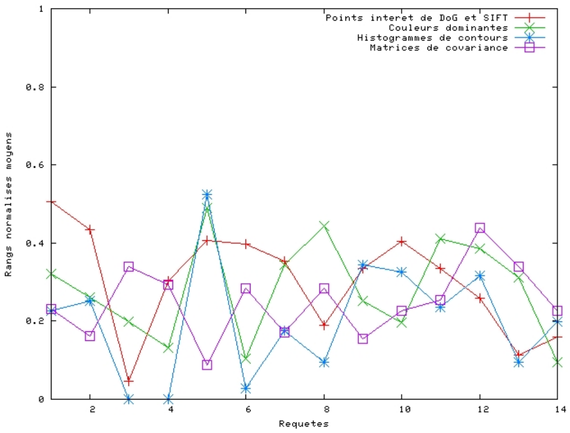
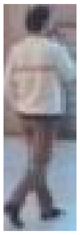
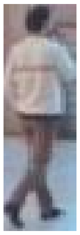
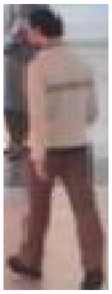
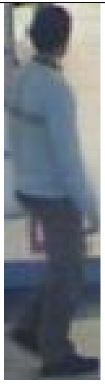
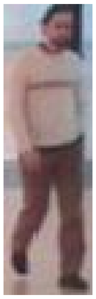


Fig. 7. Retrieval results with 15 blob queries for 53 objects of the CAVIAR project by using SIFT descriptor, edge histogram, dominant color and covariance matrix.

TABLE II
RETRIEVAL RESULT CORRESPONDING TO DOMINANT COLOR (DC), EDGE HISTOGRAM (EH), COVARIANCE MATRIX (CM) AND SIFT DESCRIPTORS. THE QUERY BLOB IS IN THE LEFT WHILE THE FIVE RELEVANT BLOBS AND THEIR RANKS ARE IN THE RIGHT.

Query	Retrieved objects				
					
DC	1	36	31	25	49
EH	1	21	45	35	25
CM	1	7	24	2	10
SIFT	1	39	31	28	29

2) Object matching evaluation: For object matching evaluation, we conduct two experiments to compare three object matching methods in the state of the art: the method of Calderara et al. [9], the method based on Hausdoff distance of Ma et al. [10] and the method based on EMD of Le et al. [6].

The first experiment corresponds to the retrieval scenario: "The security staff want to know whether a person similar to a query image appears in the scene at any other time". We have chosen 247 indexed persons as query. The query person is compared with the 810 indexed persons in Video1. The retrieval scenario in the second experiment is: "The security staff want to know whether a person observed by a camera is observed by another camera". The 54 indexed persons

[1] http://www.ist-caretaker.org/

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

of Video2 become query persons. These query persons are compared with the 810 indexed persons of Video1. In order to validate the retrieval results, we adopt the evaluation measure proposed in [17]: the Average Normalized Rank. It is defined as follows:

$$\widetilde{Rank} = \frac{1}{NN_{rel}}(\sum_{i=1}^{N_{rel}}(R_i) - \frac{N_{rel}(N_{rel}+1)}{2}) \qquad (1)$$

where N_{rel} is the number of relevant results for a particular query, N is the size of the test set, and R_i is the rank of the ith relevant results. $\widetilde{Rank}$ is zero if all N_{rel} are returned first. The $\widetilde{Rank}$ measure is in the range 0 (good retrieval) to 1 (bad retrieval), with 0.5 corresponding to a random retrieval. As we mention above, the method of Le et al. [6] gives better results in most of the cases.

TABLE III
AVERAGE NORMALIZED RANKS OBTAINED BY TWO MATCHING METHODS IN TWO EXPERIMENTS.

exp	Le et al. method [6]	Ma et al. method [10]	Calderara et al. [9]
exp1	0.13	0.15	0.334
exp2	0.32	0.37	0.405

III. SURVEILLANCE VIDEO RETRIEVAL AT THE OBJECT AND EVENT LEVEL

Surveillance video retrieval at the object and event level aims at retrieving objects of a particular event or an event (or a sequence of events with a time constraint) of a particular object. For example, users may want to find persons who abandoned a luggage at the airport (abandoned luggage event) or they want to know whether a person wearing a red jacket enters in a forbidden zone. While the definition of object in surveillance video is relatively stable, the definion of event varies from one research team to another. Therefore, a quantitative evaluation of the approaches belonging to this category is still unrealizable. In this section, we summarize and analyze the theory aspect of the surveillance video retrieval approaches at the object and event level. We divide approaches in the state of the art into two categories. Approaches in the first category support retrieval facility for a fixed and limited event types while that in the second category allows to define new events from the existing ones.

The work presented by Stringa et al. [4] belongs to the first category. In this paper, the authors have proposed a system for retrieving abandoned objects detected in a subway station. Two kinds of retrieval units are supported in this work. The first retrieval unit is the frame where the abandoned object was detected. This frame contains the person who left the abandoned object. The second retrieval unit is the video shot. A video shot is composed of 24 frames, the last frame is the frame where the abandoned object was detected. Similar abandoned objects can be retrieved using descriptors such as position, shape, compactness, etc. Retrieval capacity is limited to abandoned object retrieval. Foresti et al. [18] have tried to expand the work of [4] by adding more types of events.

A surveillance video retrieval system based on object trajectory has been introduced by Hu et al. [19]. Objects in the scene are firstly tracked and then object trajectories are extracted. The spectral algorithm is used to cluster trajectories and to learn object activity models. Several descriptions are added to the activity models such as turn left, low speed. It allows to retrieve the indexed data by keywords, multi-objects and sketch-based trajectories. The object activity model enables to work at both the low level and the semantic level. However, the semantic level is limited to only few activities. A modification has been proposed in [20], instead of using spectral algorithm the authors have applied HSOM (Hierarchical Self-Organizing Map) to learn object activity models. Chen et al. [21] has limited accident event to one sole relation of vehicles' trajectories.

The IBM Smart Video Surveillance system presented in [22] does both video analysis and video retrieval. Video analysis includes different tasks such as object detection, object tracking, object classification, long-term monitoring and movement pattern analysis. Concerning retrieval, users are not able to define new events from recognized ones. This approach does not consider temporal relations of events and objects.

In the second category, the approach of Ghanem et al. [23] presents an extension of Petri net by adding: conditional transitions, hierarchical transitions and tokens with different types of labels. The general idea of this work is to represent a query by a Petri net whose transitions are simple events modeled also by Petri nets. By this way, a complex event can be inferred from recognized simple events. In [24], [25], Le et al. have proposed a SQL like query language for surveillance video retrieval at object and event level. Composed events are defined by combining the simple recognized events and time interval relations. While several approaches have been described for surveillance video retrieval at the object and event level, there are still a lot of works. Formulation a very complex query combining image, object and event information and spatial-temporal relations is not always easy and even unfeasible. Moreover, the matching between this kind of query and database is still an open problem.

Besides the approaches we analyze in sections 2 and 3, several works try to convert surveillance video retrieval to image retrieval. The technique of Meessen et al. [26] has removed the dynamic aspect by extracting keyframes from videos and applying relevance feedback technique on these keyframes.

IV. REMAINING CHALLENGES

Before pointing out the remaining challenges in surveillance video indexing and retrieval, a question may be raised: "Surveillance video retrieval, is it still interesting?". The answer is yes. By discussing with people working on surveillance video analysis community, we can see that there is a real need for developing surveillance video retrieval for both end users (e.g. security staff) and computer vision researchers. Moreover, the object matching method proposed for retrieval can be used for associating objects across non-overlapping cameras.

A. Object visual signature

The first challenge of surveillance video retrieval is the low quality of video and the lack of object information. In surveillance applications, cameras are usually installed far from people. For example, in the airport surveillance application, cameras are sticked on the ceiling of airport hall. With this, the number of pixels for an object (e.g. person, vehicle) is relatively small. In the surveillance application, the detected person is usually considered as a whole blob. There is a huge lack of information concerning person face and posture. Retrieving these persons based on the appearance is therefore difficult. The existing approaches support retrieval facility based on global visual appearance such as retrieving persons wearing a yellow jacket. Moreover, with the imprecise video analysis module, in a lot of cases, the detected blob contains only a small part of the person.

B. Event representation

The second challenge concerns event retrieval. Since we do not have yet a common definition of event for surveillance video, the event representation is still an open problem. At present, events are usually represented as a meta data information for example by event name. Information inference is not possible.

C. Information fusion

The last challenge relates to information fusion. Besides visual information, audio information is available in amount of surveillance applications. However, it is a lack of work dedicated to audio-visual information fusion even in video analysis module. To the best of our knowledge, none work has been done for surveillance video retrieval using both audio and visual information.

ACKNOWLEDGMENT

In this paper, we are using results from the CAVIAR and the CARETAKER EU IST FP6-027231 project. We would like to thank the people involved in both projects for sharing their results.

REFERENCES

[1] Z. Xiong, X. S. Zhou, Q. Tian, R. Rui and T. S. Huang, “Semantic Retrieval of Video - Review of research on video retrieval in meetings, movies and broadcast news, and sports”, *IEEE Processing Magazine* 3(2) (2006) 18–27.

[2] M. S. Lew, N. Sebe and R. Jain, “Content-Based Multimedia Information Retrieval: State of the Art and Challenges”, *ACM Transactions on Multimedia Computing, Communications and Applications* 2(1)(2006) 1–19.

[3] L. A. Rowe and R. Jain, “ACMSIGMMretreat report on future directions in multimedia research”, *ACMTrans. Multimedia Comput. Comm. Appl.* 1(1) (2005) 3–13.

[4] E. Stringa and C. S. Regazzoni, “Real-time video-shot detection for scene surveillance applications”, *IEEE Transactions on Image Processing* 9(1) (2000) 69–79.

[5] Brémond, F., ”Interprétation de scène et video surveillance”, Talk at AViRS 2008 (Analyse Video pour le Renseignement et la Sécurité), Paris, April 2008.

[6] Le, T.L. and Thonnat, M. and Boucher, A. and Bremond F., Appearance based retrieval for tracked objects in surveillance videos, International Conference on Image and Video Retrieval (CIVR 2009), 2009, Santorini, Greece.

[7] Anh-Tuan Nghiem and Francois Bremond and Monique Thonnat and Valery Valentin, ETISEO, performance evaluation for video surveillance systems, Proceedings of International Conference on Advanced Video and Signal Based Surveillance (AVSS 2007), 2007, London, United Kingdom, September.

[8] J. S. C. Yuk, K. Y. K. Wong, R. H. Y. Chung, K. P. Chow, F. Y. L. Chin and K.S.H. Tsang, “Object-Based Surveillance Video Retrieval System with Real-Time Indexing Methodology”, in International Conference on Image Analysis and Recognition (ICIAR’07), 5-7 Sept 2007, pp. 626-637.

[9] S. Calderara, R. Cucchiara and A. Prati, “Multimedia Surveillance: Content-based Retrieval with Multicamera People Tracking”, in ACM International Workshop on Video Surveillance & Sensor Networks (VSSN’06), Santa Barbara, California, USA, 27 October 2006, pp. 95–100.

[10] Y. Ma, B. Miller and I. Cohen, “Video Sequence Querying Using Clustering of Objects’ Appearance Models”, in International Symposium on Visual Computing (ISVC’07), November 26-28, 2007, pp. 328–339.

[11] Y. Rubner, C. Tomasi and L. J. Guibas, “A Metric for Distributions with Applications to Image Databases”, in Proceedings of Int. Conf. on Computer Vision (ICCV’98), 1998, pp. 59–66.

[12] Marcos Zúniga and Franois Bremond and Monique Thonnat, Fast and reliable object classification in video based on a 3D generic model, Proc. Of. 3rd International Conference on Visual Information Engineering (VIE 2006), 2006, Bangalore, India, September 26-28.

[13] Y. Deng and B. S. Manjunath and C. Kenney and M. S. Moore and H. Shin, An efficient color representation for image retrieval, *IEEE Trans. Image Processing*, Vol. 10, 2001, 140-147.

[14] B. S. Manjunath and Jens-Rainer Ohm and Vinod V. Vasudevan and Akio Yamada, Color and texture descriptors, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, number 6, 2001, 703-715.

[15] Dong Kwon Park and Yoon Seok Jeon and Chee Sun Won, Efficient use of local edge histogram descriptor, Proceedings of the 2000 ACM workshops on Multimedia (MULTIMEDIA’00), 2000, 51-54, New York, NY, USA.

[16] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. V. Gool, “A comparison of affine region detectors”, *International Journal Computer Vision* 65(1/2) (2005) 43–72.

[17] H. Müller, S. Marchand-Maillet and T. Pun, “The truth about corel - evaluation in image retrieval”, in In Proc. Of. CIVR, London, July 2002, pp. 28–49.

[18] G. L. Foresti, L. Marcenaro and C. S. Regazzoni, “Automatic detection and indexing of video-event shots for surveillance applications”, *IEEE Transactions on Multimedia* 4(4) (2002) 459–471.

[19] W. Hu, D. Xie, Z. Fu, W. Zeng and S. Maybank, “Semantic-Based Surveillance Video Retrieval”, *IEEE Transactions on Image Processing* 16(4) (2007) 1168–1181.

[20] D. Xie, W. Hu, T. Tan and J. Peng, “Semantic-based traffic video retrieval using activity pattern analysis”, in International Conference on Image Processing, 1, 2004, pp. 693–696.

[21] X. Chen and C. Zhang, “An Interactive Semantic Video Mining and Retrieval Platform–Application in Transportation Surveillance Video for Incident Detection”, in Sixth International Conference on Data Mining, Dec 2006, pp. 129–138.

[22] A. B. Hampapur, L. Feris, R. Senior, A. C. F Shu, Y. Tian, Y. Zhai and L. Max, “Searching surveillance video”, in IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2007), 5-7 Sept 2007, pp. 75–80.

[23] N. Ghanem, D. DeMenthon, D. Doermann and L. Davis, “Representation and Recognition of Events in Surveillance Video Using Petri Nets”, in Conference on Computer Vision and Pattern Recognition Workshop (CVPRW’04), 7, 2004.

[24] Le, T.L. and Thonnat, M. and Boucher, A. and Bremond F., Surveillance Video Indexing and Retrieval Using Object Features and Semantic Events, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 23, No. 7, 1439-1476, 2009.

[25] Le, T.L., Indexation et Recherche de Video pour la Videosurveillance, PhD thesis, Universite de Nice Sophia Antipolis, 2009.

[26] J. Meessen, X. Desurmont, J. F. Delaigle, C. De Vleeschouwer and B. Macq, “Progressive Learning for Interactive Surveillance Scenes Retrieval”, in Workshop on Visual Surveillance (VS07), 2007, pp. 1–8.