Proceedings of the 2010 IEEE
International Conference on Mechatronics and Automation
August 4-7, 2010, Xi'an, China

Metadata Extraction and Organization for Intelligent Video Surveillance System

Han Zhou and Grantham K.H. Pang *Senior Member, IEEE*

Abstract— **The research for metadata extraction originates from the intelligent video surveillance system, which is widely used in outdoor and indoor environment for the aims of traffic monitor, security guard, and intelligent robot. Various features are extracted from the surveillance image sequences such as target detection, target tracking, object's shape and activities. However, the trend of more and more features being used and shared in video surveillance system calls for more attention to bridge the gap between specific analysis algorithms and end-user's expectation. This paper proposes a three-layer object oriented model to extract the surveillance metadata including shape, motion speed, and trajectory of the object emerging in image sequence. Meanwhile, the high-level semantic metadata including entry/exit point, object duration time is organized and stored which are provided for the further end-user queries. The paper also presents the experiment results in different indoor and outdoor surveillance scenarios. At last, a comparative analysis with another traditional method is presented.**

I. Introduction

More and more video surveillance systems have been deployed in indoor scenarios like tower lobby and supermarket, as well as in outdoor scenarios like parking lot and highway. Various objects and features are detected and analyzed from the surveillance videos with a large number of algorithms for object detection and tracking. However, we should pay more attention to end-user's expectation to extract and organize the surveillance metadata from increasing amount of original videos. For instance, a security operator may not only want to be informed in case of important event detection, but also to rapidly analyze the produced metadata to understand when and from which entrance a suspect pedestrian has been identified.

In this situation, the necessity to produce and benefit from the extracted metadata [1][2] in an efficient way becomes more and more significant in the video surveillance research community. Metadata that contains object's moving trajectories, object shape, object behaviors, duration time can be extracted from the raw image sequence.[2] It can then be provided for the further event recognition application, as well as for end-user's understanding on the details of the occurred event and human activity.

Figure 1 shows the use of metadata extracted from input image sequence for user's query or subsequent event recognition. The extraction and organization of metadata becomes an important research issue in recent years. The concept of metadata is used recently for high-level event recognition application. Metadata is a kind of middle-level abstraction extracted from raw image sequence, which normally contains object's position, moving direction, shape, contour, color, and other semantic components. Low-level analytical module analyzes the video streams from a number of video management storages and extracts the metadata for storage in metadata server machine. Ultimately, the metadata is provided to the post-processing modules including high-level event recognition and real-time surveillance.

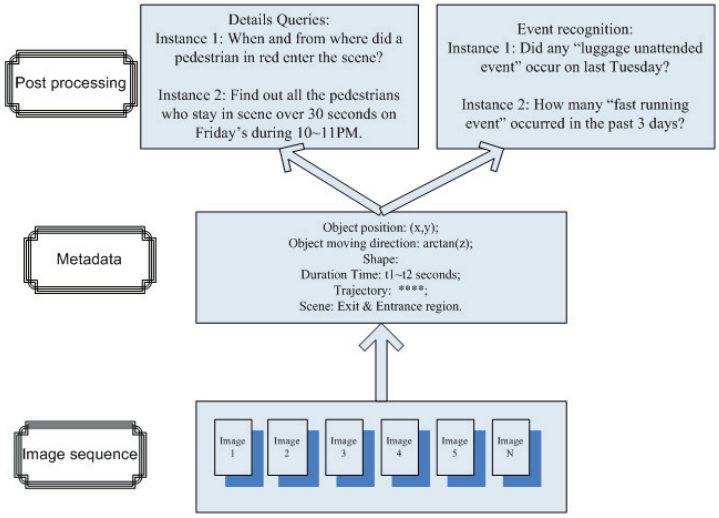

Fig.1: A system figure shows the significance of metadata extraction.

Metadata from raw images plays a significant role in video surveillance due to two advantages. First, the metadata including low-level features of moving target will significantly reduce the post-processing computation amount to massive original surveillance video. The second advantage is that the metadata can facilitate the subsequent post-processing such as event recognition and query. Due to large amount of low-level features are extracted, a well-organized metadata that analyzes and interprets these features is necessary for post-processing like complicated event recognition and details query for end-users.[6] (e.g. object recognition, abandoned luggage detection, pedestrian entry time query…)

The concept of metadata is used for the high-level event recognition application which is an important component of

Han Zhou and Grantham K.H. Pang are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Pakfulam Road, Hong Kong. (e-mail: {zhouhan, gpang}@ eee.hku.hk).

978-1-4244-5141-8/10/$26.00 © 2010 IEEE

the video surveillance system. The IP based surveillance system is consisted of cameras in sub-network, networking hardware and the centralized surveillance center. The system figure is illustrated as below. The metadata plays an important role as the video content indexing. Except for its convenience for post-processing and end-user queries, the using of metadata can also reduce the data transmission amount in the video surveillance network.

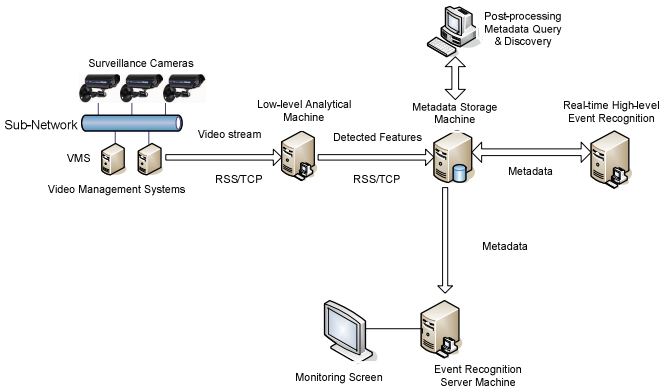

Fig.2: The status of "Metadata extraction" module in the whole event recognition system

This paper proposes an object oriented surveillance metadata organization method. The metadata is extracted in a bottom to up three layers structure and the experiment results are presented at the end of the paper. The organization of this paper is as follows. The section two gives the literature review and research advances in surveillance metadata. The section three presents the three layer surveillance metadata extraction method. The section four elaborates the metadata extraction and semantic information organization from an indoor surveillance example. The section five presents the experiment results and the comparison with another traditional metadata extraction method. The last section gives a conclusion and the future work is discussed.

II. Literature Review

There are several aspects on metadata extraction that have attracted the interests from some researchers. These include: the extraction method from image sequences, the storage organization data structure and the relationship with the high-level event recognition and queries.

C. Carincotte [1] proposed a method that meets the need for a generic, context-independent and adaptive system for storing and managing video analysis results, through the method of creating a metadata warehouse for storing the extracted middle level results from the raw image sequences. The proposed system can analyze the generated metadata over a long period, so as to discover general pattern of events/activities within the video monitored architecture.

Metadata extraction for multiple surveillance cameras was also considered in [3]. A hierarchy based database method has been proposed for the extraction and storage of metadata using a four-layer structure into a surveillance database. This paper has introduced a mechanism to generate video content summaries of object detected in terms of semantic scene models. Some research focuses on the metadata for the complicated event recognition. A survey paper on automatic event understanding [9] considered the whole automatic video event understanding as two main parts, video abstraction and event modeling. Here, video abstraction referred to intermediate metadata that was translated from video sequence inputs [9]. Moreover, the intermediate abstraction metadata provided the basis for the event modeling parts. Target tracking is the basis for event recognition which provides the characteristics of objects for high-level event or action recognition [10].

An ontology based metadata structure was proposed in [4], which gave a set of video event definitions including physical objects of the observed scene and video events occurring. Based on this ontology metadata, a video event description language was created.

R. Nevatia et al. [5,6] proposed a scheme of metadata based video event description language named VERL (Video Event Representation Language) and VEML (Video Event Markup Language) which used for the video browsing and content based video indexing. The details of VERL was given in [6] with the rules of how video index data was organized to represent and annotate the video content.

A more flexible video surveillance metadata organization method was proposed [7], where several query scenarios inspired by real events were tested that can benefit from the extracted metadata. A metadata standard for video surveillance was proposed in [8], specifically, the metadata covered a description of the surveillance system and the activity in the scene for the MPEG-7 data format. In addition to this set, appropriate descriptions for the relation between camera and scene were also considered. To summarize, the research papers [1-8] have proposed various methods and structure of metadata organization, and have used them for the event recognition and also end-user queries.

However, existing metadata producing methods organize the extracted objects according to the temporal sequence. Hence, they are not suitable for end-user's query which often focuses on an object which is in motion, and has certain activities. In addition, in situation where multiple objects have interactions, the extraction of metadata is not obvious and could be challenging. The increasing trend of considering metadata for the complicated automated event recognition can be found in many recent works. Hence, the extraction and organization of metadata need to be further investigated.

III. Proposed Three-Layer Metadata Extraction Method

In this section, we present the proposed metadata extraction method, which extracts the object features from raw image sequence and organizes the metadata under a three-layer bottom to up structure.

A. The Basic Surveillance Metadata

The emerging metadata we extract should be classified into several categories from various internal and external properties which can better fit the further application.

1) The necessary metadata includes an object's location (X-Y coordinate), moving direction, contour, color, motion state (interaction with another object).

Besides, background image, object moving route paths, and object entry and exit region are also extracted as background semantic metadata.

2) The object class of a physical object corresponds to its external nature and usually can be determined by its shape. For example, a person, a table and a car are physical objects.

3) Next, we can develop a concept of activeness, which describes the property of an object's mobility, using the weighed value to evaluate an object is static in the scene or keep moving within a time period.

Note that the concept of activeness is defined based on a time period, which is evaluated at the most recent several images (15~30) to determine whether the object is moving or not at present time point.

B. Using a Three-layer Structure to Extract and Store Object Oriented Metadata

In order to organize the metadata according to the moving objects' performance, we propose an object-oriented three-layer structure with clear functional division for the metadata extraction with multiple objects.

The metadata extraction method is divided into three layers. In the first layer, extracted image layer stores the background modeling of K continuous images (K can be a user-determined parameter in the typical range between 50 and 100). All the moving objects in the images can be retrieved easily.

The second layer is the semantic learning layer, which extracts the moving characteristics of each moving targets, including their moving trajectory, moving direction, objects moving interaction (eg. two objects merge, two objects split).

The third layer, metadata description layer, extracts the metadata of each object, and makes the extracted information well organized and convenient for further post processing and queries. Here, metadata is organized based on the object in every temporal period, (eg. every K consecutive images) and the objects are labeled with its sequential number and the moving information includes position, shape, moving trajectory, duration time, interaction status (eg. an object merges with another object, splits from another object).

As an example, the scene in Figure 4 at railway station is used to elaborate the process of extracting the metadata from the passing pedestrians. Two pedestrians get close to each other for a while, and then separate to have the different route respectively, finally exit from the scene. The metadata extraction is illustrated as the following figure 3, from the original images to the third metadata level.

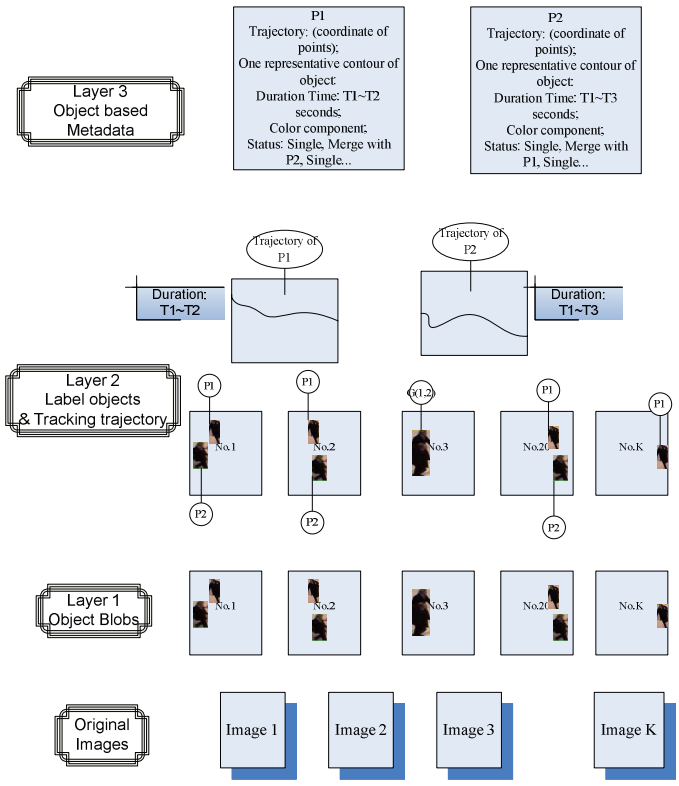

Fig.3: The metadata extraction process for the railway station scene

The metadata is organized on the basis of objects. Take the No.1 Object for example:

-Trajectory (described by consecutive points on trajectory);

-Moving direction (arctan alpha);

-A representative contour of object;

-Duration time t1~t2;

-Color;

-State: (merges with another object as a group, single, or missing);

Scene metadata is like this:

-Scene: moving route (trajectory based probability distribution);

-Entry & Exit Region: ((x1,y1),(x2,y2),(x3,y3),(x4,y4)) (four points described rectangle).

C. Object Blobs and Background Layer

This layer is the lowest layer which is extracted directly from the raw image sequence of the video surveillance system. The camera view is fixed therefore the background abstraction method is used to construct a robust background scene as well as extracting the moving object blobs. We extract the moving object blobs (a number of rectangle regions that include the object pixels) from each image. For each image, a number of rectangle object blob regions are stored in the surveillance database. For an image sequence with K continuous frames, a background scene is also stored. Compared with the storage of original raw image sequences, this strategy can save considerable storage space in the database. Moreover, the original image sequence can be

restored by the object blobs and background in this storing layer. In sum, this is a real-time storage method of each object that is detected in the image sequence.

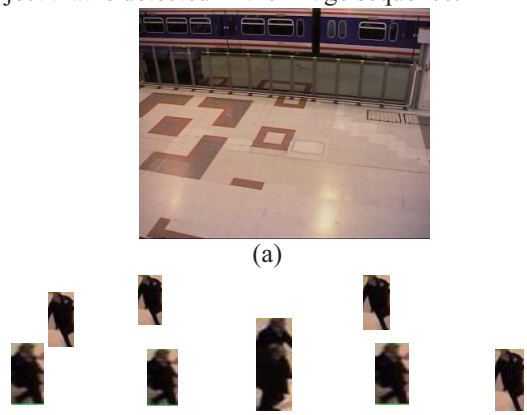

Fig.4 Background scene (a) and the extracted object blobs from five sequential frames (b)~(f)

D. *Object Labeling and Tracking Layer*

The second layer mainly completes the object labeling and tracking tasks, which involve the moving characteristics of each detected object.

First at all, the tracking algorithm is employed on each object in order to obtain the trajectory of objects. We implement the particle filter tracking algorithm on the detected object blobs at the initial frame. Then the trajectories are recorded while the pedestrians are moving. The points coordinate along the trajectory are stored in the second layer.

Second, the object labeling is conducted at this stage. Label the identity of each object refer to keep tracking a particular object target from its entering to exit the filed of view. However, for long image sequence of some surveillance scenarios, object grouping will occur where identity labeling cannot be obtained reliably. In this layer, we match each object identity by using the object trajectory that is obtained by tracking algorithm. The state of motion for each object is also described in terms of group motion, single motion.

E. *Metadata Description Layer*

In this layer, the metadata is extracted and organized based on each object which appears in the field of view. The information in this layer is generated based on the tracking and labeling results in the second layer. The entry and exit region, and routes identified from a single camera view are summarized and stored as the basic metadata of the surveillance scenario.

Based on the trajectory analysis, the metadata from every object that appears in the image sequence can be extracted. The metadata is organized based on different objects. For example, if a pedestrian walks from left to right across the field of view, the point coordinate along the trajectory, entry and exit point coordinate should be recorded. Moreover, the contour points of an object are extracted through edge detection method. The duration time of this object is also stored as the following block, where Ts and Te represent the start and end time respectively.

Except for the above, the color abstraction is summarized according to the color histogram distribution. We divide the color level into 256 different ones, and calculate the number of pixels (inside the object blob region) that belong to different colors. Ten colors are selected to represent the color abstraction, which color contains top ten largest pixel number.

Activeness level ranges from level 1 to level 5 that is calculated based on object's mobility in the recent 15~30 frames. Level one represents the object is static in the recent 15~30 frames. (around one second)

Object ID;
Trajectory points(points coordinate);
Contour (points coordinate);
Duration Time: Ts~Te;
Motion state (Group, single);
Color component (important colors);
Activeness level.

Fig.5 The figure shows the content in metadata layer which represents the characteristics of an object.

IV. Experiment Results and Comparison

A. *Experiment Result*

We present the experiment results on both indoor and outdoor surveillance scenarios. And then the extracted metadata is stored as the structure illustrated in the previous sections. All the stored metadata is based the object, and the characteristics of an object are included as the metadata. The following three scenarios are from indoor railway station and outdoor road and plaza scenarios respectively.

Fig.6 The background scene and (b)~(f) are five object blob regions from railway station image sequence

(a)

(b) (c) (d) (e) (f)

Fig.7 The background from a road scenario image sequence and (b)~(f) are the detected object blob regions in the first minute.

(a)

(b) (c) (d)

Fig.8 The background from a plaza scenario image sequence and (b)~(d) are extracted object regions.

The following one is a table that records the metadata entries which are stored based on the above three scenarios. The third layer indexing metadata takes very small size which provides the abstraction of each object detected in the surveillance scene. The last column gives the size of object blob regions that are stored in the first layer of object blobs.

Scenarios	Duration time(min)	# of object	Metadata size (byte)	Average object region size
Railway station	1.9	41	41*320byte	15.8KB
Road	2.8	35	35*320byte	4.80KB
Plaza	10.5	45	45*320byte	1.80KB

Table 1 The statistics of metadata extracted from the three scenarios

B. Comparative Analysis with the Traditional Method

We would like to compare our method with the metadata extraction method proposed in [3], in order to show the advantages of our method in terms of metadata size and query convenience for end-user.

Under the metadata extraction method of [3], the main idea is to segment the metadata into motion object layer and the movement semantic layer to combine the images for the same scene from different surveillance cameras. Our application focuses on the video stream from a single fixed camera, and subsequent queries concerns the moving object's performance. Under this situation, we consider a different module of metadata, which organize the metadata according to the moving object, not according to the image by image temporal sequence.

The difference between our method and the metadata organization of [3] is that our proposed metadata extraction is moving object based metadata structure rather than temporal based "image by image" structure. The first block diagram shows the traditional temporal sequence organization and the next one shows our proposed object oriented method.

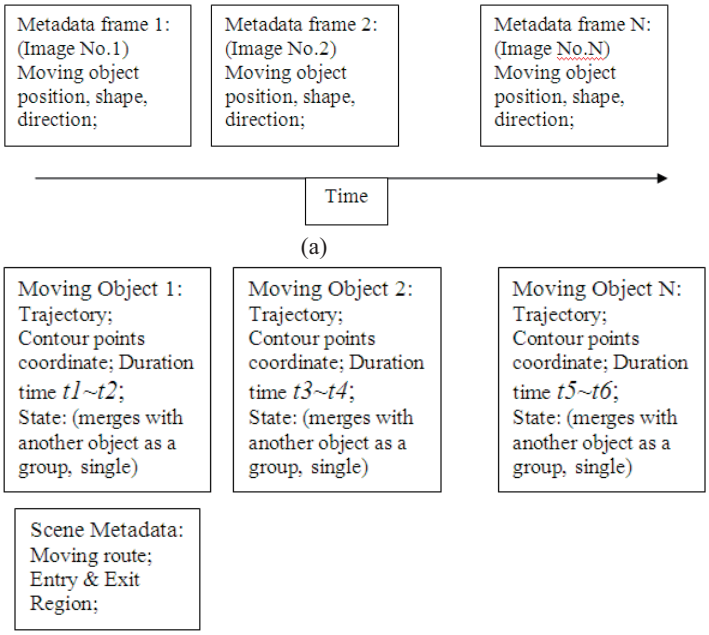

Fig.9: (a) Organization of temporal based metadata from [3] (metadata level). (b) Organization of proposed moving object oriented metadata (metadata level)

The metadata information is tagged with each object detected in the surveillance system. An important application for the extracted metadata is that the object or simple event query can be supported. The metadata scheme provides a better indexing of object motion, where an improved performance for various end-user queries can be achieved. For example, the end-user would like to find a criminal suspect with red clothes who went across the field of view from 1:00~3:00AM on yesterday morning. In order to find the image of the criminal suspect, this security staff may query with the following SQL sentence in the metadata database. The object ID is the primary key of an entry that represents a detected object. After the object ID is obtained, the blob region of the object that is stored in the "object blobs level" can be retrieved according to its entry and exit time and coordinate of location. We implement this query example on both our

proposed object based metadata organization and traditional temporal based metadata organization.

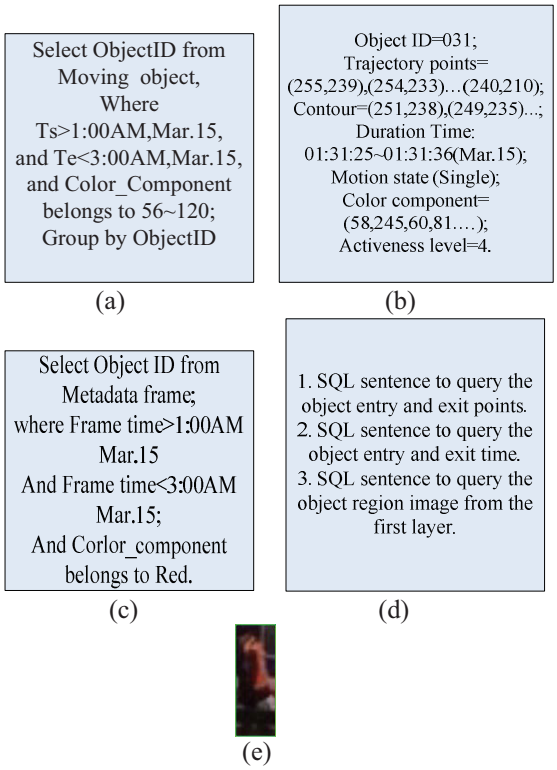

Fig.10 (a)(b) show the SQL query to find the candidate object ID with our proposed object based metadata organization and the output object metadata information. (c) shows the SQL query sentence for traditional temporal based metadata storage method. (d) shows the further query needs three more SQL sentences. (e) shows the object region that with red as a main color component.

We can find the query process for "person in red" from both our proposed metadata organization and traditional organization in [3]. From the above figures, we can see the query for "person in red" through our object based metadata organization can directly get the object information including trajectory, entry and exit points, duration time, motion state, color information and so on. However, for the temporal based metadata organization, the query process can only be achieved indirectly, which involves two steps to find the information about a particular object. First, find the object ID with the red color component. Secondly, retrieve the object information in the temporal based metadata (frame abstract) by several SQL query sentences.

From the example above, the method of metadata organization we propose with an object based metadata can be more suitable for the queries of object's activities in the field of view. The traditional method is easy and straightforward in metadata extraction period because it organizes the metadata on the basis of each continuous frame. However, when the metadata is providing for object queries, our proposed structure performs better due to its flexibility and matching for object queries. It also spends less query response time compared with the traditional metadata organization.

V. Future Work

In the wide application of video surveillance system, semantic metadata can assist the post event recognition and end-user queries. The main benefit of our metadata organization structure is that high-level activities query can be conducted based on the metadata database. Also, the metadata used in surveillance system can reduce the network load and save the transmission network bandwidth resources. For the future work, more complicated event should be queried based on more complex organized metadata. For example, pedestrian loitering, illegal access to some specified region need to be queried by end-user. Except for the above ones, probability method also can be used to describe the relationship of various metadata components.

References

[1] C. Carincotte, X. Desurmont, and A. Bastide, "Adaptive Metadata Management System for Distributed Video Content Analysis," 10th International Conference on Advanced Concepts for Intelligent Vision Systems 2008, Juan-les-Pins, France, LNCS 5259, pp. 334–345, 2008.

[2] P. Chmelar, and J. Zendulka, "Video Surveillance Metadata Management," Proceedings of 18th International Conference on Database and Expert Systems Applications (DEXA 2007), Regensburg, Germany, 2007.

[3] J. Black, T. Ellis, and D. Makris, "A Hierarchical Database for Video Surveillance Applications," IEEE International Conference on Multimedia and Expo (ICME), 2004.

[4] F. Bremond, N. Maillot, M. Thonnat and V. T. Vu, "Ontologies For Video Events," Technical Report of INRIA, Sophia Antipolis. April, 2004.

[5] R. Nevatia, J. Hobbs, and B. Bolles, "An Ontology for Video Event Representation," IEEE Workshop on Event Detection and Recognition, June 2004.

[6] R.J. Alexandre, R. Nevatia, J. Hobbs, and C. Bolles, "VERL: An Ontology Framework for Representing and Annotating Video Events," IEEE Multimedia Magazine, pp. 76-86, October-December 2005.

[7] S. Han, A. Hutter, and W. Stechele, "Toward contextual forensic retrieval for video surveillance: Challenges and an architectural approach," WIAMIS, pp.201-204, 10th Workshop on Image Analysis for Multimedia Interactive Services, 2009.

[8] J. Annesley, A. Colombo, J. Orwell, and S. Velastin, "A profile of MPEG-7 for video surveillance," AVSS, pp.482-487, 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, 2007.

[9] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Volume: 39, Issue: 5, On page(s): 489-504, Sept., 2009.

[10] H. Zhou, M. Taj and A. Cavallaro, "Target detection and tracking with heterogeneous sensors," IEEE Journal of Selected Topics in Signal Processing, Vol. 2, N0. 4, 2008, pp.503-513.