VIGILANT: A Semantic Model for Content and Event Based Indexing and Retrieval of Surveillance Video

Kamel Zerzour, Gaël Frazier and Farhi Marir
School of Computing and Information Systems,
Kingston University
Penrhyn Road,
Kingston Upon Thames KT1 2EE
United Kingdom
Email: {k.zerzour, frazier, f.marir }@kingston.ac.uk

Abstract

This paper presents a semantic video-object model for efficient storage, indexing and content/event-based retrieval of real-time surveillance video without reverting to the constant re-interpretation of source and thus avoiding time-consuming analysis of every video surveillance query. Based on the work on object tracking carried out at the Digital Image Research Centre (DIRC) at Kingston University [Orwell99a, Orwell99b], the model automatically extracts moving objects such as car, identifies their properties and labels events (e.g. a car entering the park) in a car park environment. A Video-Object (VO) is an abstraction of a Video Shot (a sequence of video frames filmed either from a fixed camera position or using a coherent camera motion such as panning, rotation and zooming). A combination of object oriented modelling and description logic knowledge representation techniques will be used in the realisation of VIGILANT video object model. Object-Oriented Database (OODB) will model the static aspects of a video shot using the concepts of classes and relationships. A Description Logics (DLs) knowledge base will be employed to handle the dynamic and incremental content description of videos. This video object model will be used to index extracted features and recognised objects provided by the Image Processing module.

1 Introduction and background Work

Research in visual surveillance has recently been boosted by the availability of more sophisticated methods for tracking simple movements. Objects involved in such movements are usually recognised at video capture time. Yet, efficient interpretation and labelling of events still have not caught up primarily due to inherent temporal properties of videos where an event is associated with an extended temporal context expanding over a sequence of video frames (Video Shot).

Recently, video surveillance industry, particularly in the UK, has experienced growth rates of over 10 percent per year. CCTV camera systems have been installed on company premises, stores of city and around councils. the total annual UK market is believed to be valued at £2billion. The bulk of this expenditure is spent on hardware i.e. cameras, recording equipment and control centres. Currently, the overwhelming bulk of video systems involve multiplexing CCTV channels onto single 24-hour video tapes. The extensive range of commercial and public safety applications include intruder alarms, collation of shopping behaviours, automatic event logging for surveillance and car park administration. Assuming that more sophisticated systems may be marketed, the issue of how to efficiently manage and intelligently search through the gigabytes of video footage generated each day from surveillance and monitoring will depend on the type of information stored and how efficiently it is indexed. Presently a complex query such as (Display all fast moving red cars entering the park between 3pm and 5pm on thursday) could not be easily answered.

The need for speedy retrieval of information from a plethora of video cameras has become a central focus for a multidisciplinary research collaboration involving Database, Computer Vision and Artificial intelligence. Database research concentrated on providing models to mostly handle static aspects of data video (with few or no support for the dynamic aspect), and indices are usually created offline. From the Image Processing point of view, video (and image) features are created automatically during image (frame) capture. These features usually include motion vectors and, in most cases, an object-recognition module is employed to depict object attributes (identity, colour, etc.).

Content Based Image Retrieval (CBIR) is a potential player with an emphasis on how to efficiently manage the knowledge extracted from video processing and how to automatically capture and annotate the video content for efficient retrieval. Although CBIR technology is beginning to move from laboratories to marketplace, it still lacks maturity and is not yet being used on a significant scale. Firstly, this is attributed to the inherent text-based (keyword) mothods of Document Retrieval. Secondly, most current state-of-the-art image indexing and retrieval systems rely heavily on image primitive features (level 1) such as colour, texture or spatial relationships. Research into retrieval by derived attributes (logical feature retrieval: level 2),such as object identities, is only just starting in earnest. Attempts were made into level 3 retrieval (abstract attributes such as happiness) but no significant research has yet been reported, as the issues involved are dauntingly complex [Eakins99, Goble96]. To address the limitations of current CBIR systems and adapt them to build robust Content-Based Video Retrieval (CBVR) systems, the proposed model recognise the need for synergy between databases, image processing communities and other knowledge representation fields such as Natural Language Processing (NLP) and Description Logics (DLs).

The Database indexing realm includes the works carried by [Hjelsvold95] who developed an EER video model that captures video data structure (sequence, scene, and shot) and supports thematic indexing based on manually inserted annotations (Person, Location, Event). They also provided a set of operations to model video data. [Oomoto94] investigated the appropriateness of existing OODB systems for Multimedia data modelling and suggested the need for more modelling tools to handle video data as conventional OODBMS suffer from three main limitations: (1) Video data is raw and its use is independent from why it is created, (2) Video description is incremental and dynamic, and finally The need to deal with data descriptors sharing when modelling overlapping meaningful scenes.[Smoliar94] proposed a content-based video indexing and retrieval architecture in which a DBMS is used coupled with capabilities to parse, index and retrieve (and browse) video clips. Semantic indexing was attained through a frame-based knowledge base (similar to a record in traditional databases) in the shape of a restricted tree form. The database also includes textual descriptions of all frames present in the knowledge base.

Objects may also be classified on their motion by extracting trajectories which are used with other primitives (colour, shape) as keys to index moving objects [Sahouria97]. Description Logics is also used to build Terminology Servers to handle video data content in a bottom up approach [Goble96]. DLs proved beneficial in terms of dynamic, incremental classification, knowledge extensibility and evolution, precise and imprecise querying abilities. Recently [Ivanov99], Standard Stochastic Context-Free Grammar (SCFG) were employed to contextually label events in an outdoor car park monitoring. Objects and events are incrementally detected using a confidence factor with events being detected on the basis of Spatial Locations when objects tend to enter or leave the scene. But no storage issues were addressed. [Romagnino98] modelled object interaction by Behaviour and Situation agents using Bayesian Networks. The system applies textual descriptions for dynamic activities in the 3D world by exploiting two levels of description: Object level where each detected object is assigned a behaviour agent, and Interaction level using situation agents.[Regazzoni98] developed a system to detect abandoned objects in unattended railway stations. The system is able to reveal to the operator the presence of abandoned objects in the waiting room. It also has the capability to index video sequences based on an event detection module which classifies objects into: abandoned, person, structural change (e.g. chair changes position) or lighting effect, with the aid of a Multilevel Perceptron Neural Network.

The remainder of this paper is organised as follows: section 3 presents the proposed model for which an architecture is adapted (and discussed in section 4) to reflect both static and dynamic aspects of surveillance video. The image processing module is introduced in section 5 whereas retrieval of video content and events is discussed in section 6, which is followed by conclusions and future challenges.

2 A Global View of the Proposed System

Similar to [Lahlou95], a Video Shot specification consists of three main parts: a Description to represent content information (such as start/end time), a Composition part to identify the different constituents of a Video Object (these are objects each with its own constituents and states) and finally an Event to capture relationships between a VO's components.

Physically, a Video Shot is made up of number of frames, each of which is composed of a background, and a number of objects. Initially the background is either the empty parking lot or a car park with a number of objects that happen to be static when recording starts. Objects depicted in each frame are assigned two states. Objects with moving status are instances of an object that is still being tracked (an object

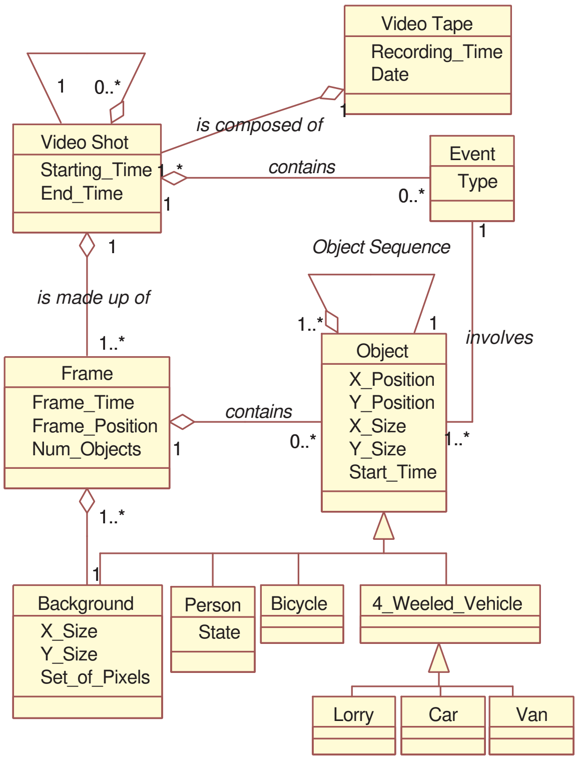

Figure 1: The VIGILANT video object model represented in UML

that either starts moving in the current frame or one whose tracking has started in previous frames). The frame background implemented here is dynamic, so if a moving object stops for a period of time (e.g. 5 seconds) it becomes part of the background.

From the event point of view, the proposed model views video shots at two levels. If no events are detected a video shot is the same as a video tape and is hence composed solely of frames where only background is present. In general a surveillance video tape consists of a number of video shots (in general more than one camera is used and they are allowed to move to have a wider view). The model also allows for the possibility of video shot overlapping and shots containing other shots. Events are triggered, interpreted and tagged dynamically with an aid of an object recognition module which determines object identities. An event can be viewed as a signature (trail) of a particular object spanning over a number of frames, where the same objects tends to change position (and may be appearance depending on location with respect to the camera), speed but not (logical) identities (an object retains its logical identity throughout the lifetime of the event it is involved in). A video object model is presented in figure 1 using Unifed Modelling Language (UML). Various design decision have been made, the discussion of which is left to the conclusion.

3 System architecture

The proposal solution is to develop a dynamic descriptive Video-Object model reflecting the evolution of video content in real time. It uses object-oriented modelling to present the hierarchical structure (static) of the video content enhanced with description logics to model the incremental aspect of that video content (dynamic). Three complemetary types of schemas need to be implemented. Figure 2 illustrate the main components of the system's architecture.

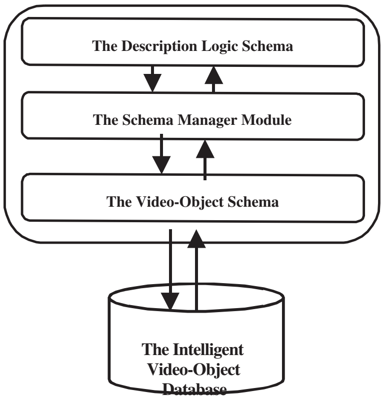

Figure 2: Architecture of the VIGILANT system

3.1 Video-Object Schema

Video content is difficult to model within a structured data model such as the relational or object-oriented mainly due to the dynamic changes of this content. The proposed solution is to develop a dynamic descriptive video object model reflecting the evolution of the content as camera shots is taken. It uses object-oriented (OO) modelling to present the hierarchical structure (static) of the video content enhanced with Description Logics (DLs) to model the incremental aspect of that video content (dynamic).

The video-object model proposed here extends the EMIR system developed by [Lahlou95] and models the video shots in terms of three main parts: The description part which holds attributes such as start and end times of the main video-object. The Composition part is a hierarchical structure which represents all the component objects of the main video-object "car", their sub-components (a car door) and their states (a door is open). The event part describes and labels the significance of the event depicted from a video scene. In this part, the significance of the event is used as a relationship relating two or many objects of the main video-object. For example the event "person gets in car" relates "person" and "car" through the "gets in" relationship.

Existing Object-Oriented modelling tools have been described in [Oomoto94] as to allow representation and management of complex objects, handle object identities, provide encapsulation and inheritance mechanisms. A video shot is a sequence of consecutive frames (images) which appear to have been continuously filmed and exhibit some contextual consistency e.g. "a car entering a car park". Shots are the fundamental unit of manipulation, having their own state and behaviour. The most appropriate tool to model the static properties of a Video Shot is the object-oriented data modelling, thanks to its powerful aggregation and inheritance concepts and its abstraction power to handle complexity. Yet, they are still limited in dealing with video data due to the follwing restrictions:

- Video data is raw and its use is independent from the reason it is created,
- video content description is both changing and incremental,
- the provided inheritance is unable to accurately model metadata sharing for overlapping video shots.

A successful Video Object data model should also be able to interpret and label events depicted in surveillance video.DL knowledge base is proposed to enhance the obtained video-object model.

3.2 The Description Logic Schema

DLs, otherwise known as Terminology Logics, are descendants of the KL-ONE language. They have been studied extensively in the field of Artificial Intelligence, especially in Natural Language Processing (NLP) [Borgida95].

Description Logic (DL) is selected to augment the video-object with further knowledge in order to model the dynamic changes of the video. This is done for several reasons [Goble96]. Firstly, DLs offer a natural approach to flexible extensible schema and indexing for complex data and is equipped with both formal semantics and inference mechanism. Recent applications, which have used DL based knowledge to describe and classify multimedia documents, have been reported to be successful. Secondly, DLs are powerful in the in the sense that they rely on the explosion of automatic subsumtion of compositional descriptions with respect to one another. Thus the classification process is performed dynamically. Thirdly, Modelling using DLs can easily extend and evolve by increasing the set of concepts and / or roles reflecting the nature of databases Fourthly, DLs are very good in dealing with precise and imprecise querying. This comes useful in querying knowledge bases where users are not familiar with the content and structure of data. (eg. Retrieve clips of all red cars entering the park after 12:00 am, and what user is really interested in is clips of all cars entering the parts at that time). This is important when the result is empty in which case one might attempt to generalise the query until something is returned and hence allowing retrieval of objects fitting general description. And finally, the description itself represents precompiled answers to all possible queries. That is indices for events, and hence frame sequences, are provided by the concept definition itsdelf, which form the key to video retrieval.

The DL knowledge base will represent the Video-object model in terms of concepts, roles and individuals. This process is facilitated by the fact that DLs and OO data models share similar modelling concepts. DLs model an application domain in terms of concepts (OO classes), roles (OO relationship) and individuals (OO objects).

The obtained DL video-model will address the limitation of the Video-object in dealing with unknown objects and events by implementing an incremental process of identifying objects (content) or relationship between objects (event). Once the process classifies an object or an event as unknown, it will be given an initial generic description, which is refined and changed as further information become available. This information is used by a Schema Manager Module to determine automatically whether the object or event is already modelled in the Video-Object Model. If the object or event is recognised its position in the Video-model hierarchy is automatically determined and is then classified dynamically. If it remains unidentified, the schema manager will make changes to the video-object schema using the DL inference engine and the schema evolution mechanism to introduce a new object or event either to the DLs Knowledge Base or the video object database. For example suppose the process has to identify "a person is getting in the car" event. At first it could not differentiate it from the event "a person is getting out of the car" until further information such as "the person disappeared" or the "person is leaving the car" is processed.

DLs will be employed to handle limitations of the OO data models and to deal with the dynamic aspects of video objects. However, using solely DLs to model video data has its shortcomings. Firstly, reasoning about classes cannot be performed independently from the class definition due to the way both Terminological Box (T-Box) and Assertional Box (A-Box) work. Secondly, DLs are unable to deal with a large number of concept definitions and instances as it is the case of a video repository (storage in terms of giga / terabytes).

3.3 Schema manager

The Schema Manager Module aims at providing communication channels between the DLs inference engine and OO schemas to reflect the dynamic nature of video content. Another equally important role is dealing with the new developments that could occur to an unknown object. As soon as an unknown object or scene in the DL knowledge base is fully recognised, the Schema Manger transfers the recognised object's instance into the video-object database. However, sometime the unknown object remains unrecognised in which case the schema manager will use the description or structure provided by the DL inference engine to either automatically update the video-object schema and include the new object using the schema evolution process, or leave unrecognised objects residing in the DLs knowledge base.

This challenging task will be achieved through the design of a mapping mechanism between the DLs inference engine and OO Video-Object Schema. Since the interpretation domain for the DL consists of atomic objects, while each instance of an object-oriented model is assigned a structured value, the mapping mechanism needs to implement new concepts and functions to allow translation from one schema to the other. This way, DLs inability to handle large number of concepts and instances will be addressed through: (1) Enhancing the DL inference engine with facilities to access secondary storage (where both the concepts that compose the DL schema and the instances derived from the schema might be stored) and also (2) Improving the performance problems with the reclassification and recognition of instances whenever conceptual schemas or instances are changed.

The design of the mapping mechanism between DLs inference engine and video database will extend the work done by [Marir93] where he extended first-order logic (PROLOG-like) with relational data management facilities to access external storage. This was achieved through the development of an algorithm that transforms the first order logic (Horn Clause) into equivalent relational algebra expressions which could be evaluated during query processing using algebraic operations such as selection, union, Cartesian products and join operations and SQL.

4 Image Processing and Object Recognition

At the heart of the system is Image Processing and Object Recognition module which identifies object identities (car, person,etc.), object primitive properties (colour, texture, etc.) and object behaviour (motion trajectories) [Orwell99a, Orwell99b]. The Agent paradigm is used where cameras are assigned camera-agents which detect and track regions of interest. Such information is then used to update object-agents. By comparing properties, two objects agents can infer that they have the same referent (cameras are observing the same entity) and thus merge identities. Object trajectories are classified with an activity type with reference to ground plane agents.

5 Video Content and Event Indexing and Retrieval

5.1 Real-Time Video Indexing Module

Video Content Indexing can be classified into two categories. Indexing of individual objects depicted in a video scene, which include information such as identity, colour, shape and spatial relationships. Assigning tags (keys) to video shots is referred to as Event-Based Indexing (e.g. a car entering a car park). Before a shot and its objects can be indexed, its boundaries must be determined. Various methods for detecting shot boundaries have been employed, but the most widely used technique is pixel-intensity histogram comparison [Smoliar94] which detects changes from one shot to the next involving sudden changes in screen image content and camera angle. The next step is to select a single representative frame (key frame or r-frame) from each shot and apply the same still image techniques (colour, texture, shape, etc.) including object motion.

Key-frame techniques are not suitable for applications such as Car Park Surveillance as the distortion of image content varies considerably even from one frame to another. In cases as such the problem of key-frame selection becomes a tracking problem where moving objects are analysed in all the frames they appear in, in order to build a coherent content description including motion, colour and shape. The problem of vehicle and human tracking has recently become a hot research topic involving a multidisciplinary collaboration between Database, Image Processing and Artificial Intelligence communities.

The Digital image Research Centre (DIRC) at Kingston University has implemented an event-detection and tracking algorithm that detects and encodes visual events captured by a security camera [Orwell99a, Orwell99b]. The object and the event streams are parsed automatically to recover the visual bounds and pixels representing moving objects (cars, persons, etc.) in the park. The output of this process will be used by the Video-Index Module to either populate the video-object database (if the object or the events are well recognised) or to enrich the DL knowledge base (if the objects or the events are not yet known). As soon as an instance of a video object or event is created and stored in the video-object database or the DL knowledge base, the Video-Index will generate automatically a logical pointer to the digital file containing the video shot (an image or a video clip). To eliminate redundant video indices in the database, the index-generating process will only be triggered when the video-Index Module detects a change of content or events in the car park.

The instances and the indices generated will reflect the three parts of the video-object model: description (such as colour of a car), composition (such as the number of cars in a car park car) and events (such as moving car). Based on these three levels of indices, the user can later query the database and retrieve the pertinent images or video clips.

5.2 content and Event-Based Retrieval

Video retrieval is a topic of increasing importance. While the requirements of video users can vary considerably, video queries are classified into three levels of abstraction (reflecting the classification process of the Video-Index Module): primitive features such as colour or shape, logical features such as the identity of objects shown, and abstract attributes such as the significance of the scenes depicted [Eakins99, Aigrain96]. The description, composition and event parts proposed for the design of the video-object model reflect these levels of abstraction.

Current Content-Based Image Retrieval (CBIR) techniques are used to break up long videos into individual shots, extract still keyframes summarising the content of each shot, and search for video clips containing specified types of movement. The effectiveness of all current CBIR systems is inherently limited by the fact that they can operate only at the primitive feature level. None of them can search effectively for "a car entering a park" though some semantic queries can be handled by specifying them in terms of primitives. While Content-based image retrieval (CBIR) systems currently operate effectively at the lowest of these levels, a variety of domains, including crime prevention, medicine, architecture, fashion and particularly video surveillance demand higher levels of retrieval.

To respond to this urgent need, this work proposes to research and implement queries for the second and third level of abstraction. For this, it is proposed to develop a real-time Video-Object Query Language (Video-QQL) which will uses the semantic content, indexes and annotation generated by the Real-time Index Module. The Real-time Video-OQL module will not be limited to the first level of abstraction i.e. (primitive features) but respond to complex queries such as “display all the video clips showing red cars entering the car park between 11am and 2pm”. This could be expressed in OQL-Video as:

```
SELECT VideoObject = VideoClip
FROM VideoClip.Cars AND VideoClip.CarPark
WHERE VideoClip.Event = ‘‘car entering the car park’’ AND
      VideoClip.Time = IN (11am, 3pm) AND Car.Colour=‘‘Red’’.
```

To achieve such complex queries, the Video-OQL will combine the facilities provided by the object database Object Query Language (OQL) and the description logic (DL) descriptive language. This combination of Object database and DL queries will also widen the power of Video-OQL to deal with exact queries as in database query languages as well as with incomplete and imprecise queries.

6 Conclusion

This paper has presented a conceptual model and an architecture of a surveillance system. There are various tradeoffs in designing the model using UML. The first of which is the choice of video content representation. UML tools are very suited for complex object modeling based on the class concept, yet they seem inadequate to cater for various operations specifically related to video data such as overlapping, intersection and video shots containing other video shots. No direct means of modeling an update operation are also available, as it is the case in creating an object sequence to represent the fact this is history tracing the same objects (one identity) which seems to change some of its related attributes (position, time, colour distribution, orientation, etc.). What is really meant by the Aggregation on the class OBJECT is an update operation. One might argue the case of having a rigid structure (e.g. object sequence) in which all instances of the same object involving in a labeled events are inserted but then how to model the fact that the first instance of the object is included as it is but later instances are updated versions of the first instance?. This leads us to the second tradeoff: having a more abstract model with fewer classes and associations versus explicitly providing rigid structures whenever existing modeling tools cannot cope and hence losing abstraction. On one end of the scale, too much abstraction leads to model portability, implementation-indepenence, universal understanding, extendibility easiness and fewer concepts and relations. On the other end of the spectrum, excessive abstraction certainly gives way to management, maintenance, expressiveness and mapping problems.

Although the OO and DL models seem to complement each other, this work will be facing two main challenges. The first is mapping between DL and OO models or schemas. Tight coupling is preferred in this system to allow on-demand traffic of instances between the video object database and the DLs knowledge base on demand. Fully recognised objects and events are transferred to the video object database permanently (persistent objects and events). Unknown objects (and events) reside in the knowledge base until further information about their content is acquired. If still unrecognised, a decision has to be made whether to transfer them to the video object database or leave them in DLs knowledge base. The second challenge is addressing the limitation of DL in accessing external storage when the number of DL concepts (class) individuals (objects) grows and requires external storage.

References

[Aigrain96] P. Aigrain, H. Zhang , and D. Petkovic, *Content-Based Representation and Retrieval of Visual Media: A State-of-the-art Review*, Multimedia Tools and Applications, 3(3), pp 179-202, 1996.

[Borgida95] A. Borgida, *Description Logics and Data Management*, IEEE Transactions on Knowledge and Data engineering, 7(5), pp 671-782, 1995.

[Eakins99] J.P. Eakins, M. E. Graham , *Content-Based Image Retrieval*, Report to JISC Technology Applications Programme, Jan. 1999.

[Goble96] C. A. Goble, C. Haul and S. Bechhofer, *Describing and Classifying Multimedia using The Description Logic GRAIL*, SPIE Conference on Storage and Retrieval of Still Image and Video IV, San Jose, CA, 1996.

[Hjelsvold95] R. Hjelsvold, R. Midtstraum and O. Sandsta, *A Temporal Foundation of Video Databases*, Proceedings of the International Workshop on Temporal Databases, Zurich, Switzerland, Sep. 1995.

[Ivanov99] Y. Ivanov, C. Stauffer, A. Bobick and W. E. L. Grimsom, *Video surveillance Interactions*, Second IEEE International Workshop on Visual Surveillance, Fort collins, Colorado, USA, June 26 1999.

[Lahlou95] Y. Lahlou, *Modelling complex objects in content-based image retrieval*, In Proceeding of Storage and Retrieval for Image and Video Database III, Vol. 2420, pp 104-115, San Jose, CA.

[Marir93] F. Marir, *A Compiled Inference Approach to the Integration of Logic and Relational Database systems*, PhD thesis, University of Salford, June 1993.

[Oomoto94] E. Oomoto, and K. Tanaka, *OVID: Design and Implementation of a Video-Object Database System.* IEEE Transaction on Knowledge and Data engineering, Vol. 4, No 4, August 1994.

[Orwell99a] J. Orwell, P. Massey, P. Romagnino, D. Greenhill, and G. A. Jones, *A multi-agent Framework for Visual surveillance*, Proceedings of the 10th International conference on Image Analysis and Processing, September 27-29, 1999, Venice, Italy.

[Orwell99b] J. Orwell, P. Romagnino, and G. A. Jones, *Multi-camera Colour Tracking*, Proceedings of the second IEEE International Workshop on Visual Surveillance, June 26 1999, Fort Collins, Collorado, Illinois, USA.

[Regazzoni98] C. S. REGAZZONI, and E. STRINGA, *Content Based Retrieval and Real-Time Detection from Video Sequences acquired by Surveillance systems*, IEEE International conference on image Processing, ICIP98, 3(3),Chicago, Illinoi, USA, 4-7 Oct. 1999.

[Romagnino98] P. ROMAGNINO, T. TAN, and K. BAKER *Agent Oriented Annotation in Model Based Visual Surveillance*, In ICCV, pp 857-862, Bombay, India, 1998

[Sahouria97] E. SAHOURIA, *Video Indexing Based on Object Motion. Image and Video Retrieval for National security Applications: An Approach Based on Multiple Content Codebooks*, The MITRE Corporation, Mclean, Virginia, 1997, USA.

[Smoliar94] S.W. SMOLIAR and H. ZAHNG, *Content-Based Video Indexing and Retrieval* IEEE Multimedia 1(2), pp 62-72, 1994.