# A Conceptual Model for Surveillance Video Content and Event-Based Indexing and Retrieval

Farhi Marir, Kamel Zerzour , Karim Ouazzane and Yong Xue

Knowledge Management Group (KMG), School of Informatics and Multimedia Technology, University of North London, Holloway Road, London N7 8DB, UK
{f.marir, k.zerzour, k.ouazzane,Y.xue}@unl.ac.uk

**Abstract.** This paper addresses the need for a semantic video-object approach for efficient storage and manipulation of video data to respond to the needs of several classes of potential applications when efficient management and deductions over voluminous data are involved. We present the VIGILANT conceptual model for content and event-based retrieval of video images and clips using automatic annotation and indexing of contents and events representing the extracted features and recognised objects in the images captured by a video camera in a car park environment. The underlying video-object model combines Object-Oriented modelling (OO) techniques and Description Logics (DLs) Knowledge representation. The OO technique models the static aspects of video clips and instances and their indexes will be stored in an Object-Oriented Database. The DLs model will extend the OO model to cater for the inherent dynamic content descriptions of the video, as events tend to spread over a sequence of frames.

## 1. Introduction and Related Work

Recently, video surveillance industry, particularly in the UK, has experienced growth rates of over 10 % 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 £2billions. 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 videotapes. The extensive range of commercial and public safety Content-Based Image/Video Retrieval applications includes law enforcement agencies, intruder alarms, collation of shopping behaviours, automatic event logging for surveillance and car park administration. It also include art galleries and museum management, architectural and engineering design, interior design, remote sensing and management of earth resources, geographic information systems, medical imaging, scientific database management systems, weather forecasting, retailing, fabric and fashion design, trademark and copyright database management.

The problem of providing suitable models for indexing and effectively retrieving videos based on their contents and events has taken three main directions. Database research concentrated on providing models to mostly handle static and aspects of data video (with little or no support for the dynamic aspects of video), and classification is process indices are usually performed offline. From the Image Processing point of

P.M.A. Sloot et al. (Eds.): ICCS 2002, LNCS 2329, pp. 41–50, 2002.

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 such as identity, colour and texture. The need for a synergy between database and computer vision approaches to handle video data is inevitable. This collaborative approach involves database modelling, Vision/Image Processing, Artificial Intelligence and Knowledge Base Systems and other research areas [1].

The Database indexing realm includes the works carried by [4] 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. [6] investigated the appropriateness of the existing object-oriented modelling for Multimedia data modelling. He recognised that there is a need for more modelling tool to handle video data as conventional object modelling suffer from three main limitations when dealing with video images and scenes. The video data is raw and its use is independent from why it is created, the video description is incremental and dynamic, and finally the overlapping of meaningful video scenes. Content-based video indexing and retrieval architecture was proposed in [13]. This architecture extends a database management system (DBMS) with capabilities to parse, index and retrieve or 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.

With regards to Image Processing, 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 [2]. Description Logics is also used to build Terminology Servers to handle video data content in a bottom up approach [3]. DLs proved beneficial in terms of dynamic, incremental classification, knowledge extensibility and evolution, precise and imprecise querying abilities. Recently, standard stochastic context-free grammars (SCFG) have been employed to contextually label events in an outdoor car park monitoring [5]. 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. [10] 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 levels where each detected object is assigned a behaviour agent, and Interaction level using situation agents. [9] 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, and structural change (e.g. chair changes position) or lighting effect, with the aid of a Multilevel Perception Neural Network.

## 2. The VIGILANT Video Object Model

VIGILANT is an end-to-end intelligent semantic video object model for efficient storage, indexing and content / event-based retrieval of *(real-time)* surveillance video in a car park environment. The proposed system diagnoses the problems of content extraction, (semantic) labelling and efficient retrieval of video content. The system is intended to automatically extract moving objects, identify their properties and label the events they participate in (e.g. a car entering the car park).

- Capture: A suitable means for capturing surveillance video footage with associated semantic labelling of events happening in the car park had to be implemented. An unattended camera was set up to record events (e.g. a car entering the car park) and a frame grabber is needed to process this information and make it available for the next stage: indexing and content extraction.
- Segmentation and Content Extraction: The most desirable segmentation of surveillance video material is clearly separation into individual meaningful events (a person getting off the car) with participating objects (a car and a person). Semantic video segmentation is a highly challenging task and usually segmentation based on simpler breaks of the video (e.g. a camera break) are opted for [12]. This involves examining the structure of the video and using statistical measures (e.g. colour histograms) [13]; cuts are determined from scene to scene. The VIGILANT projects attempts to exploit the frame-to-frame object tracking as a means of providing shot cuts, whereby each new object appearing in the scene gives rise to a potential event which ends when the tracking of that particular object finishes (e.g. object leaves the camera view). At each frame various statistical features (colour, position, time, speed) are extracted from object at each frame and the content (especially identities of objects and their activities) is incrementally built as objects are tracked from frame to frame using DLs knowledge base (events) and Neural Networks (objects).
- Content / Event Indexing: Different indexing schemes have been researched [1] and these are classified into various abstraction levels depending on the types of features being extracted. Conventional database indexing tools have proved inadequate for video labelling due to the dynamic and incremental nature of video content [6]. The VIGILANT system will couple Object Oriented Database (OODB) tools with Description Logics (DLs). The former will be employed to model the static aspects of video (contextually independent content such as object identity and maybe colour) using concepts of classes and relationships. The latter will deal with dynamic contents that spread over a sequence of frames (object activities) by means of concepts and roles. Various issues will need to be addressed at this stage primarily concept tractability, DLs reasoning tools and the ability of the DLs knowledge base to access secondary storage.
- Retrieval: The way video clips are retrieved is strongly tied to the way in which they are indexed. Current retrieval practice relies heavily on syntactic primitive features such as colour, texture, shape and spatial arrangement, and little work has been done into indexing (and hence retrieval) based on semantic contextual features (events) [1]. To respond to this need, this work proposes to research and

implement queries for the second and third level of abstraction. We proposed to develop a real-time Video-Object Query Language (Video-QQL) which will uses the semantic content, indexes and annotations generated by the Index Module. The 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 video clips showing red cars entering the car park between 11am and 2pm". (a full description of the VIGILANT's proposed UML modelled can be found in [14]).

## 3. VIGILANT's Architecture

The complete architecture of VGILANT is shown in Figure 1 with the parts to be addressed in this paper in **bold**. The Video-Indexing Module parses the VIGILANT test file, which encodes the extracted features and recognised objects, to either populate the video-object database (if the object and events are well recognised) or to populate the description logics knowledge base (if not yet known). Then indices are generated in each video-object to reference a frame or a sequence of frames (clips). The Schema Manager Module will implement a mechanism and channels of communication between the video-object and description logics (DL) schemas to reflect the dynamic changes of the video content. The Real-time Video-OQL Module will provide visual facilities, through the VIGILANT user interface, for the user to access and browse retrieved images or video clips of a particular scene using the indices and content of the video already stored in the video-object database and DL knowledge base.

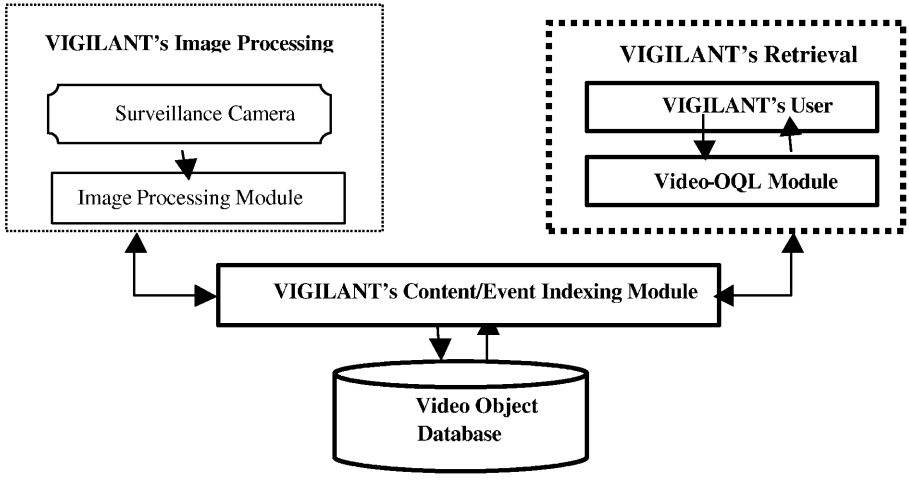


**Fig. 1.** VIGILANT's Proposed Architecture

## 4. VIGILANT Content and Event Indexing

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 [7, 8]. 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.

Figure 2 summarises the work of the collaborating three schemas (video-object, description logics and the schema manage).

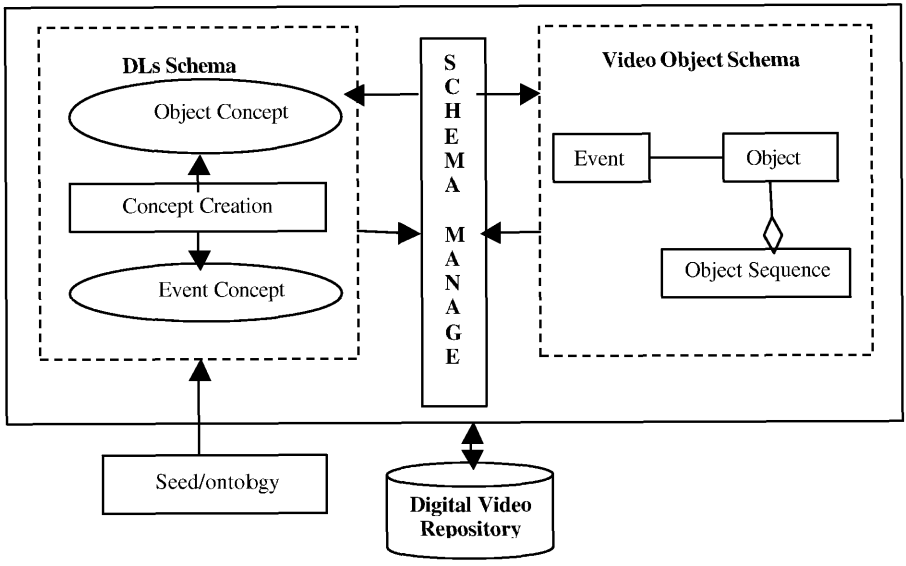


**Fig. 2.** The VIGILANT's Proposed Content/Event Indexing Module shows the collaborative work between both the DLs schema and the Video Object schemas through the Schema manager

### 4.1 The Schema Manager and The Video-Object Schema

Video content is difficult to model within a structured data model such as 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 VIGILANT's both Video Object and Schema Manager's schemas are described in [14], we here further describe the DLs schema.

### 4.2 The Description Logics Schema

The *DLkb* domain knowledge represents concepts held by objects in the class, and is defined by the Equation

$$DLkb(C) = <F[fn,method],\ R[rn,rolebvr],CM[qc,val]> \quad (1)$$

where $C$ is a concept in the knowledge base representing an object of a given class (moving objects or events). *Fn* is an object feature (e.g. colour, type,...) which is extracted using a procedure *method*. *R* is a relationship connecting a feature to a concept (e.g. has_colour) and *role* is the role's semantic behaviour (e.g. has_colour). *CM* is a condition mapper that converts the condition value(s) of the query into a certain data *value* whose data type is the same as that of the data values of the *F* (afternoon gets converted into Time > 12:00 am).

Two basic concepts are needed for the VIGILANT's project to reflect the two directions of the undertaken research (content and event based retrieval). Content indexing is represented by *moving object concepts* (car, person) and events by *event concepts*.

Object features include object identity, colour, speed, position within a particular frame, speed and size. These attributes are available from the image-processing module, except object types which require utilising the Semantic Neural Net. When an object appears for the first time a **hypothesis** is made about its identity, then using other syntactic object features (colour, height, width, and area) the hypothesis is strengthened until at the end of the tracking process the hypothesis becomes a **thesis**. Whenever an object is processed a new version of that object concept is created in the DLs domain knowledge.

Event concepts are created in a similar manner. The first version of the object concept calls for an event concept to be created with the following features:
(1)A temporary event type, which is then strengthened as the tracking, continues over time. (2)The event start time is the first object's time. (3) The end time is the last object version's time, and (4)An object sequence containing all the versions of one object (or many) involved in that particular event.

Equations (2) and (3) show a formal definition of a moving-object and an event concept, while in figures 4 and 5 are samples of the definition of the concept Event in both OO and DLs environments

<MovingObject>
Feature{[Type, Ext_Type],[Colour, Ext_Colour], ...} (2)
Role{[Type, Has_Type], [Colour, Has_Colour],...}
Condition_Mapper{[Red,[120,24,167], [fast, 90km/h],...}

<Event>
Feature{[Type, Ext_Type],[StartTime, Ext_STime], ...} (3)
Role{[Type, Has_Type], [STime, Has_Has_STime],...}
Condition_Mapper{[Type,[Car_Entering,Person_Leaving,...],}

```
Class: VideoShot{
StartTime, EndTime, BackgroundClour, NumOfObjects, etc;
Car("door open", "window close", etc);
Person("walking", "getting in Car", etc);
//Event part- Relationships between objects
Person,Car ("get in", "get out", ect);
Car, CarPark("Enter", "leave", etc);
}//End of videoShot Class
```

**Fig. 3.** Example of typical Class concept Definition of an Event

```
(defind-concept VideoShot(AND
  (define-concept Colour(aset red green yellow ...)(ALL StartsAt Num)
  (ALL EndsAt Num)(ALL HasNumberOfObjects Num)
  ;Composition part
  (ALL ContainsComp VideoShotComp)
  (define-concept State (aset CLODE OPEN)
  (define-primitive-concept Door CarComp)
  (define-primitive-concept Window CarComp) etc...
  ;Event part
  (ALL ContainsEvent Event)
  (define-concept EventType (aset GetIn GetOut Enters Leaves ...))
  (define-concept GetIn (ALL Event (ALL HasEventType(aset GetIn))
                                   (ALL HasFirstInteractor Person)
                                   (ALL HasSecondInteractor Car))) etc
```

**Fig. 4.** Example of typical DLs concept Definition of an Event

## 5. VIGILANT's Content and Event Based Retrieval

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.EventType = "Car Entering Car Park" AND
      VideoClip.Time = IN(11am, 3pm) AND
      Cars.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 logics (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.

In the VIGILANT project, similar to [11], a query Q is modelled as:

$$Q=dm(rf[ac,c]) \tag{4}$$

where $dm$ is the domain or the set of classes to be retrieved as a result of the query $Q$, $rf$ is the retrieval function, $ac$ is the attribute(s) for which the condition(s) is/are specified, and $c$ is the condition value(s) of the attribute(s).an example query is "find video clips of cars". This is expressed as:

```
ObjectSequence(find_equal[VideoTape.VideoShot.Frame.Object.Type,Car])
```

This is regarded as a primitive query and more sophisticated queries can be built by combining primitive queries. Consider the following query statement: "find video

clips of red cars entering the car park between 11:00 am and 12:00 am", this can be expressed in our model as:

```
ObjectSequence(

    equal_type[videoTape.VideoShot.Frame.Object.Type,Car]&&

    equal_color[videoTape.VideoShot.Frame.Object.Color,Red] &&

    equal_event[videoTape.VideoShot.Event.Type,Entering] &&

    time[videoTape.VideoShot.Frame.Object.StartTime,11am&12am]
```

## 6. Conclusions

The VIGILANT's proposed model couples Description Logics Domain Knowledge schemas which deal with the static content with Object-Oriented schemas which handle the evolution of video content. Challenges associated with this scheme include DLs inability to handle large volumes of data. This 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. Another equally important challenge will be the mapping of both OO and DLs schemas.

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.

## 7. References

1. J.P. Eakins, M.E. Graham, "Content-Based Image Retrieval, Report to JISC Technology Applications Program", 1999.
2. 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.

3. C. A. Goble, C. Haul and S. Bechhofer, "Describ ing and Classifying Multimedia using the Description Logics GRAIL", SPIE Conference on Storage and Retrieval of Still Image and Video IV, San Jose, CA, 1996.
4. R. Hjelsvold, R.Midtstraum and O. Sandsta, "A Temporal Foundation of Video Databases", Proceedings of the International Workshop on Temporal Databases, 1995.
5. Y. Ivanov, C. Stauffer, A. Bobick and W. E. L. Grimsom, "Video Surveillance Interactions", Second IEEE International Workshop on Visual Surveillance, 1999.
6. 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, 1994.
7. 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, 1996.
8. J. Orwell, P. Romagnino, and G. A. Jones, "Multi- Camera Colour Tracking", Proceedings of the second IEEE International Workshop on Visual Surveillance, 1999.
9. 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), 1999.
10. P. Romagnino, T. Tan, and K. Baker, "Agent Oriented Annotation in Model Based Visual Surveillance", In ICCV, pp 857-862, 1998
11. A. Yoshitaka, S. Kishida, M. Hirakawa and T. Ichikawa, "Knowledge Assisted Content-Based Retrieval for Multimedia Databases", IEEE '94.
12. 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 79-202, 1996.
13. S.W. Smoliar and H. Zhang, "Content-Based Video Indexing and Retrieval", IEEE Multimedia 1(2), pp 62-72, 1994.
14. K. Zerzour, F. Marir and J. Orwell, A Descriptive Object Database for Automatic Content-Based Indexing and Retrieval of Surveillance Video Images and Clips, ICIMADE'01 International Conference on Intelligent Multimedia and Distance Education, Fargo, ND, USA, June 1-3