OVIS: ontology video surveillance indexing and retrieval system

Mohammed Yassine Kazi Tani, Abdelghani Ghomari, Adel Lablack, Ioan Marius Bilasco

Int J Multimed Info Retr

REGULAR PAPER

OVIS: Ontology Video Surveillance Indexing and Retrieval System

Mohammed Yassine Kazi Tani[1].
Abdelghani Ghomari[1].
Adel Lablack[2].
Ioan Marius Bilasco[2].

Abstract Nowadays, the diversity and large deployment of video recorders results in a large volume of video data, whose effective use requires a video indexing process. However, this process generates a major problem consisting in the semantic gap between the extracted low-level features and the ground-truth. The ontology paradigm provides a promising solution to overcome this problem. However, no naming syntax convention has been followed in the concept creation step, which constitutes another problem. In this paper, we have considered these two issues and have developed a full video surveillance ontology following a formal naming syntax convention and semantics that addresses queries of both academic research and industrial applications. In addition, we propose an Ontology Video-surveillance Indexing and retrieval System (OVIS) using a set of Semantic Web Rule Language (SWRL) rules that bridges the semantic gap problem. Currently, the existing indexing systems are essentially based on low-level features and the ontology paradigm is used only to support this process with representing surveillance domain. In this paper, we developed the OVIS system based on the SWRL rules and the experiments prove that our approach leads to promising results on the top video evaluation benchmarks and also shows new directions for future developments.

Keyword Video Surveillance Ontology, Video-Indexing, Crowdsourced events, Semantic-Gap, Naming Syntax Convention, OVIS System, SWRL Rules.

Mohammed Yassine Kazi Tani (Corresponding author)
yassine.kazi@gmail.com

Abdelghani Ghomari
ghomari65@yahoo.fr

Adel Lablack
adel.lablack@univ-lille1.fr

Ioan Marius Bilasco
marius.bilasco@univ-lille1.fr

1 RIIR Laboratory, Computer Science Department, Exact Sciences and Applied Faculty, University of Oran1 Ahmed Ben Bella, Oran, Algeria.

2 Research Center in Signal Informatics and Automatic of Lille (CRIStAL), University of Lille 1, France.

1 Introduction

During the last few years, the semantic multimedia indexing process becomes a major research topic in computer vision and machine learning, due to the huge increase in the size of recorded data and the diversity of application domains like sport, broadcasting, news, cooking, surveillance etc. The need to find an effective tool to index and store this large volume of data for future uses must be satisfied. Therefore, many scientists explore new ways to improve the existing approaches or to develop new ideas in the video indexing domain. Currently, there are two dominant categories of indexing approaches. The first one consists of using low-level features in an automatic system where the second one uses metadata or keywords in a manual high-level system. Even if the combination of those two approaches could offer an efficient indexing system, it does not overcome completely the semantic gap problem.

Generally, the semantic gap outlines the differences between the video sequence information perceived by human experts and the interpretation of the results obtained from low-level analyzers. Several works used an ontology-based approach to handle this problem. For instance, Kless et al. [1] present a thesaurus or taxonomies as the best method for the creation of ontology. Atta et al. [2] develop a framework based on a network of scalable ontologies that index a large repository of special effects video clips. This proposed framework enables intelligent retrieval for the film post-production domain. Mariano et al. [3] realize a system to answer ontological queries that include many specific optimizations. On one hand, they exploit the ABox (assertion component) skills that generally represent the assertion component or instances of the class. On the other hand, they respect the TBox (terminological component) properties that generally describe a system in terms of controlled vocabulary. Scherp et al. [4] propose the notion of a core ontology, like a system based on the logical notion of reducibility, rather than on the distinction between generic and domain ontologies. Benmokhtar et al. [5] use an ontology paradigm integrated with a neural network approach for detecting a concept purpose. Rector et al. [6] consider annotation using existing ontologies as a good practice. Smith et al. [7] present a specific theory of ontology called Ontological Realism to build high-quality ontologies, using both philosophical views (i.e. the study of the existing entities and the way that they are related to each other) and computer science ones (i.e. the

conceptualization of a domain). Hernandez-Leal et al. [8] propose the use of ontologies as a tool to reduce the semantic gap between low and high-level information, and present the foundations of ontology to be used in an intelligent video surveillance system.

Our main contributions in this paper are as follows:

- To our knowledge, this is the first work that introduce the idea of following a unified ontology structure to formalize both the objects and the actions as well as the events in surveillance domain. We have also modeled the non-sequential relationship between events using Allen's interval algebra.
- Secondly, we extend the video surveillance domain representation by considering new concepts characterizing events in the industrial domain, whereas the approaches proposed in the literature focused only on academic aspects.
- Our ontology is more complete compared with those proposed in the literature and gives a large coverage of important objects and events in the surveillance domain.
- Fourth, this is the first work that proposes a diversity of applications, which are not limited to indexing purposes, but also applies to scene description, benchmark creation, etc.
- Finally, we contribute the use of the proposed ontology in a rule-based indexing and retrieval approach, by generating SWRL rules. These rules are deployed in the middle and high-level step of an event detection process, that are supplied with low-level descriptors, rather than using the classical descriptors and classifiers in all event detection steps.

The remainder of this paper is organized as follows: Section 2 focuses on related works. In Section 3, we present our ontology with its naming syntax convention and semantics. Section 4 describes a comparative study with other ontologies. Section 5 presents the various domains of application of our ontology. The OVIS system architecture is illustrated in Section 6. Preliminary of the Video-Surveillance Ontology Indexing and Retrieval system developed here are given in Section 7. The final section provides some concluding remarks.

2 Related works and background

In video surveillance applications, an ontology could be used in the indexing process to support the detection of an event such as an abnormal behavior, crowd situations of people or traffic monitoring. Indeed, the description of a domain covered by the ontologies and the reasoning results that are generated increase the accuracy of the indexing process. Several approaches have been proposed using an ontology.

Soner et al. [9] use an ontology to extract instances from a document corpus, and add them to the knowledge basis. Till et al. [10] handle the problem of using a variety of languages and propose a distributed ontology language DOL that allows to use its own preferred ontology formalism considering the interoperability with the others. Ballan et al. [11] recognize events in broadcast news and video surveillance domains by embedding knowledge into the ontology. Bagdanov et al. [12] use a multimedia ontology that contains visual prototypes representing each cluster that acts as a bridge between the domain and the video structure ontology. They present a system that gives a solution to the semantic gap between the high-level concepts and low-level descriptors. Bertini et al. [13] classify the events and the objects that are observed in video sequences by adding new instances of visual concepts to their ontology through updating mechanisms of the existing concepts. This approach used in both generic and specific domain descriptors attempts to identify visual prototypes that represent elements of visual concepts. To overcome the problem of manual rules creation by a human expert, Bertini et al. [14] proposed an adaptation of the First Order Inductive Learner (FOIL) technique for Semantic Web Rule Language (SWRL) [15], called FOILS. Xue et al. [16] propose an ontology-based surveillance video archive and retrieval system. Lee et al. [17] classify and index video surveillance streams through the creation of the framework called Video Ontology System (VOS). Snidaro et al. [18] use a set of rules in SWRL language for event detection purposes in the video surveillance domain.

The problem that arises in the use of an ontology paradigm to support the indexing process is to find the best way for the creation of this ontology. Moreover, some previous works have used the ontology tool and demonstrated their efficiency in helping and managing the indexing and retrieval process. However, they have based their experimental studies on events that consider only one or two relevant objects in a video clip. In the contextual cases, they consider events such as an abandoned or stolen object; whereas in the moving cases, they consider events like a person who walks from right to left, an airplane flying, and so on. The problem that arises is how to ensure the efficiency of the ontology in the indexing and retrieval process when the user requires multiple object events or crowd sourced events considering a set of relevant objects (e.g. queries about a regular group of people walking, a group of people running, a group of people splitting, etc.).

Calavia et al. [19] developed an intelligent video surveillance ontology system that analyzes objects movements and identify abnormal and alarming situations. However, the domain application covered by the documentation is not consistent with the ontology representation. Papadopoulos et al. [20] proposed a genetic algorithm for optimizing the size of each ontology element (e.g. concepts, etc.). In this way, they consider the variable relevant importance of global and local information to detect the different ontology elements. Nevertheless, the relationship named “some/some” is used instead of “all/some”. Sawsan et al. [21] constructed a video movement ontology for automatic annotation of human movement’s purposes in the classic Benesh notation. However, it is not clear whether their ontology is formal or not, and there is something wrong in the use of the “Is-A relation” in a non-transitive way as the relation between the two concepts (i.e. media and video) in multimedia representation part. Trochidis et al. [22] proposed a well-structured ontology approach to model life events described as a graph of connections between concepts with representing a particular domain. Nevertheless, this approach has some limitations about mainstream ontology and its application in video analysis. Bohlken et al. [23]

considered the problem of high-level scene interpretation suggesting a novel architecture based on the generation of rules from an OWL-DL ontology. However, the concept of vehicle entering a zone is not conceptual, because it represents an action between vehicles and zone concepts.

Nevatia et al. [24, 25] developed two languages called VERL (Video Event Representation Language) and VEML (Video Event Markup Language), for describing an ontology of events, and annotating instances of the events respectively. However, a confusion occurs between the object language that describes the referent in the subject domain and the meta-language that defines this object language. Bai et al. [26] presented a video semantic content analysis framework based on ontology. A high-level concept is described referring to this domain of application and combined with the MPEG-7 standards for expressing low-level content analysis algorithms. Nevertheless, this ontology confuses between the relation "Is-A", and "Instance-Of". For example, the combination of many algorithm instances with the "Is-A" relation in an ontology replaces the "Instance-Of" relation. SanMiguel et al. [27] proposed an ontology-based approach to represent the prior knowledge of a video event analysis consisting of two types of knowledge: the application domain and the analysis system. The domains knowledge involves all the high-level semantic concepts (objects, events, context, etc.), while the system knowledge includes the abilities of the analysis system (algorithms, reactions to events, etc.). However, this ontology determines only the best visual analysis framework (or processing scheme), and does not handle the inference for object tracking and event detection.

The detection process is often represented using three levels: low-level, middle-level and high or semantic one. Different approaches could be found in the literature using descriptors in the low and middle level and classifiers in the semantic one. Utasi et al. [28] proposed a statistical descriptor approach detecting three kinds of events: regular activity, running and splitting. This approach consists in using a background extraction technique followed by calculating the optical flow of foreground pixels. However, it does not detect many events like walking and formation. Chan et al. [29] used a model based on global properties to detect events such as walking, running, splitting, formation and local dispersion. Their approach characterizes the crowd flow using a dynamic texture. However, this model does not process overlapping that could occur between events.

Other semantic based approaches participate on TRECVID Surveillance Event Detection (SED) task 2016. Markatopoulou et al. [30] proposed a system for surveillance event detection based on fisher vector encoding method and SVM models to learn how to separate each activity from the others. However, this approach detects many false alarms. Zhao et al. [31] used different approaches to detect surveillance events, and their overall system consists of two parts: the retrospective part and the interactive part. The retrospective part implements Pedestrian detection, Pedestrian tracking and event detection. The interactive part, determines the events after fixing the false and missing rate. However, this method considered a limited number of events.

After a deep analysis of all the problems noted above, we introduce an innovative approach in this work, by creating an ontology and implementing the OVIS indexing and retrieval system that considers all the above observations. In this paper, we have reconsidered our previous works [32, 33] where we presented only our SWRL rules based approach allowing to handle a video surveillance ontology to detect a single or multiple objects events. In the present work, we have improved and extended our previous approach by implementing a complete Video-Surveillance ontology with a very precise step creation syntax. This extension describes numerous objects and events that can appear in a video surveillance domain. Furthermore, the creation of our ontology is more complete considering new concepts that characterize events in the industrial domain. Moreover, we have not based our approach only on indexing purposes, but also in scene/video description and benchmark creation domains. Finally, we used our ontology to prove the efficiency of our approach. We have generated SWRL rules for event detection, and used them in both middle level and high or semantic level, rather than using the classical descriptors and classifiers in all event detection steps. These SWRL rules use results that are supplied by low-level descriptors for event detection purposes. We have also tested the performance of these rules by experimenting videos from the PETS 2012 challenge [34] and SED task from TRECVID challenge [35]. The PETS challenge represents multiple view sequences that handle different crowd activities and contain multiple objects events (e.g. group walking, group splitting, etc.). It allows providing the existence of each event, in these sequences, with start/end as well as transitions between these different events. The main goal of TRECVID (The TREC Video Retrieval Evaluation) is to promote progress in content-based analysis and retrieval from digital video via open metrics-based evaluation. The Surveillance Event Detection (SED) task focuses on developing new approaches able to detect observations of different events. It consists in a subset of 10 hours of videos recorded using multi-camera derived from Gatwick airport. Seven events are identified: PersonRuns, CellToEar, ObjectPut, PeopleMeet, Embrace, PeopleSplitUp, and Pointing. In our work, we focus on three events: PersonRuns (Running), PeopleMeet (Formation) and PeopleSplitUp (Splitting).

3 Ontology hierarchy description

Our ontology creation based on a naming syntax convention includes most of the existing concepts appearing in a Video-Surveillance domain. Indeed, our approach is an improvement of SanMiguel et al. [27] work, where we used the same modelling that defines the high-level relationships between objects and events to compose single and multiple-object activities. We believe that this modelling is the most suitable to represent the video surveillance domain. However, regarding the different levels that compose ontologies, we use this modelling only in the high level expressed as level 2 "L2" in Table 1 and Table 3 below. We also introduced new concepts representing events used in the industry like intrusion.

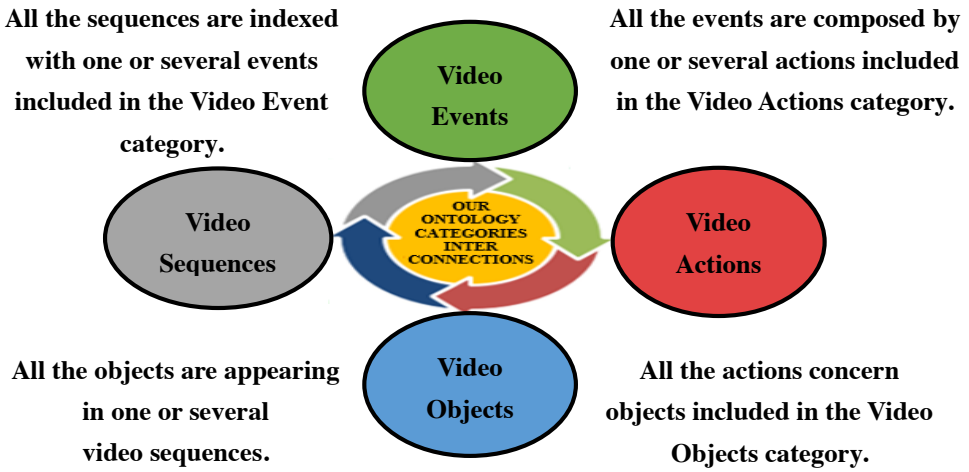

Fig. 1: Interconnection between the four main categories of concepts in the proposed ontology.

Fig. 2: Illustration of the "Is-a" relation representing the four main categories of our ontology.

3.1 Semantics of our ontology

Generally, the semantic interpretation of video sequences is the critical step in an indexing process. It corresponds to the translation of the low-level features extracted from the visual analysis module into the video sequence meaning. Here, we used an ontology paradigm as a tool characterizing a video surveillance system. Our semantic ontology is represented by different interacting concepts where each concept carries one or more properties as a "Data_Property" described in detail below.

3.1.1 Ontology Concepts

The concepts of the ontology proposed here correspond to the categorization of the video surveillance domain, regarded as generalization/specialization relationships. In order to have a complete representation of all the objects or events that can happen in a video surveillance domain, we formally divide the ontology into four main categories concepts representing Video_Actions, Video_Events, Video_Objects and Video_Sequences.

Figure 1 and Figure 2 above show how the four main categories of concepts are linked together, where each one forms an interconnection with the other. First, all the available video sequences in the video database must be indexed with one or more concepts that appear in the Video Events category. A relationship exists between Video Events and Video Actions since an event is a composition of one or more actions. Furthermore, another connection exists between Video Actions and the Video Objects category. Actually, scene description is formed by the Video Object components category where the actions occur. Finally, Video Sequences category encloses objects that belong to the Video Objects category. In the following, we describe each category.

Video Objects

The different kinds of objects represent the principal entities that interact in the video sequences. The Video_Objects category involves all the objects that can appear in a video surveillance scene. A large variety of objects interacts with each other to create a video surveillance action. According to their mobility skills, the objects can be assembled into two main categories. The Contextual Objects having no mobility skills and the Mobile Objects with mobility skills. Eventually, we can add a third category representing an image region of interest (ROI), characterized by Low-Level Features. It represents all the data extracted from the visual analysis module. Table 1 below illustrates these three categories dividing the Video Objects into five levels (from L1 to L5).

- The category of Contextual Objects is divided into two sub-classes representing Fixed Objects (that can never be moved by other objects) and Portable Objects (that have the possibility of moving with other objects). Fixed Objects correspond to those of human creation (e.g. air conditioners, panels, etc.) and natural objects (e.g. grass, land, etc.). Portable Objects represent all general and self-using objects (e.g. a box, a chair, a cellphone, etc.).
- The category of Mobile Objects encloses four sub-classes that have the ability of self-moving, such as animals, airplanes, trains, boats, humans and traffic. Traffic corresponds to bicycles and motorcycles. The Human subclass corresponds to a human entity like a person or Group-Of-Person.
- The category of Low-Level Features includes all the results of the low-level analysis module that could be used to help the indexing process for event detection purposes, such as blocks, bounding box, frame, etc.

Video Actions

In video surveillance domain, the action concept represents the behavior of different objects detected in the video sequence in a time frame window. This category includes the actions that can be expected in video surveillance events. Therefore, several varieties of occurring objects can produce multiple kinds of actions. Generally, we can expect five categories of objects that belong to Video Objects category: airplane, boat, train, traffic (road traffic) and humans. In our ontology, we divide these actions into different sub-classes according to the nature of the different objects interactions:

- Interactions with environment: such as human walking, human stopping, airplane landing, etc.
- Interactions with humans: human attacking, human meeting, etc.
- Interactions with objects: such as a human breaking an object, etc.

As described in Section 3.2, Table 2 below presents a summary of the different sub-classes of the Video Actions parts of our ontology, illustrated at different levels (L1 to L3). We separate all the actions according to their degree of priority. The first sub-class degree of the Video Action category expressed in level 2 represents the action actors and properties with which they interact (e.g. Human_And_Objects_Actions, etc.). The second sub-class degree expressed in level three corresponds to the actions themselves. Among them, we note those that are widely reported in the literature such as (Split, Met and Ran) and those related to the industrial domain. Thus, for example walking and running events represent generally a group-Of-person that interacts with the environment by multiple walked and ran actions. Intrusion event related to industrial domain represents the attempt of a person to enter a restricted zone. Considering new concepts that characterize events in the industrial domain, is one of the objectives of our proposal work contribution.

Table 1: Video_Objects Hierarchy.

L 1	L 2	L 3	L 4	L 5
Video_ Objects	Contextual_ Objects	Fixed_Objects	Human_Creation _Objects	Air-Conditioner / Building / Electrical-Pole / Equipment / Floor / Panel / Parking-Lot / Road / Stairs / Stairs-Barrier / Wall / Zone / Glass Barrier
			Natural_Objects	
		Portable_Objects	General_Using	Box / Chair / Door / Plant / Reception-Desk / Surveillance-Camera / Table / Window / Curtain / Sofa
			Self_Using	Cellphone / Document / Luggage (Bag, Suitcase)
	Mobile_ Objects	Airplane		
		Boat		
		Train	Long-Distance-Train / City-Tramway / Underground	
		Animal		
		Human	Person / Group-Of-Person	
		Ground_Traffic	Bicycle	
			Ground-Vehicle	Bus / Car / Truck
			Motorcycle	
	Low_Level_ features	Bounding-Box (BB)		
		Frame		
		Major_Bounding-Box (MBB)		
		Temporary_Bounding-Box (TBB)		
		Temporary_Group-Of-Person (TGP)		
		Blocks(B)		

Table 2: Video_Actions Hierarchy.

L 1	L 2	L 3
Video_ Actions	Train_And_Environement_Actions	Fire_detected / Arrival_detected / Stationed
	Train_And_Human_Actions	Upped/ Downed / Crossed_Forbidden_Zone
	Boat_And_Environement_Actions	Navigated
	Boat _And_Human_Actions	Threw_bag
	Boat _And_Objects_Actions	Approached_bank / Drived_away_the_bank
	Airplane_And_Environement_Actions	Crashed_Off / Flew / Landed / Took_Off
	Airplane_And_Human_Actions	Disembarked / Embarked
	Airplane_And_Objects_Actions	Registered_Luggage / Took_Luggage
	Human_And_Objects_Actions	Downed_Stairs / Read _Document / Broke_Object / Browsed_Object / Counted_Ground-Vehicule / Counted_Human / Crossed_Virtual-Line / Left_Luggage / Put_Object / Removed_Luggage / Rested_On_Chair / Smoked_Cigarette / Upped_Stairs
	Human_And_Human_Actions	Attacked / Chased / Evacuated / Fell_Down / Flow_Opposed / Formed / Fought / Helped / Hit / Local_Dispersed / Met / Split / Stole / Talked / Waited
	Human_And_Environement_Actions	Appeared / Counted_Speed / Disappeared / Entered_Area / Face_Recognized / Fell / Intruded / Left_Area / Loitered / Overcrowded / Ran / Skateboarded / Slipped / Stopped / Trespassed / Walked
	Traffic_And_Object_Actions	Crashed_Object
	Traffic_And_Environement_Actions	Parked

Video Events

The video event concept represents a composition and succession of one or several actions appearing in a video sequence. In the present ontology, the Video_Events category encloses all the different events that could happen in a video stream. Each event representing the formation of actions encloses one or several relevant objects that interact with each other.

Table 3 describes the four levels (L1 to L4) representing the different subclasses of the Video Events parts in our ontology. Each level corresponds to a degree of priority, as described in Section 3.2. The first degree is related to the number of relevant objects and divides our Video Events category into two main subclasses representing respectively single and multiple object events. The second degree is related to the nature of objects represented by seven types: Group-Of-Person, Multiple_Ground-Vehicle, Airplane, Train, Boat, Person and Single_Ground-Vehicle. The final degree corresponds to the interaction between these objects and the other concepts such as human, environment or objects.

Video Sequences

The Video-Sequences category represents the class of all the videos indexed by the OVIS system and the instances represent the Video Database.

After presenting the Ontology concepts part, in the following we describe DataProperty and ObjectProperty parts.

3.1.2 Ontology DataProperty

The Data_Property represents the real information related to individual's concepts. In our ontology, Data Property includes all the properties related to one or more concepts. Table 4 displays the DataProperty hierarchy divided into three levels (L1 to L3). The top level is split into seven sub-classes related to the types of DataProperty, such as Event properties, Frame Properties, etc Each of them enclosing one or more data properties like Event_Place (if the event represents indoor or outdoor events), Number_Frame, etc.

Table 3: Video_Events Hierarchy.

L 1	L 2	L 3	L 4
Video_ Events	Multiple_Objects_ Events	Group-Of-Person_Events	Interaction_Group-Of-Person_And_Environement Interaction_Group-Of-Person_And_Human
		Multiple_Ground-Vehicle_Events	Interaction_Multiple_Ground-vehicle_And_Objects
	Single_Objects_ Events	Air-Plane_Events	Interaction_Airplane_And_Environement Interaction_Airplane_And_Human Interaction_Airplane_And_Objects
		Train_Events	Interaction_Train_And_Environement Interaction_Train_And_Human
		Boat_Events	Interaction_Boat_And_Environement Interaction_Boat_And_Human Interaction_Boat_And_Objects
		Person_Events	Interaction_Person_And_Environement Interaction_Person_And_Human Interaction_Person_And_Objects
		Single_Groud-Vehicle_Events	Interaction_Single_Ground-Vehicule_And_Environement

Table 4: Top_Data_Property Hierarchy.

L 1	L 2	L 3
Top_Data_Property	Event_Properties	Event_Place / Nature_Event
	Detected_Objects	Bottom_Left_Point_X / Bottom_Right_Point_Y / Detected_In_Frame / Direction / Ended_F / Height / ID / Leaving_Object_Way / Major_BB / MBB_True / Number / Number_Of_Person / Posture / Speed / Started_F / Top_Left_Point_X / Top_Right_Point_Y / Weight / etc.
	Entering_Exit	
	Frame_Properties	Number_Frame / Number_BB_In_Frame / Number_MBB_In_Frame / Started_MBB / etc.
	Type	
	Time	
	Video_Sequence_Properties	Video_URI / Number_Of_Frame / Started_F_Formed_Event / Ended_F_Formed_Event / Took_Place_Before_R / Started_R / etc.

3.1.3 Ontology ObjectProperty

The Object_Property concerns the concepts of the ontology interactions and is divided into three levels as shown in Table 5. A complete representation of all interactions between our ontology concepts is obtained by subdividing the top level into three subclasses reflecting the interaction between the objects. We can consider two categories of objects like Humans and Objects (referring to sub-categories of Video Objects other than humans); while, the interactions represent Human_Against_Human, Human_Against_Objects, and Objects_Against_Objects. In the last level, each type of interaction encloses its ObjectProperty, such as Asked_Direction, Walked_Around, Detected_In, etc. For describing some examples of DataProperty and ObjectProperty related to Objects, the Figure 3 presents properties of Group-Of-Person and Bounding-Box Objects.

3.2 Syntax of our ontology

We have created a naming syntax convention composed of several rules to obtain a complete and consistent ontology. First, all the new concepts created in the Video_Event Category must be named in the same way as the previous ones in this category. Each concept is composed of three parts:

- The interaction (each concept starts with an interaction name).
- The name of objects existing in the Video Object category.
- The property with which this object interacts.

Furthermore, all the concepts of our ontology are generalized, and their details are classified in the form of Object_Properties and Data_Properties. As a typical example, the notion of time, posture, position and interaction are represented as Object_Properties or Data_Properties, and not under the sub-event class. Moreover, the concept naming duplication must be avoided and Multiple_Event separated from Single_Event, Human_Actions from Objects_Actions, as well as Person Events from Objects Events. In addition, the Data_Properties and the Object_Properties must be generalized and the duplication avoided. For instance, instead of having a Data_Property called Name-Of-Animals for the concept animal and another one called Name-Of-Building for the concept building, we introduce a generalization and call the DataProperty "Name," using it for both concepts. The events and actions must be separated according to their degree of priority. Concerning the events, the first degree is attributed to multiple or single object, the second degree is linked to the nature of the object (Group-Of-person, Person, Ground-Vehicle, etc.), and the third one represents the object interaction (with Human, with Environment or with Object). As far as the actions are concerned, the first degree is set according to the action actors and the concepts with which they interact, while the second degree is set with the action itself.

In the present work, the Group-Of-Person Object is regarded as the formation of two or more individuals. To be more concise, each word in the formation of a concept, an Object_Property or a Data_Property, must start with a capital letter. Relationships are determined between concepts

Table 5: Top_Object_Property Hierarchy.

L 1	L 2	L 3
Top_Object_Property	Human_Against_Human	Asked_Direction / Chased /Formed_Final_Meta_Group / Attacked / Formed_With / Had_Diff_End_Position / Had_Different_Direction / Had_Same_Start_Position / Had_Started_Meta_Group / Helped / Hit / Met_With / Pushed / Split_With / Spoke_With / Stole / Walked_With / etc.
	Human_Against_Objects	Walked_Around / Attempted_To_Open /Browsed_On / Downed / Left / Loitered_In / Occurred_In / Put / Rested_On / Stood_Near / Upped / etc.
	Objects_Against_Objects	Detected_In / Crashed_With / Flew_In / Landed_In / Parked_In / Represented / Took_Off_From / etc.

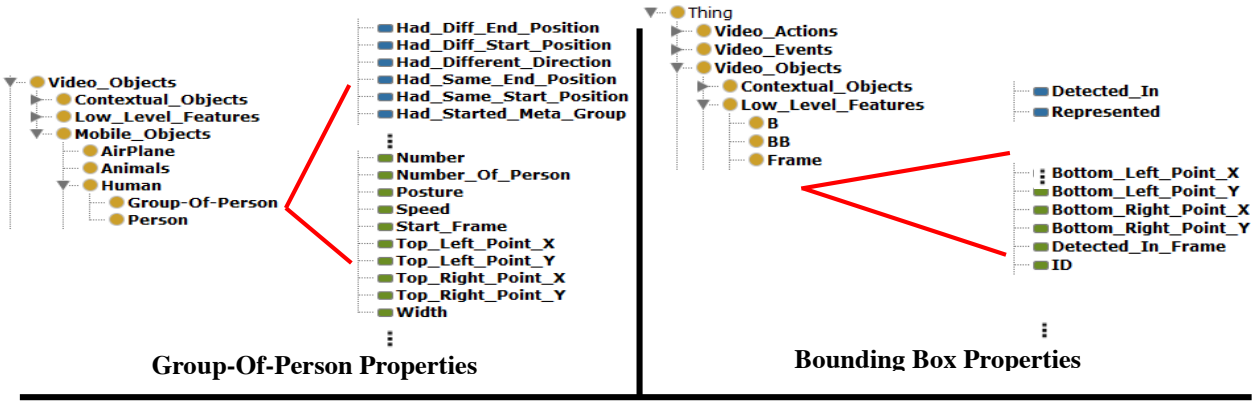

Fig. 3: Group-Of-Person and Bounding Box properties Illustration.

(Object-Property) in the three categories: Object-Against-Object, Human-Against-Object, and Human-Against-Human. The new concepts added to the Video Actions Category are made of one word representing the action, except for the category object, where we need to specify the interaction of our concepts (Human, Ground, Vehicle, Airplane). The nature of interaction with an object in the category Video_Actions must be specified if the action does not consider all types of objects. Moreover, all the concepts of our ontology are created by unifying the words with an underscore (_), except for the composed objects, linked with a hyphen (-).

4 Comparison with other ontologies

Table 6 shows a comparison of the present ontology with the others in literature. Each version has advantages and weaknesses related with the domain representation covering, consistency, etc. As far as our ontology is concerned, it is essentially based on avoiding all these negative points, as explained in Section 3. The aim of this work is to develop a strong and efficient ontology for applications in various domains as discussed in Section 5.

Our ontology uses a naming syntax convention and semantics for consistency, formalism, conceptualization and sufficiently clear relationships between concepts. Moreover, we create a complete video surveillance ontology that includes most of the events arising from both the research and industrial domains. Furthermore, our ontology handles the small domain representation covering problem. Finally, our ontology has been developed for future usages like inference rules (SWRL) which enhance their efficiency with new knowledge in event detection as described in Sections 6 and 7.

Table 6: Comparative study between our ontology and other ones.

Metrics / Points / Ontologies	Consistent	Formal	Large domain representation covering	Conceptual	Separate IS-A from INSTANCE-OF relations	Use inference rules (SWRL)
Our Ontology	OK	OK	OK	OK	OK	OK
Calavia et al. [19]		OK	OK		OK	OK
Sawsan et al. [21]	OK		OK	OK		OK
Trochidis et al. [22]	OK	OK		OK	OK	OK
Bohlken et al. [23]	OK		OK		OK	OK
Bai et al. [26]		OK	OK	OK		OK
SanMiguel et al. [27]	OK	OK		OK	OK	

5 Case studies in various domains

We have proposed a complete and consistent ontology that covers major video surveillance events. This complete knowledge representation could be useful in various domains. In the following sub-sections, we present some interesting application domains that could use the proposed ontology.

5.1 Benchmark creation

A benchmark represents challenges accepted and practiced by the scientific community to solve problems in various domains. In video surveillance, most of the benchmarks like PETS, TRECVID, etc. handle events detection. The formalism of our ontology could help these benchmarks in the selection process of the events appropriately.

5.2 Scene description

Scene description represents all the objects that act/appear in the scene. These objects form either background objects (objects acting for a long time in the scene) and/or foreground objects (new objects appearing in the scene).

Recently, a particular attention is given to the process of describing images automatically. Kuznetsova et al. [36] proposed to consider the task of image description as a retrieval problem, and create a hand-designed approach able to describe images in a wild field. It is based on retrieving similar captioned images from a large database, before generating new description by generalizing and recomposing the retrieved captions. This approach involves typically an intermediate generalization step to remove the specificity of a caption that is relevant only in the retrieved image such as the name of a city. The model reported by Socher et al. [37] uses dependent representations and neural networks to embed images and sentences together, into a common vector form. This approach shows how to map sentence representations from recursive networks into the same space as images. Vinyals et al. [38] demonstrated the effectiveness of storing contextual information in a recurrent layer, and developed a generative approach based on a combination of Convolutional and Recurrent Neural Networks, to generate image captions monitoring their output on the image features extracted by a convolutional neural network. This approach uses the MS COCO dataset that contains 5000 images with 40 reference sentences to enhance the accuracy of automatic measures.

Kiros et al. [39] used two separate pathways (for images and text) to define a joint embedding, even if they can generate text. They proposed a different architecture using the hidden state of an LSTM (Long Short-Term Memory) encoder at time T as the encoded representation of the length T input sequence. It maps this sequence representation and combines it with the visual representation of a modern visual Convnet model, a joint space is obtained with a separate decoder predicts words.

Mao et al. [40] opened new prospects for bi-directional methods that retrieve images based on a textual input, or sentences from a given image. They developed powerful methods of jointly learning from image and text inputs to form higher-level representations from models such as convolutional neural networks (CNNs). They tested their methods on object recognition and word embedding taken from a large-scale text corpus. They proposed a system using Convolutional Neural Networks to extract image features, and Recurrent Neural Networks for sentences, with an interaction performed in a multimodal common layer.

Figure 4 shows a precise segmentation of two scenes extracted from PETS challenge. The scene contains static elements that do not change over time (i.e. buildings, grass, electric poles, roads, trees, car parks, restrictive roads). All these concepts belong to our ontology.

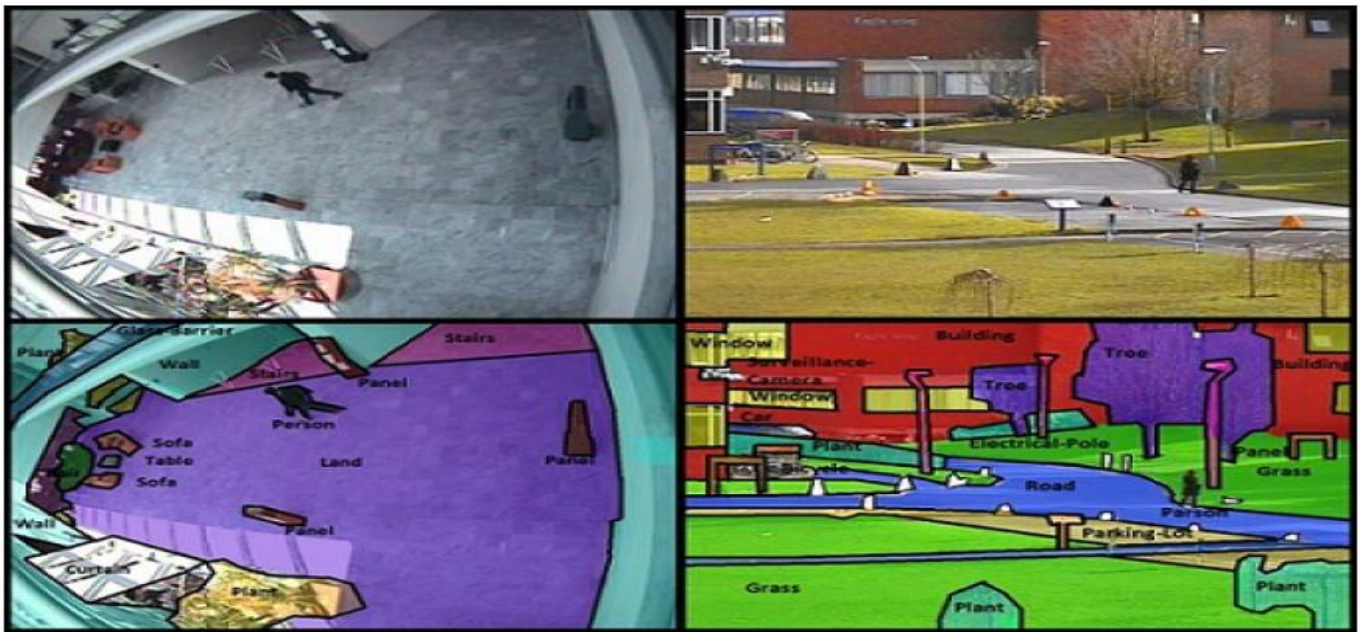

Fig. 4: Scene description from PETS 2004 and PETS 2012 challenges.

5.3 Video description

Like image description, the video description process represents operations that describe objects acting in a video clip. However, unlike the images that are static, the videos require the information dealing with dynamic and temporal changes of the structure along with translation into a natural language description.

Many works have led to greater exploration of video description applications. Yao et al. [41] developed detailed models using Long Short-Term Memory (LSTM) description capable to select the most relevant temporal segments in a video and to incorporate 3D CNN by generating sentences. In their work, two types of encoders are tested: one is a simple frame-wise application of the pre-trained convolutional network, while the other is a 3D convolutional network. Rohrbach et al. [42, 43] used an approach based on statistical machine translation that produces descriptions of videos containing several people cooking in the same kitchen, with the possibility to go from an intermediate semantic representation to sentence generation. Sentences are generated starting from a semantic role representation of high-level concepts such as the actor, action and object. Venugopalan et al. [44] applied the neural approach to static image caption generation, and used an LSTM decoder type for automatic video description generation tasks. They used a convolutional neural network to extract appearance features from each frame of an input video clip. Due to its complete and coherent representation, our proposed ontology could be used easily in all video description works.

5.4 Video Event Indexing

Video indexing offers advanced computer vision capabilities to efficiently and automatically categorize and search events in large datasets. It describes the process of events detection in the video surveillance domain. The consistent and the diversity of our ontology in terms of video surveillance concept representation can incorporate video event indexing and retrieval systems. In Section 7, we present an application of our ontology to video surveillance event indexing, using the PETS 2012 and TRECVID 2016 datasets We have selected five event recognition tasks to depicts the efficiency of the proposed OVIS system.

6 OVIS system architecture

The ontology approach is the effective way to support the event indexing process in the video surveillance domain. It represents the core module in the global architecture of the OVIS system as shown in Figure 5. Its main purpose is to ensure the video sequence indexing process from the first step using the blobs extraction module for extracting blobs bounding box features to the last step of events identification and video sequences indexation.

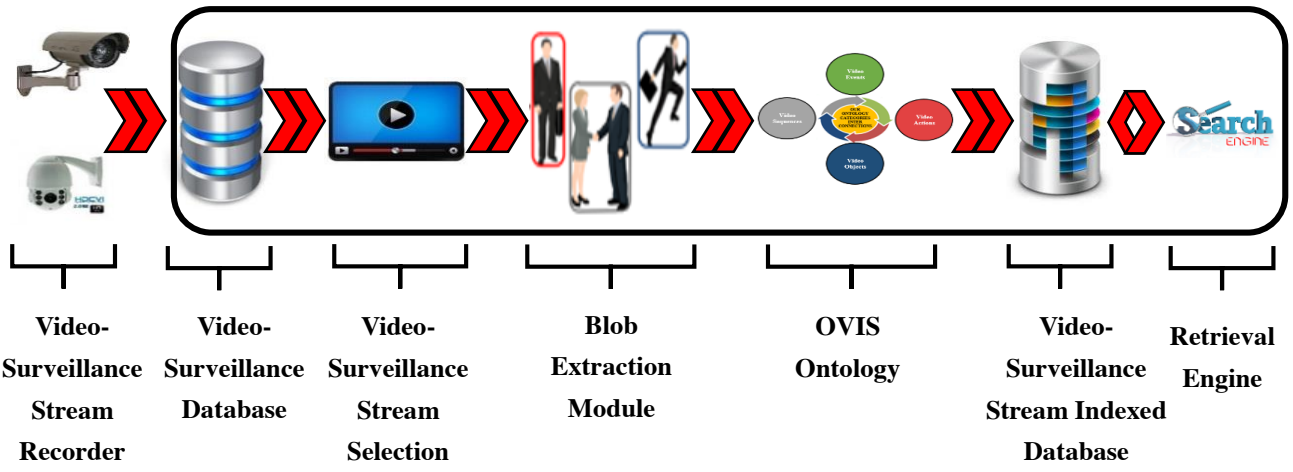

Fig. 5: OVIS system Architecture.

The indexing process of video sequences shown in Figure 5, starts when the video analysis module extracts the different blobs bounding boxes from the video sequence using a background subtraction method with some low-level properties such as Top Left Point X, Top Left Point Y, Width, and Length. The ontology considers these bounding box features as an input, and organize them to create Data Property and Object Property, frame, video sequence, etc. Then, the reasoner of our ontology gives them the correct order using a set of SWRL rules, index this video sequence into the appropriate video event class according to the behavior of its objects with start and end event frames. Finally, the video surveillance stream will be indexed and stored in the database for future needs. For retrieval purpose, the OVIS system allows the search of all indexed videos in the video surveillance stream indexed database using key words expressed as events names (e.g. walking, running, splitting, formation and local dispersion). For example, if we like to retrieve walking events, we will use walking as keyword and the OVIS system will return all sequences indexed with a walking event.

6.1 Blobs extraction module

Blobs represent features that usually have a large coverage area and have proved to be better than points, corners or edges, due to the full occlusion of the subject. Several algorithms could be used to collect the blobs. The background subtraction algorithm will classify the pixels of the input image into foreground and background. The blobs are extracted by collecting the foreground pixels that belong to a single connected component. Optical flow can be used by extracting the characteristics of each pixel in each motion image. These flows are then grouped into blobs with coherent motion and are modelled by a mixture of multivariate Gaussians. The optical flows are useful to characterize each moving pixel according to certain features of the flows' vectors. In the present work, a background subtraction method is used to extract blobs that occur in each frame with their bounding boxes. These features represent the input of our SWRL approach for event detection purposes.

For video analysis, we have used the OpenCV [45] library to extract low-level features of blobs bounding boxes such as top left coordinates, height and width. Therefore, all these features will be used as an input of the OVIS system.

6.2 Methodology for populating the ontology

Our solution is open-source based on the Pellet reasoner and Protege2000 (5.0.0 version) application. However, the biggest challenge was to fill the population of the ontology. We have performed a reverse engineering of Protege2000 to understand how to create the automated filling of individuals with Data_Property and Object_Property features. In each OWL document generated by Protege2000, we found different properties like Individuals, Data-Property, and Object-Propert, etc. Therefore, the solution represents the creation of the parser that reads and extracts different data from the output file of the image/video processing. Then, the parser opens and includes the right tag of the OWL file based on our ontology modelling, the different individuals represented as bounding boxes generated with their Data_Properties and Object_Properties. Finally, the population of the ontology (individuals represented as bounding boxes generated from image/video processing) was already filled with their "Data_Property and Object_Property" opened with the new generated OWL file.

6.3 SWRL rules

To test the efficiency of the proposed approach, different events are addressed together with more than 300 SWRL rules (see web link[1] for some example of SWRL rules), such as:

- Group running and walking events: In each image, the motion magnitude identifies the difference between these two events. For instance, a high-magnitude event means running, while a low-magnitude event means walking. The detection is performed by defining an experimental threshold or using a classifier with a feature such as the average speed of movement. In our case, we used an experimental threshold.
- Group formation and splitting events: The position, orientation and speed of the groups are the main factors determining the accuracy of the events.
- Group Local dispersion events: the positions and the evolving size of the group over frames determine the accuracy of the events.

We used the rule plugin of Protege [46] to write the inference rules of our engine in the SWRL language, and the Pellet reasoner [47] to infer all the events. These SWRL rules are divided into three categories: distance, tracking and event rules.

6.3.1 SWRL Distance rules

They consist of generating all major bounding box in each frame of the video sequence. These rules check distances between detected bounding boxes in the current frame. Neighboring Bounding boxes are grouped into a major one.

Figure 6 depicts an example of a situation for grouping two bounding boxes detected with the blobs extraction module into a major one. An SWRL distance rule verifies whether these bounding boxes could be grouped into the same major bounding box or not. The Pellet Reasoner takes the decision of inferring or not across the right side and checks the left side of the SWRL rule (before the arrow).

6.3.2 SWRL Tracking rules

These rules consist of generating all the different Group_Of_Person instances using the results of major bounding boxes created by the previous rules (SWRL Distance rules) to detect the start/end position of each group and other parameters in the video sequence.

Figure 7 illustrates an example of a situation describing the tracking of the identified group GPY between two successive frames (FZ and FZ+1).

6.3.3 SWRL Event rules

This category of rules is used to detect the appropriate events. Therefore, the behavior of the group identified in the previous category (SWRL Tracking rules) is analyzed. An example of a splitting event is illustrated in Figure 8 below where an event SWRL rule is used for verifying whether the identified group GPZ splits or not into two groups (GPX and GPY) between two successive frames (FZ and FZ+1).

After presenting the three categories of our SWRL rules, the next paragraph demonstrates all the strategies used for reasoning and inferring the different events.

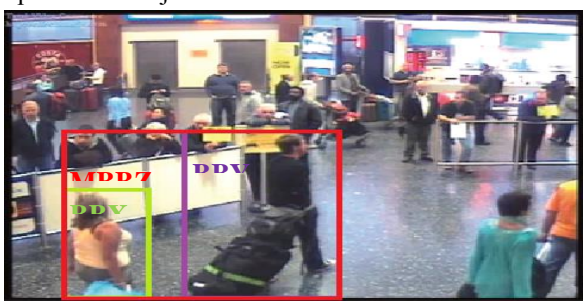

Fig. 6: An illustration of a situation for grouping two bounding boxes into a major one (TRECVID 2016).

Fig. 7: An illustration of checking if the MBB detected in frame FZ+1 represents the same GPY in frame FZ (PETS 2012).

[1] http://ovis-system-information.000webhostapp.com/

Fig. 8: An illustration of a situation for checking if the group GPZ is split or not (PETS 2012).

6.3.4 Reasoning strategies

In the aim of inferring the different events presented above, we create different SWRL rules with a strategy that follows all these 16 steps in the reasoning process:

1) Inferring the four coin points of each bounding box (x, y) detected with the low-level extractor module.
2) Inferring the major bounding box in the case of frame containing only one bounding box.
3) Inferring the four coin points of each major bounding box (x, y) (case of frame containing only one bounding box).
4) Inferring an ID for each bounding box detected with respect to its position in the frame (the one who is most to the right, will have the ID number one, then the one who comes to his left will have the ID number two and so on).
5) Inferring the majors bounding boxes in the case of frame containing two bounding boxes.
6) Inferring the four coin points of each major bounding box (x, y) (case of frame containing two bounding boxes).
7) Inferring the majors bounding box in the case of frame containing three bounding boxes or more:
   7.1) Comparison between Bounding Boxes having as ID number one and two to extract all the blocks with their MIN and MAX (x, y).
   7.2) Comparison of the generated blocks with different Bounding Boxes of frame and inferring an FID (Final ID) for each Bounding Box.
   7.3) Comparison between Bounding Boxes with final ID number one and final ID number two and extract either an MBB (Major Bounding Box) in the case of a large distance or a TBB (Temporary Bounding Box) in the case of a small distance.
   7.4) Comparison between a TBB and the rest of Bounding Boxes with respect to the order of the final ID and extract either an MBB (Major Bounding Box) or a new TBB (Temporary Bounding Box) with the same strategy of distance noted in the last point.
   7.5) Inferring the four coin points of each major bounding box (x, y) generated above.
8) Measuring the centroid of each Major bounding box.
9) Search the first frame that contains a Major Bounding Box and inferring a TGP (Temporary Group of Person).
10) Inferring groups with their properties: Started Frame, Ended frame, in the normal case.
11) Inferring Groups with their properties: Started Frame, Ended frame, in the case of final frame.
12) Inferring Groups with their properties: Started Frame, Ended frame, in the case of empty frame.
13) Search the first frame that contains a Major Bounding Box after browsing an empty frame.
14) Inferring relations between detected groups.
15) Checking if it is not a false detection.
16) Inferring the type of event.

7 Results and discussions

To test the efficiency of the OVIS system inspired by TRECVID and PETS, we selected five events (walking, running, split, formation and local dispersion). We developed an application in the Java environment that handles all the steps of our indexing and retrieval system. The process started with the selected video and ended with the indexing results. The tests were performed on a machine with Intel Core I7 CPU and 16 GB RAM, under Windows 8. We considered three types of evaluations to check the performance of the OVIS system:

7.1 Evaluation based on the events

The first type of evaluation was based on the number of events returned by the OVIS system; it is carried out by many metrics, such as Precision, Recall, F-measure, FP (False Positive), FN (False Negative), TP (True Positive) and TN (True Negative). We considered these measures as follows:

- Precision = Number of detected videos that contain the event / Number of videos indexed with the event.
- Recall = Number of detected videos that contain the event / Number of all videos in Database that contain the event.
- F measure = 2 * ((Precision * Recall) / (Precision + Recall)).
- TP = Number of videos indexed with the event and containing it.
- TN = Number of videos not indexed with the event and not containing it.
- FP = Number of videos indexed with the event and not containing it.
- FN = Number of videos not indexed with the event and containing it.

We selected two challenges (PETS 2012 and TRECVID 2016) to evaluate our event detection system.

In the first challenge, we used 16 videos (with 4240 frames) that were used to describe different events and we carry out all the metrics noted above.

In the second one, an 11-hours subset of the multi-camera airport surveillance domain evaluation data is used with the following evaluation metrics: select Precision, Recall and F-measure.

7.1.1 PETS 2012 Challenge

Table 7 shows the indexing results for each event. We consider 13 walking, 8 running, 4 splitting, 4 formation and 4 local dispersion events as the ground truth videos, given by the first column. The second column displays the number of videos that the OVIS system indexes in each event; the final column shows the number of videos containing the effective events among those indexed by OVIS.

Table 7: OVIS indexing results of the five different events.

Results / Events	Number-of-all-video-in-Database-that-really-contain-the-event (ground truth)	Number-of-video-returned	Number-of-video-returned-that-really-contain-the-event
Walking	13	11	11
Running	08	09	05
Split	04	06	03
Formation	04	06	03
Local dispersion	04	04	04

Discussion 1

In Table 8, we summarized the statistics of the obtained data from the full dataset. On one hand, the events walking and local dispersion provide excellent results and reach 100 % of precision. Therefore, these results mean that the number of detected videos by OVIS system that contains walking and local dispersion events is equal to the number of videos indexed with these events. On the other hand, the events of running, splitting, formation exceed 50% of precision. So, we can conclude that the number of detected videos by OVIS system that contains running, splitting, and formation events is equal to at least the half of the number of videos indexed with these events. As illustrated, the event local dispersion reaches also an excellent result of 100% in recall. This means that the OVIS system does not miss any local dispersion event. Moreover, the recall of event walking, running, splitting and formation provides good result and is at least above 62%. Following the result of precision and recall, the F-measure metric expresses the relation precision/recall. Consequently, the F-measure metric provides excellent result in walking and local dispersion events and a good one in the rest of events. The metrics (FP, FN, TP and TN) give the accuracy of the indexing process generated by of the OVIS system. Thus, this accuracy is considered as excellent when the number of FP and FN is very low and the number of TP and TN is very high. As described in Table 8, these metrics provide good results in general. For example, walking event generates good results with 3 videos not indexed with the event and not detected among 3 videos (TN), 11 videos indexed with the event and containing it among 13 videos (TP), 2 videos not indexed with the event and containing it (FN) and no video not indexed with the event and containing it (FP). This is an evidence that the OVIS system using SWRL rules with their different categories (Distance rules, Tracking rules, Event rules) successively, opens new prospects without using traditionally methods (SVM, KNN, etc.). The use of this complete video surveillance ontology is efficient in this indexing process that represents one among other domain applications addressed by this ontology.

Table 8: The obtained results for the five different events.

Measures / Events	Precision	Recall	F-measure	FP	FN	TP	TN
Walking	100%	84%	91%	00	02	11	03
Running	55%	62%	58%	04	03	05	04
Splitting	50%	75%	60%	03	01	03	09
Formation	50%	75%	60%	03	01	03	09
Local dispersion	100%	100%	100%	00	00	04	12

Comparison with other approaches

For the aim of evaluating the OVIS system based on SWRL rules, we compared it with two other studies [28, 29] that use the same video sequences from the PETS 2012 challenge dataset and handle the same event detection purpose: Group Walking, Group Running, Group Splitting, Group Formation and Group Local Dispersion.

Table 9 compares the results obtained with the OVIS system and those reported in [28, 29], using the Dataset PETS 2012. Figure 9 illustrates the Precision/Recall of our approach compared with the others.

Table 9: Comparison of different detection event approaches, NC (Not Communicated), ND (Event Not Detected).

Event	Metrics	The OVIS System	Utasi et al. [28]	Chan et al. [29]
Group Walking	Precision	1	ND	0,87
	Recall	0,84	ND	NC
Group Running	Precision	0,55	0,99	0,75
	Recall	0,62	0,99	NC
Group Splitting	Precision	0,5	0,6	0,74
	Recall	0,75	1	NC
Group Formation	Precision	0,5	ND	0,6
	Recall	0,75	ND	NC
Group Local Dispersion	Precision	1	ND	0,8
	Recall	1	ND	NC

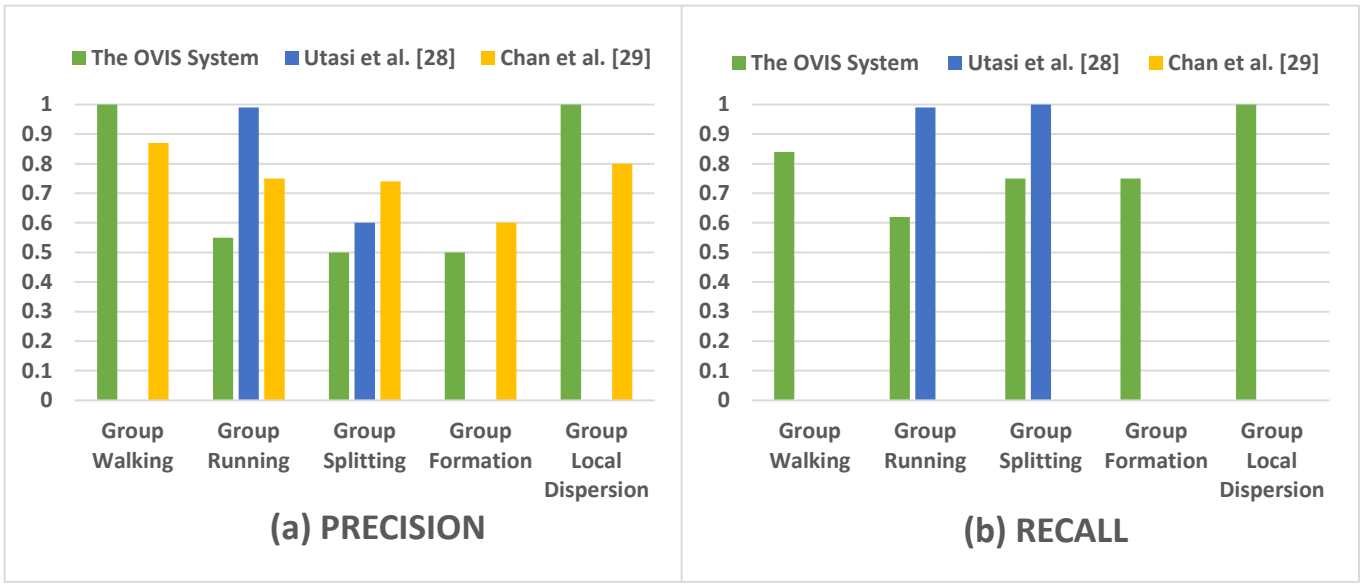

Fig. 9: Comparison of different detection event approaches under graphical forms, (a) Precision, (b) Recall.

Our approach based on SWRL rules detects all events, while Utasi et al. [28] method detects only two events (Group Running and Group Splitting). They did not try to identify the missing events although they claim that their approach was able to detect the other events without providing any indications. The method of Chan et al. [29] cannot include two events simultaneously while OVIS offers the possibility of processing and detecting two or more events at the same time, as illustrated in Figure 11 below. Furthermore, the approach of Chan et al. [29] does not provide some important results as indicated by NC in Table 9. Even if the precision of [29] is better than OVIS in few cases, they did not present their recall.

These observations lead us to conclude that the negative points noted above (not detecting all events, not processing two events simultaneously and not providing recall results) prove that our approach based on SWRL rules and created from the presented ontology is strong. We could also add new events just by creating the associated SWRL rules.

7.1.2 TRECVID 2016 SED task

For the aim of evaluating our OVIS system in various contexts, we perform a comparison with two other approaches [30, 31] that participate in the SED task of TRECVID 2016 challenge. The SED task regroups seven types of events:

PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp, Embrace, and Pointing. In our case studies, our OVIS system with it different SWRL rules handles only three of them: PersonRuns (Running), PeopleMeet (Formation), PeopleSplitUp (Splitting). In this way, we compare results obtained from our OVIS system with those presented in [30, 31] representing these three events.

Table 10: OVIS indexing results of the three different events.

Results / Events	Number-of- event (ground truth)	Number-of-all-event-returned-by-OVIS	Number-of-correct-event-returned-by-OVIS
People Meet	323	411	303
People Split Up	176	203	173
Person Runs	63	97	58

Table 10 shows the indexing results of PeopleMeet, PeopleSplitUp and PersonRuns events returned by our OVIS system. In our case, the OVIS system based on SWRL rules inferring method return the events of Formation, Splitting and Running and we consider them as PeopleMeet, PeopleSplitUp and PersonRuns respectively. We consider 323 PeopleMeet, 176 PeopleSplitUp and 63 PersonRuns events as the ground-truth videos, given by the first column. Each event was detected with its start and end frames that provide the correct detection or false alarm event.

Table 11 illustrates the results obtained with the OVIS system and those reported in [30, 31], using the Dataset TRECVID 2016 SED task, where Figure 10 shows the Precision/Recall metrics of our approach compared with the others.

Table 11: Comparison of different detection event approaches.

Event	Metrics	The OVIS System	Markatopoulou et al. [30]	Zhao et al. [31]
People Meet	Precision	0,74	0,02	0,34
	Recall	0,94	0,92	0,18
	F-measure	0,83	0,04	0,23
People Split Up	Precision	0,86	0,01	0,32
	Recall	0,98	0,98	0,2
	F-measure	0,92	0,02	0,25
Person Runs	Precision	0,6	0,01	0,67
	Recall	0,92	0,97	0,35
	F-measure	0,73	0,02	0,46

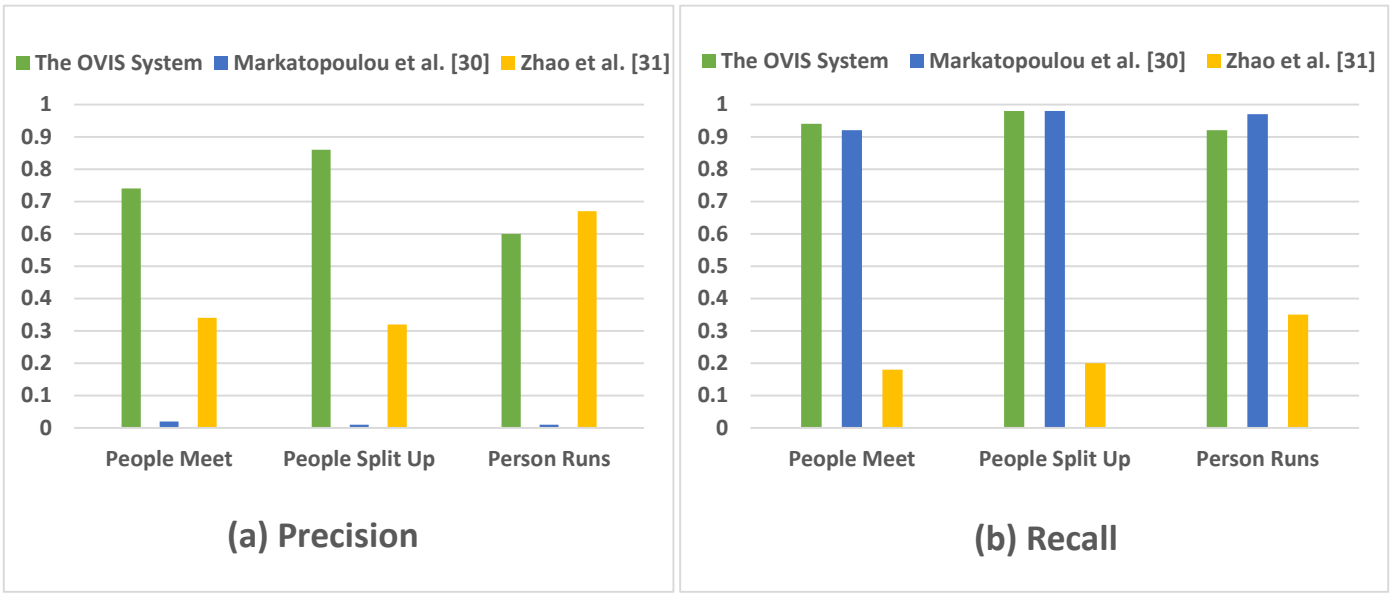

Fig. 10: Comparison of different detection event approaches under graphical forms, (a) Precision, (b) Recall.

First, our approach based on SWRL rules has a better ratio Precision/Recall (F-measure) compared with those exposed in [30, 31]. This point means that our system detects the majority of events related to ground truth without detecting a great number of false alarms. Secondly, the approach Markatopoulou et al. [30] detects the majority of correct event like our OVIS system but at the same time a great number of false alarms. We can conclude that this approach does not miss a large number of correct events but their precision is very low compared with OVIS system. Third, the method of Zhao et al. [31] expresses acceptable results in precision metrics but at the same time misses a great number of correct detection events compared with OVIS system. These observations lead us to conclude that all weaknesses detected in [30, 31] (miss a great number of events, detect a great number of false alarms) prove that our approach based on SWRL rules and created from this ontology is strong. We could also add new events like CellToEar, ObjectPut Embrace, and Pointing or other ones by creating the associated SWRL rules.

7.2 Evaluation based on the frame timing (PETS 2012)

The second type of evaluation is based on frame timing using an overlay (in frame number) of 10 frames to evaluate the performance of our system. To meet this end, three kinds of situations are considered:

- Too Early: Our system detects events before they really start (ground truth) with 10 frames.
- On Time: Our system detects events exactly when they start.
- Too Late: Our system detects the events after they really start with 10 frames.

Figure 11 represents the frame timing events of sixteen videos (Ground Truth and results returned by the OVIS system). The annotation for ground truth is done manually by watching the entire video, and represents in each case the events that occur with their start/end frames.

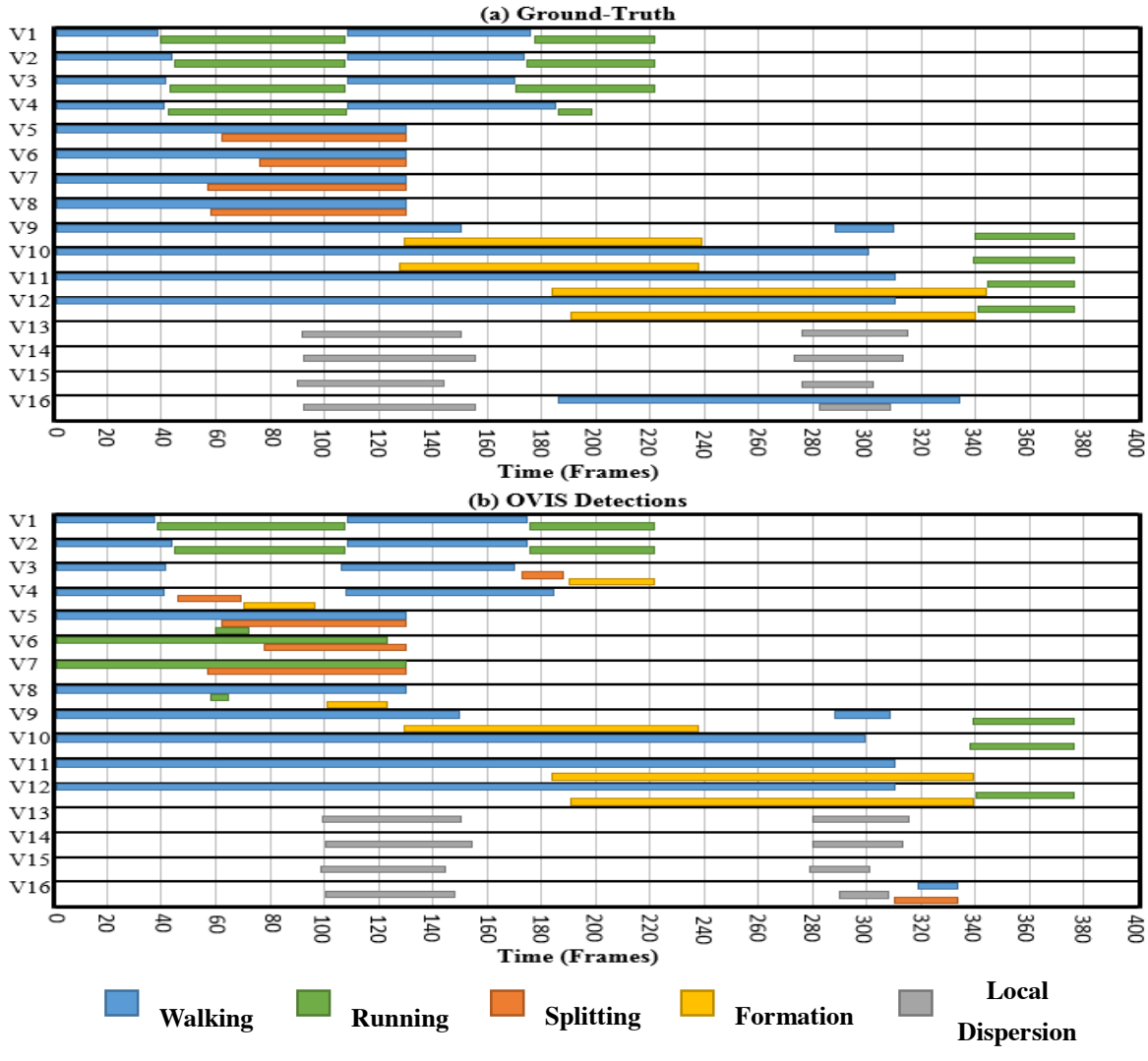

Fig. 11: Frame timing events of sixteen videos (Ground-Truth and results returned by the OVIS system).

Discussion 2

Figure 11 above illustrates the results for the frame timing events obtained from the output of the OVIS system compared with the ground truth. The first advantage is the detection of at least one correct event in each video sequence analysed by our system. Moreover, the second positive point is that all the correct events are detected on time without exceeding the overlay of 10 frames. These positive points demonstrate that our approach (based on the three types of SWRL rules from blobs features until event detection) works correctly at least once, and detects the right event at the right moment, where [28, 29, 30, 31] fail in this task. However, the weakness of the OVIS system is the occasional events confusions that generate incorrect events detections. This is generally due to these two examples cases:

(1) The emergence of objects (not relevant) detected in the video sequence by the low-level feature analysis module; their behaviour leads our system to detect a wrong event. For example, when an image/ video processing detects a moving tree shadow and considers it as bounding box of pertinent blobs. This wrong detection leads OVIS system to detect a wrong event in general.

(2) The incorrect splitting and formation of detected objects caused only by their walking or running speed. For example, when in the same Group-Of-Person, we detect different human speeds and in presence of walking or running events, OVIS can detect a wrong Splitting or Formation event.

We are confident that these weaknesses can be solved in a future work by adding new SWRL rules and improving initial blob detection and characterisation. We not only expect to improve the obtained results, but also predict and extend the tests for detecting new events.

Table 12: Illustration of the inferring results representing Allen's relations provided by the OVIS system

Video_Sequences	Allen's Relations detected by OVIS
Video 1	Took_Place_Before_R / Took_Place_After_R / Met_R / Met_By_R
Video 2	Took_Place_Before_R / Took_Place_After_R / Met_R / Met_By_R
Video 3	Took_Place_Before_R / Took_Place_After_R / Met_R / Met_By_R
Video 4	Took_Place_Before_R / Took_Place_After_R / Met_R / Met_By_R
Video 5	Finished_R / Finished_By_R / Overlapped_R / Overlapped_By_R / Took_Place_During_R / Contained_R
Video 6	Overlapped_R / Overlapped_By_R
Video 7	Finished_R / Finished_By_R
Video 8	Took_Place_During_R / Contained_R / Took_Place_Before_R / Took_Place_After_R
Video 9	Overlapped_R / Overlapped_By_R / Took_Place_Before_R / Took_Place_After_R
Video 10	Took_Place_Before_R / Took_Place_After_R
Video 11	Overlapped_R / Overlapped_By_R
Video 12	Overlapped_R / Overlapped_By_R / Met_R / Met_By_R
Videos 13 /14/ 15	
Video 16	Took_Place_Before_R / Took_Place_After_R / Met_R / Met_By_R

7.3 Evaluation based on Allen's interval algebra (PETS 2012)

The third type of evaluation is based on Allen's interval algebra [48] for modelling many possible situations such as "X takes place before Y", "X overlaps Y", "X meets Y", …. (Where X and Y represents two different events detected in the same video sequences). For this purpose, another kind of SWRL rules were developed, and aim to infer the different relations between events detected in the same video sequence. The results of these SWRL rule were saved as a Boolean Data_Property related to Video_Sequences individuals (For example of relation: Took_Place_Before_R). Therefore, all the relations expressed as Data_Property inferred with these SWRL rules take true as value. For inferring the different relations between events detected in the same Video_Sequences, the start frame and end frame of each event represents the key aspects. Furthermore, the detection of five frames between two different successive events was considered in our case as "Met_R / Met_By_R" relations, and therefore up to five frames as "Took_Place_Before_R / Took_Place_After_R" relations. However, when the Video_Sequences contain only one event detected in several times, like Videos (13, 14, and 15) in Table 12, no relations was inferred.

Therefore, we expect seven possible relations with their opposite situations:

- X took place before Y: Where event X finished before when event Y started.
- Y took place after X: Where event Y started after when event X finished.
- X met Y: Where event X finished at the same time when event Y started.
- Y met by X: Where event Y started at the same time when event X finished.
- X overlapped Y: Where event X started before event Y and event Y finished after event X.
- Y overlapped by X: Where event Y started after event X and event X finished before event Y.
- X started Y: Where event X started at the same time with event Y and finished before event Y.
- Y started by X: Where event Y started at the same time with event X and finished after event X.
- X took place during Y: Where event X started after event Y and finished before event Y.
- Y contained X: Where event Y started before event X and finished after event X.
- X finished Y: Where event X started after event Y and finished at the same time with event Y.
- Y finished by X: Where event Y started before event X and finished at the same time with event X.
- X equalled to Y: Where event X started and finished at the same time with event Y.

Discussion 3

Table 12 above illustrates all the Allen's relations obtained from the output of the OVIS system. The first advantage is the detection of all the correct relations in the sixteen video. For example, in Video 12 "V12" illustrated in Figure 11 above (OVIS Detections part) handles three kinds of events (Walking, Formation, and Running). The different relations that we can observe is that:

- First, event Walking started before event Formation and event Formation finished after event Walking (X overlapped Y) and (Y overlapped by X).
- Second, event Formation finished at the same time when event Running started (X met Y) and (Y met by X).

However, all these Allen's relations are inferred by the OVIS system like described in Table 12 (Video 12). The second advantage is no detection of any relation in presence of only one event detected in a video sequence, is done. For example, video 13, video 14 and video 15 illustrated in Figure 11 above, we can find only Local-Dispersion event. Consequently, any relation was inferred by OVIS system as presented in Table 12. These positive points demonstrate that our approach based of SWRL is able to handle correctly Allen's relations whereas this point is not generated or presented also in [28, 29, 30, 31].

8 Conclusions

Nowadays, Video Surveillance systems are part of our daily life, because of their role to ensure security and safety (i.e. allowing human behaviour to be studied among the population). Thus, many research works have tried to develop an efficient system to index a very large volume of data accurately. In the present work, we have proposed a complete and coherent Video Surveillance Ontology and a rule-based approach to detect multiple object events or crowd events (e.g. Group walking, Group splitting, etc.). In fact, we have described the link between the four main categories composing our ontology (Video_Sequences, Video_Objects, Video_Events, Video_Actions), that are in interaction.

Our Video Surveillance ontology covers a very large number of objects and events happening in the video surveillance domain, as well as exhibiting a large dimension taking into consideration new concepts that represent events in the industrial domain.

Furthermore, we implemented the OVIS indexing and retrieval system, based on a complete Video Surveillance Ontology and SWRL rules in the middle and high level of indexing process. Moreover, we tested OVIS with videos selected from the PETS 2012 and TRECVID 2016 Challenges. In this way, we obtained very promising results.

Moreover, the strengths of our approach are as follows:

- The competitive level results obtained with the different types of evaluations such as: evaluation based on the events, evaluation based on the frame timing, evaluation based on Allen's interval algebra.
- The facility of creating and using SWRL rules or adding new ones when new events occurs. This point allows the research community to address many future prospects in the domain ontology-based video surveillance indexing and retrieval systems.

The weaknesses of our approach stem from the requirement of manual reproduction of SWRL rules when new events occur and the lack of the uncertainty management by the OVIS system in image/video processing.

In our future work, we will extend the OVIS system by considering other events that could occur in the video surveillance domain. This will be possible by adding new SWRL rules and testing them using other datasets. In addition, we plan to use the neuronal network formalism to reproduce others SWRL rules, and use the Shannon's normalized entropy function for modelling the uncertainty associated to image/video processing.

References

1. D. Kless, L. Jansen, J. Lindenthal and J. Wiebensohn, (2012) "A method for reengineering a thesaurus into an ontology," Frontiers in Artificial Intelligence and Applications (FAIA), pp. 133-146.

2. A. Badii, C. Lallah, M. Zhu and M. Crouch, (2009) "The dream framework: Using a network of scalable ontologies for intelligent indexing and retrieval of visual content," in International Conference

on Web Intelligence and Intelligent agent Technology (WI-IAT), pp. 551-554.

3. M. Rodrguez-Muro and D. Calvanese, (2012) "High performance query answering over dl-lite ontologies," in International Conference on Principles of Knowledge Representation and Reasoning (KR), pp. 308-318.

4. A. Scherp, C. Saathoff, T. Franz and S. Staab, (2011) "Designing core ontologies," Journal Applied ontology, vol. 03, pp. 177-221.

5. R. Benmokhtar and B. Huet, (2011) "An ontology-based evidential framework for video indexing using high-level multimodal fusion," Multimedia Tools and Applications (MTAP), vol. 55, no. 3, pp. 1–27.

6. A. Rector, S. Brandt, N. Drummond, M. Horridge, C. Pulestin and R. Stevens, (2012) "Engineering use cases for modular development of ontologies in owl," Journal Applied ontology, vol. 02, pp. 113-132.

7. B. Smith and W. Ceusters, (2010) "Ontological realism as a methodology for coordinated evolution of scientific ontologies," Journal Applied ontology, vol. 03, no. 4, pp. 139-188.

8. P. Hernandez-Leal, H. J. Escalante and L. E. Sucar, (2017) "Towards a Generic Ontology for Video Surveillance," Applications for Future Internet.

9. S. Kara, Z. Alan, O. Sabuncu, S. Akpnar, N. K. Cicekli and F. N. Alpaslan, (2012) "An ontology-based retrieval system using semantic indexing," Information Systems Journal, vol. 04, pp. 294-305.

10. T. Mossakowski, C. Lange and O. Kutz, (2013) "Three semantics for the core of the distributed ontology language," in International Joint Conferences on Artificial Intelligence (IJCAI), pp. 3027-3031.

11. L. Ballan, M. Bertini, A. Del Bimbo and G. Serra, (2010) "Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies," Multimedia Tools and Applications (MTAP), vol. 02, pp. 313-337.

12. A. D. Bagdanov, M. Bertini, A. Del Bimbo, G. Serra and C. Torniai, (2007) "Semantic annotation and retrieval of video events using multimedia ontologies," in International Conference on Semantic Computing (ICSC), pp. 713-720.

13. M. Bertini, A. Del Bimbo, C. Torniai, C. Grana and R. Cucchiara, (2007) "Dynamic pictorial ontologies for video digital libraries annotation," in 1st ACM Workshop on The Many Faces of Multimedia Semantics, pp. 47-56.

14. M. Bertini, A. Del Bimbo and G Serra, (2008) "Learning ontology rules for semantic video annotation," in 2nd ACM Workshop on Multimedia Semantics, pp. 1-8.

15. M. OConnor, H. Knublauch, S. Tu, B. Grosof, M. Dean, W. Grosso and M. Musen, (2005) "Supporting rule system interoperability on the semantic web with swrl," in 4th International Semantic Web Conference (ISWC), pp. 974-986.

16. M. Xue, S. Zheng and C. Zhang, (2012) "Ontology-based surveillance video archive and retrieval system," in 5th International Conference on Advanced Computational Intelligence (ICACI), pp. 84-89.

17. J. Lee, M. H. Abualkibash and P. K. Ramalingam, (2008) "Ontology based shot indexing for video surveillance system," in Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering, pp. 237-242.

18. L. Snidaro, M. Belluz and G. L. Foresti, (2007) "Representing and recognizing complex events in surveillance applications," in IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 493-498.

19. L. Calavia, C. Baladrn, J. M. Aguiar, B. Carro and A. Sanchez-Esguevillas, (2012) "A semantic autonomous video surveillance system for dense camera networks in smart cities," sensors, pp. 10407-10429.

20. G. T. Papadopoulos, V. Mezaris, I. Kompatsiaris and M. G. Strintzis, (2007) "Ontology-driven semantic video analysis using visual information objects," in International Conference on Semantic and Digital Media Technologies, pp. 56-69.

21. S. Saad, D. D. Beul, M. Said and M. Pierre, (2012) "An ontology for video human movement representation based on benesh notation," in IEEE International Conference on Multimedia Computing and Systems (ICMCS), pp. 77-82.

22. I. Trochidis, E. Tambouris and K. Tarabanis, (2007) "An ontology for modeling life-events," in IEEE International Conference on Services Computing (SCC), pp. 19-20.

23. W. Bohlken and B. Neumann, (2009) "Generation of rules from ontologies for high-level scene interpretation," Lecture Notes in Computer Science, pp. 93-107.

24. R. Nevatia, J. Hobbs and B. Bolles, (2004) "An ontology for video event representation," in Computer Vision and Pattern Recognition (CVPR), pp. 119-128.

25. A. R. J. Francois, R. Nevatia, J. Hobbs, R. C. Bolles and J. R. Smith, (2005) "VERL: an ontology framework for representing and annotating video events," IEEE Multimedia, vol.12, pp. 76-86.

26. L. Bai, S. Lao, W. Zhang, G. J. F. Jones and A. F. Smeaton, (2008) "Video semantic content analysis framework based on ontology combined mpeg-7," Lecture Notes in Computer Science, pp. 237-250.

27. J. C. SanMiguel, J. M. Martinez and A. Garcia, (2009) "An ontology for event detection and its application in surveillance video," in IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 220-225.

28. A. Utasi, A. Kiss and T. Sziranyi, (2009) "Statistical filters for crowd image analysis," in Performance Evaluation of Tracking and Surveillance workshop,at CVPR, pp. 95–100.

29. A. B. Chan, M. Morrowand and N. Vasconcelos, (2009) "Analysis of crowded scenes using holistic properties," in 11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS).

30. Z. Zhao, M. Wang, R. Xiang, S. Zhao, K. Zhou, M. liu, S. He, Y. Zhu, Y. Zhao and F. Su, (2016) "BUPT-MCPRL," at TRECVID.

31. F. Markatopoulou, A. Moumtzidou, D. Galanopoulos, T. Mironidis, V. Kaltsa, A. Ioannidou, S. Symeonidis, K. Avgerinakis, S. Andreadis, I. Gialampoukidis, S. Vrochidis, A. Briassouli, V. Mezaris,

I. Kompatsiaris and I. Patras, (2016) “ITI-CERTH,” at TRECVID.
32. M. Y. Kazi Tani, A. Ghomari, H. Belhadef, A. Lablack and I. M. Bilasco, (2014) “An ontology based approach for inferring multiple object events in surveillance domain,” in IEEE Science and Information Conference (SAI), pp. 404-409.
33. M. Y. Kazi Tani, A. Ghomari, A. Lablack and I. M. Bilasco, (2015) “Events detection using a video-surveillance ontology and a rule-based approach,” in Computer vision + ONTology Applied Cross-disciplinary Technologies workshop (CONTACT) in conjunction with European Conference in Computer Vision (ECCV), pp. 299-308.
34. PETS. Pets 2012 challenge [online]: http://www.cvg.reading.ac.uk/PETS2012/a.html
35. TRECVID. TRECVID 2016 challenge [online]: http://www-nlpir.nist.gov/projects/tv2016/tv2016.html
36. P. Kuznetsova, V. Ordonez, T. Berg and Y. Choi, (2014) “Treetalk: Composition and compression of trees for image descriptions,” Transactions of the Association for Computational Linguistics (TACL), pp. 351-362.
37. R. Socher, A. Karpathy, V. Q. Le, C. D. Manning and A. Y. Ng, (2014) “Grounded compositional semantics for finding and describing images with sentences,” Transactions of the Association for Computational Linguistics (TACL), pp. 207-218.
38. O. Vinyals, A. Toshev, S. Bengio and D. Erhan, (2014) “Show and tell: A neural image caption generator,” arXiv:1411.4555.
39. R. Kiros, R. Salakhutdinov and R. S. Zemel, (2014) “Unifying visual-semantic embeddings with multimodal neural language models,” arXiv:1411.2539.
40. J. Mao, W. Xu, Y. Yang, J. Wang and A. L. Yuille, (2014) “Explain images with multimodal recurrent neural networks,” arXiv:1410.1090.
41. L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle and A. Courville, (2015) “Describing videos by exploiting temporal structure,” in IEEE International Conference on Computer Vision (ICCV).
42. A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal and B. Schiele, (2014) “Coherent multi-sentence video description with variable level of detail,” in German Conference on Pattern Recognition (GCPR).
43. M. Rohrbach, W. Qiu, I. Titov, T. Stefan, M. Pinkal and B. Schiele, (2013) “Translating video content to natural language descriptions,” in IEEE International Conference on Computer Vision (ICCV).
44. S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney and K. Saenko, (2014) “Translating videos to natural language using deep recurrent neural networks,” arXiv:1412.4729.
45. OpenCV. The OpenCV API [online]: http://docs.opencv.org/3.3.0/
46. Protege. The protege project [online]: http://protege.stanford.edu
47. E. B. Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur and Y. Katz, (2003) “Pellet: A Practical OWL-DL Reasoner,” Journal of Web Semantics.
48. J. F. Allen, (1983) “Maintaining knowledge about temporal intervals,” Communications of the ACM, 26, pp. 832–843.