Generic Video Surveillance Description Ontology

Wael F. Youssef
IRIT, Paul Sabatier University, CNRS, SAMOVA Team
Toulouse, France
wael.youssef@irit.fr

Siba Haidar
Lebanese University
Faculty of Sciences
Beirut, Lebanon
siba.haidar@ul.edu.lb

Philippe Joly
IRIT, Paul Sabatier University, CNRS, SAMOVA Team
Toulouse, France
joly@irit.fr

***Abstract*— In this study, we present an automatic generic ontology for video surveillance description, to be used as a high-level layer in video-surveillance systems. We considered the temporal dimension of the video, using appropriate features and classification methods. Our ontology introduces six main classes; one of which is a representation for generic scene types, divided into twelve subtypes according to the number of moving objects before and after the interaction. This ontology was used to fulfill an automatic textual description of video surveillance, focusing mainly on interactions between objects.**

***Keywords*— Ontology, Video surveillance, Automatic video description, Video objects interaction, Scene type.**

I. Introduction

We live in the age of big data; one big data source is the multimedia content, which includes the video surveillance domain. The progress in hardware and telecommunication technologies and capability of storage has resulted in a rapid increase of the available amount of huge video surveillance streams. For that, the deployment of video surveillance systems worldwide has grown exponentially in recent years, especially, when video surveillance is considered to be an ubiquitous feature for the security and safety in the modern fight against crimes. Visual surveillance is now used to monitor the security of sensitive areas. It is a tool for crime reduction and risk management. Currently, hundreds of large cities have video cameras already installed in the streets and around the important cites; public places, schools, banks, highways, department stores, shopping malls, transport infrastructures, railway stations, hospitals, government buildings, commercial premises and borders. Videos are a rich and complex source of information. Cameras surveillance systems produce large amounts of video data which are used live or stored for future use. The main concern is how to extract, automatically, the useful information out of video surveillance data.

Years of video surveillance are recorded. If those records can be automatically processed, analysed and classified, this would give us an access to a huge amount of useful data. How people act and interact with the world around them and with each other? Video surveillance provides real-time data about behaviour happening in the present, not just the past — whether it is traffic, public places, or indoors incidents and others.

To fulfil such a need, video content analysis paradigm is shifting from a fully human operator model to a machine-assisted and automated model. The computer vision and artificial intelligence community are seeking to develop automated systems for real-time monitoring and archives investigation of contents understanding. Detecting an incident is an easy task for human, but it is very hard for machine. Therefore, the real need is to extract meaningful information efficiently from the huge flow or storage of video data in order to produce high-level scheme or descriptions of the activities occurring in the area under surveillance, and as mentioned, especially the interaction between those objects.

Additionally, as video surveillance recordings are most of the time only indexed with rough descriptors like time, camera ID, and some photometric parameters, there is an urgent need to develop intelligent methods for effective storing, indexing, organizing, data mining and retrieval.

Although there are many ways to represent the content of video in current video research, there still exists a big gap between users and systems. It is worth to mention that significant results have been reported in the literature on many fields concerning video analysis and understanding in general and for automated video surveillance in particular. To simplify the problem adding some assumptions may significantly improve the results but will limit the applicability in real world. Most of the researches often have specific limitation, they are designed for particular set of objects, and actions in a specific context, and no generic multimodal framework to achieve system robustness in multiple contexts, object types and actions performed. However, the lack of precise and generic models for video content representation and the complexity of video processing algorithms make the development of fully automatic video content analysis a challenging task. This challenge, which often referred as the semantic gap, is corresponding low-level spatio-temporal features that can be automatically extracted from video data with high-level semantic concepts. This causes the existing systems and approach to be too non-flexible and can not satisfy the need of video applications at the semantic level. So the use of domain knowledge is very necessary to enable higher level semantics in automatic parsing. This is where “Ontology” enters the scene. Ontologies are a powerful mechanism for structuring, organizing and reusing knowledge; also it is a way to reduce the semantic gap in video processing between low level descriptors and the domain of interest.

In this study, we present our ontology named "Video-Surveillance-Description Ontology". It is an automatic generic approach for video surveillance description, and designed to be used as a generalist high-level layer for video analysis, principally in a video-surveillance system.

Next, we present the state of art of many ontologies in the domain of video surveillance. Then, we present our own ontology.

II. Related Works

Different domains exist for the development of video analysis and description ontologies, especially for Human behaviour recognitions and interaction analysis. According to [1], in their survey on ontologies, they distinguished between:

1. **Data-driven approaches,** also known as probabilistic approaches, focusing on the branch of pattern recognition and machine learning, for recognition of human activities and the detection of anomalies during their performance using the information provided by sensors to build, infer, or calibrate a behaviour model. Machine-learning techniques have been extensively used with this purpose, and, more specially, probabilistic models, data mining, and inductive learning.
2. **Knowledge-based approaches,** include deterministic tools to model semantics and require a more accurate and refined activity knowledge representation. By opposition to the previous ones, deterministic techniques do not rely on learning. Instead, they use *a priori* knowledge to model the events to recognize.

A. Ontology on contextual information and context-aware

A significant amount of research on ontology has been done for the structural representation and recognition of contextual information [1], activities and interactions. Also, different taxonomies were proposed in different surveys [2], [3], like CONON (CONtext ONtology) [4], the Pervasive Information Visualization Ontology (PIVOn) [5] , the Context Aggregation and REasoning (CARE) middleware [6], and the fuzzy ontology [7] and [8]. A wide range of factors are used to classify previous work in human motion analysis and video understanding, such as: model-based vs. non-model based, functionality (tracking, pose estimation, and movement recognition), human-object interactions and group activities, action and activity recognition and classification, complex activities recognition and behaviour understanding, etc.

B. Ontology in the domain of video surveillance

In the domain of video surveillance, various approaches of ontologies and algorithms were used to address different stages of the problem. Video surveillance has its own set of most significant entities, terms, hierarchies, and relations. Due to the huge set of possible cases combined with the flexibility of description, the definition of unique video surveillance ontology is very ambitious and probably unfeasible. Nonetheless, a set of actions, events and entities can be selected due to their importance. The surveillance community has made some proposals for action, event, human activity and behaviour ontologies. Some shared concepts can be found among the following ontologies; also some ideas intersect with our proposed ones.

Video Surveillance Online Repository (VISOR) [9] is a platform for annotating, and retrieving surveillance videos, which used as a support tool for different projects. It contains a large set of multimedia data and corresponding annotations. VISOR provides a list of video surveillance concepts. The main concept of dividing between context and content is shared between many ontologies, including ours.

In [10] a behaviour ontology is proposed, mainly based on set of scene models linked by time relations. The set of scene models contains set of object models where low level data is specified, set of object relations, and set of object conditions. Some of the concepts of the object Model intersect with ours.

More recent works can be found in [11], ViVA ontology proposes three main classes Content, Context and System. VIVA was designed with the OWL format. Also, concepts concerning place, weather, location, and object may meet our same objectives, as follow; some of those influence our ontology.

The problem of action detection, recognition, understanding and especially the description in videos is acquiring an increasing importance, due to its applicability to a large number of applications. Moreover, because of the nature of the problem, it is necessary to consider the temporal dimension of video, requiring thus appropriate features and classification methods to deal with it, while taking into account the variability in the execution of events and actions, and the variety of scene types or contexts.

In order to work with a format well fitted to our later needs, simple to use and flexible at the same time, we created our own ontology, where concepts are hierarchically structured and defined.

III. Proposed Ontology

The varied nature of objects participating to a scene, the variety of scene types or contexts, and the complex nature of the object behaviours, actions and interactions and in the execution, requires an abstract level of information to reduce the size of the description scope. This work presents an ontology-based method that combines low-level primitives of objects basic features, like size, color, locations, speed and others, that should allow to intelligently deriving more meaningful high-level information.

In order to realize the knowledge-based and automatic generic description of video surveillance introduced in the previous section, the knowledge for video analysis is abstracted. Among many distinctive characteristics for this ontology, we mention that it:

1- focuses mainly on the objects interactions, nonetheless it is expendable.
2- presents detailed propositions about the interaction, from a methodologic and systematic approach.
3- is not directed by the results of the automatic analysis, and there is no pre-assumption or condition which restricts this ontology.
4- targets mainly the level of generic and abstract description, but it can be applied to any scene type or context.
5- shall be convenient to describe real interactions during incidents as they appear in CCTV control rooms reports.
6- focuses on new concepts concerning mediation, action at a distant and close interaction, deformable and non-deformable objects, and others.

Our proposed ontology, named "Video-Surveillance-Description Ontology" mainly describes the concepts that relate video, objects, and actions. It has been designed to be used as a generalist high-level layer for a video analysis, principally in video-surveillance system. It proposes six main classes: Object, Video, Context, Activity, Scene_Type and Descriptor (see Figure 1).

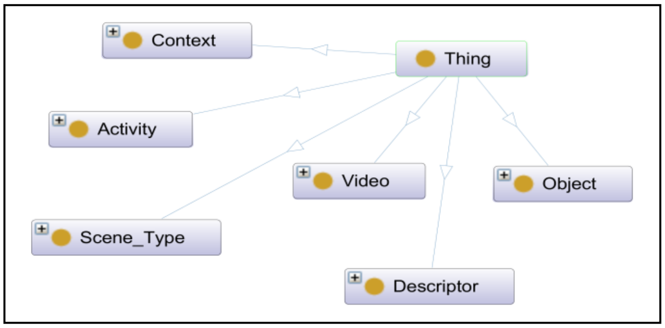

Fig. 1. Main classes: Object, Video, Context, Activity, Scene Type and Descriptor.

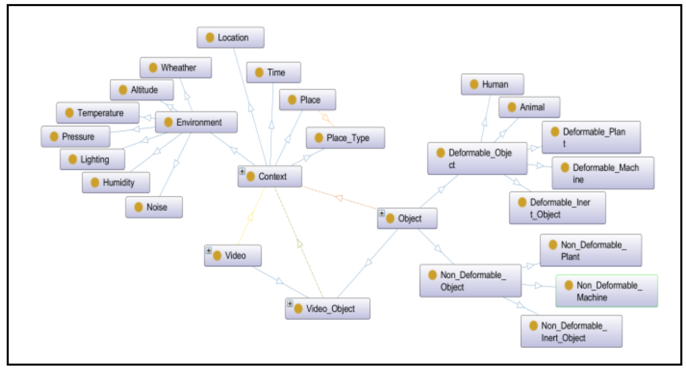

Fig. 2. The Video, object, video_object, and context classes.

A. *Context*

This class contains all the elements that provide information about the real context. For example: the GPS coordinates, the place where the action happens which can have two types: Indoor or Outdoor, the environment (weather, altitude, temperature, pressure, lighting, humidity, noise) and the time class.

B. *Object*

The **Object class** represents instances of humans, animals, plants, machines and all other inert objects. This class can represent all what exists in an environment. One of the most important features that can rule the way that an object can do the action, interaction or reaction is its deformability. The deformability criteria is mainly deduced from the object shape and is based on the degree of deformation [12]. From the surveillance point of view, non-deformable objects actions or reactions during an interaction are easy to detect, analyse, understand and maybe predict.

We chose to group all objects in two general sub-classes, deformable and non-deformable objects, depending on the properties of their appearance in the video. Humans and animals are generally "deformable". Plants, machines and inert objects are most of the time non-deformable.

For more information on how our proposed algorithm performs this classification, the reader is referred to [13].

C. *Video*

In visual surveillance systems, the cameras are mainly fixed. An object exists in a context. As the same object may appear several times in the same video, each appearance will be considered as an instance in Video_Object class. So, the Video_Object class is a subclass of video and object classes. This instance is delimited from the first moment of that appearance to the last one.

A **sub_object** is mainly used for deformable objects, for example for articulated segments of human and animal bodies, or parts of machines, etc. A video object can have several sub_Objects. As we may have many **states** for each appearance, each of the states describes the object / video_object state. Similarly many of the states can be taken for each of the sub_objects to create a sub_object_state. The number of states depends on the time of appearance, time of disappearance, and the suitable frame difference that we should take. In plus, for each state, the video object state can have many features like form, surface, displacement, speed, trajectory and many others.

D. *Activity and Action*

Different taxonomies are used for describing an action. We can find, among others, the terms operation, gesture, action, event, activity, and behaviour. So far, there is not a unique standard ontological definition of those notions or concepts.

Most of the authors agree to define human action as the simplest unit in human activity. A difference should be made between the terms human behaviour, events and activity to differentiate between the concepts of what a human is doing in the environment (activity), and the purpose or meaning it could have (behaviour). An Event is the occurrence of an activity in a particular place during a particular time interval. The Behaviour is a description of activities and events within a specific context. However, its refinement into various granularity levels is common to the usage of many notions in the various application domains.

Activities, according to [14], are units of life which are organized into three hierarchical layers. The top layer is the activity itself, which is oriented toward a motive, corresponding to a certain need. The motive is the object that the subject ultimately needs to attain. An activity it is understood as purposeful, transformative, and developing interaction between actors ("subjects") and the world ("objects") [15]. An activity is hierarchically structured into **Actions**, which are conscious processes directed at goals. In case of two or many objects, an action begins when one of those objects has the intent to perform an action.

Another important concept is the mediation. The main distinctive features of humans, such as language, culture and society, the production and use of advanced tools, etc., all involve mediation; here we note the mediation of information as the most important one among interactions. They represent different aspects of the same phenomenon, that is, the emergence of a complex system of structures and objects, both immaterial and material which serve as mediating means embedded in the interaction between human beings and the world and shaping the interaction. In cultural-historical psychology, mediation is, arguably, the most important concept of all; it serves as the cornerstone of the activity theory as a whole [15].

An example of a mediation is a human shooting another object (human, animal, …). In this case the bullet can be considered as the mediation. We may equally well consider the linguistic interaction as a transmission of information, for example saying "Hello". Without any word, when two humans are carrying together an inert object, the information is passed by the inert object itself. In the opposite, when two animals are fighting, or when two animal are following each other, there is no mediation between the two objects or unmediated action. We can consider that the implicit information helps both objects to coordinate their interaction.

In the case of one, two or many objects, and where the action/interaction is unmediated, or at least well noticed visual mediation by the application, we distinguish between two action types:

a- There is no physical contact: then we consider "action at a distance" or "far action/interaction", for example: when two objects are running together, or when two humans are saluting each other, etc.

b- There is physical contact: then we consider "close action/interaction", for example when an object is turning on a fire, or when two humans are fighting.

An Action is a series of operations done by an object on nobody, object, or many objects. The operations are considered the lower-level units implementation of the action.

Operations are routine processes providing an adjustment of an action to the ongoing situation. They are oriented toward the conditions under which the subject is trying to attain a goal.

Accordingly to the object states, the **Interaction States** can document the state of interaction at a related moment (existence, type and level of aggressiveness).

We present the relations between components and action. But those relations can be the same for activity and operation, or for the interaction state. We mention that:

- An object or video_object or sub_object can have an action/interaction, and an action is done by an object or video_object or sub_object.
- A video contains an action - an action is viewed in a video.
- A video_object_state or a sub_object_state is a part of an action, and an action can have instances of a video_object_state or a sub_object_state.

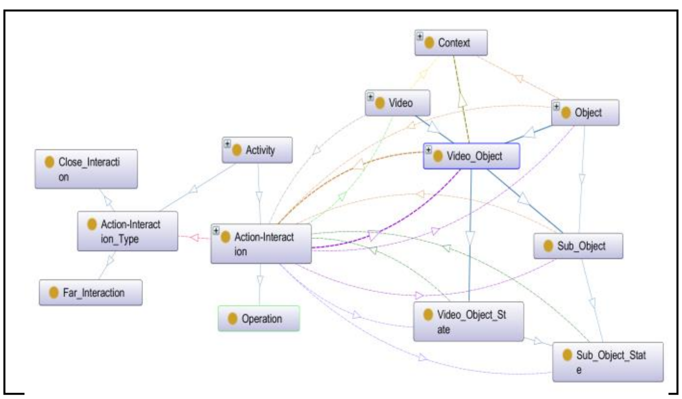

Fig. 3. The Sub_object, Video_Object_state, and Sub_object_state, and Action-Interaction classes.

E. Scene Type

To define a methodologic and systematic approach to describe the video scene types especially the action between video objects in video surveillance, we identify twelve types according to the number of moving objects and to their characteristics before and after the action.

1) ***0 Object (Scene without any moving objects)***: *when no objects are moving in the scene, still the environment is exposed to context changes. Example: day to night, etc.*

2) ***1 Object → 1a (Single object, no interaction with the environment)***: *when a single object is moving in the scene without any interaction. Examples: human walking, or doing sports, car passing, etc.*

3) ***1 Object → 1b (Single object, interaction with the environment)***: *when a single object is moving in the scene without any interaction with another moving object but mainly changing and interacting with the environment (background). Examples: person switching on the lights, smoking, crashing an ATM machine etc.*

4) ***1 Object → 1c (Single object, interaction with the inert objects of the environment)***: *when a single object is moving in the scene without any interaction with another moving object but changing and interacting with the inert objects of the environment; and doing so changing its characteristics either gaining (good influence) or losing (bad influence) some. Examples: person handling a box, person removing a wall picture, etc.*

5) ***1 Object → 2a (moving object trigger an inert object)***: *when a single moving object in the scene, at a given moment, performs an action with another inert object, and make it moving. Examples:, one ball hitting another fixed ball, person opening a door, etc.*

6) ***1 Object → 2b (moving object divides into 2)***: *when a single moving object in the scene, at a given moment, divides into 2 objects. Examples: person jumping out from a car, person removing his hat, etc.*

7) ***2 Objects → 1a (moving object stops another moving object)***: *when there are two moving objects in the scene and, at a given moment, one object does an action that stops the other object. Examples: a moving car hits a moving person, etc.*

8) ***2 Objects → 1b (2 moving objects merge into 1)***: *when there are two moving objects in the scene that, at a given moment, merge into one single object. Examples: a person jumping on a moving skateboard, etc.*

9) ***2 Objects → 0 (2 moving object stops after interaction)***: *when there are two moving objects in the scene that, at a given moment, interact and stop moving. Examples: two cars collide and stop, etc.*

10) ***2 objects → 2a (2 moving objects without interaction)***: *when there are two moving objects in the scene without any interaction between them. Examples: two cars passing near each other, two human passing by without any far or close interaction, human and animal co-appear in a scene without any kind of interaction, etc.*

11) ***2 Objects → 2b (2 moving objects with interaction)***: *when there are two moving objects in the scene that, at a given moment, interact, and then continue. Examples: two cars passing near each other trying to avoid a collision, two human walking together, two human saluting each other, two human boxing, etc.*

12) ***Many Objects → Many Objects (Group of moving objects with interaction)***: *when there are many moving objects in the scene, interacting together at a given moment, and continuing after. We do not consider here many objects in the scene so that the interaction can be divided in couples.*

This category is meant to describe scenes with a crowd. Anyway, this category may be divided into many other ones, but as it is not our field of interest, we preferred to keep it as one category. Examples: group fighting, or cheering, etc.

Concerning the **Scene_Sub_Type**, we may introduce more detailed interaction categories, such as: at distance or physical, Aggressive or Peaceful.

F. Description

This class is intended to describe the whole scene from objects to action/interaction and context, according to the scene type and sub_type. It contains two main sub_classes: Abstract_Description, and Semantic_Description. Those descriptions of a scene can be done using two methods:

1- Holistic method: this method takes the whole scene as one single closed box. It does not require for example the localization of body parts, object or action identification; the most important is what happens. Using this method, we consider all the possible combinations of actions/interactions in order to recognize, later, which one is the closest to this scene action. It is considered that the actuator actuated and action as a single box.

2- Detailed method: it is the study of each element of the scene, where the identification of each object, sub-object, action, operation, element apart, is required.

Then, the scene description, according to the scene type and sub-type and the method used, can be a generic abstract (context free) or a much more semantic text where the context has a big influence.

In figure 4 we present the abstract description used in this study. To generate a semantic description one can add to an instance of this abstract description the information taken from the context, like location, time, place, and place, etc.

For example:

- Abstract description: At frame 201, "Deformable" object "1" enters the scene, from "C" spot, on the "Left Middle" of the "Outside" area of the camera field of view, heading "Up Left", having respectively "regular" form, "small" surface, and "slow" speed.
- Semantic description: In 10/11/2018, at 11:35:22, a Human "1" enters the scene, from the conjunction "Verdun-Dunant", on the side of Verdun street, heading toward "Alfred Nobel" Street, having respectively small body, and "slow" speed.

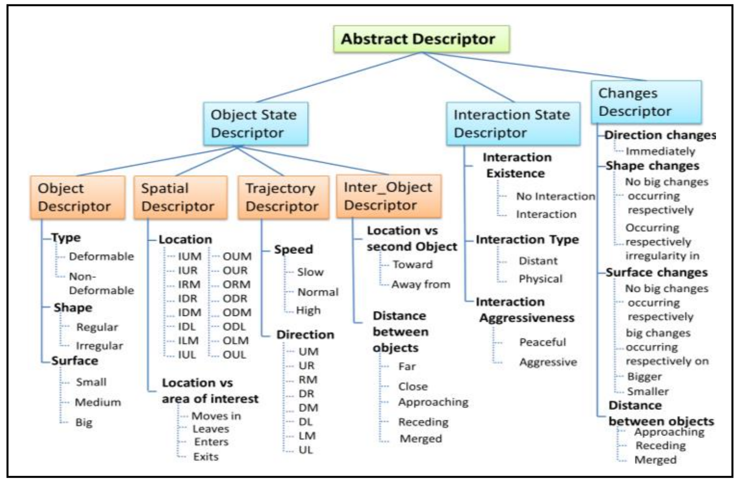

Fig. 4. Abstract description, having in the location and direction: U (Up), M (Middle), D (Down), R (Right), L (Left), I (inside), and O (outside)

Finally, we present all mentioned components of the "Video-Surveillance-Description Ontology" in Fig. 5.

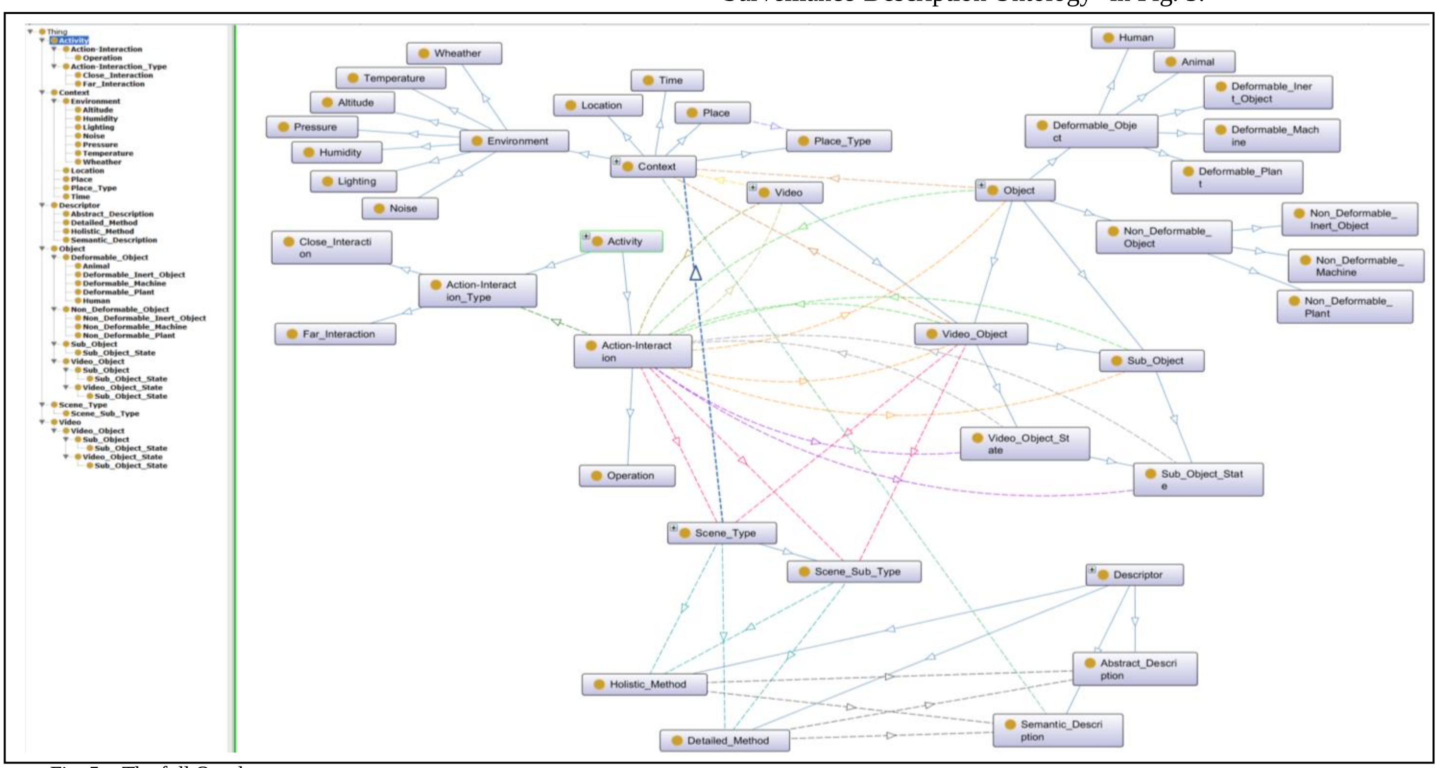

Fig. 5. The full Ontology summary.

TABLE I. EXEMPLE OF VIDEO SURVEILLANCE GENERATED DECRIPTION: TAKEN FROM THE SCENE "FIGHT_RUNAWAY2", OF THE DATABASE "CAVIAR"[16], WHERE TWO PERSON ENTER THE SCENE (FRAMES 193, AND 202), BEGIN TO APPROACH AGGRESSIVELY (FRAME 290), THEN START FIGHTING (FRAME 321), AFTER THAT THEY RUN AWAY (FRAME 468), FINALLY THEY EXIT THE SCENE (FRAMES 488, AND 491).

Frame number	States Descriptions
193	**Object 1:** "*Deformable*" object "1" enters the scene, *from "C" spot,* on the "*Left Middle*" of the "*Outside*" area of the camera field of view, heading "*Up Left*", having respectively "*regular*" shape, "*small*" surface, and "*slow*" speed.
202	**Object 1:** "*Deformable*" object "1" moves, *in "C" spot,* on the "*Left Middle*" of the "*Outside*" area of the camera field of view, heading "*Up Left*", "*No big changes occurring respectively on*" its shape, and "No big changes occurring respectively on" its Surface, and having respectively slight "Increasing" of its Speed.
	Object 2: "*Deformable*" object "2" enters the scene, *from "A" spot,* on the "*Down Right*" of the "*Outside*" area of the camera field of view, heading "*Up Left*", having respectively "*regular*" shape, "*medium*" surface, and "*Normal*" speed.
	Object 1 and Object 2: The two objects are respectively "far", No Interaction occurs between them.
290	**Object 1:** "*Deformable*" object "1" moves, in "F" spot, on the "*Left Middle*" of the "*Outside*" area of the camera field of view, heading immediately "*Right Middle*", "Occurring respectively irregularity in " its shape, and "big changes occurring respectively on" its Surface having now "Bigger" one, and having respectively considerable "decreasing" of its Speed.
	Object 2: "*Deformable*" object "2" moves, in "F" spot, on the "*Left Middle*" of the "*Inside*" area of the camera field of view, heading immediately "*Up Middle*" , "Toward" the object "1", "Occurring respectively irregularity in " its shape, and "big changes occurring respectively on" its Surface having now "Bigger" one, and having respectively considerable "increasing" of its Speed.
	Object 1 and Object 2: The two objects are respectively "Approaching", A "Distant" "Aggressive" Interaction occurs between them.
321	**Object 1:** "*Deformable*" object "1" moves, in "F" spot, on the "*Left Middle*" of the "*Inside*" area of the camera field of view, heading "*Left Middle*", "No big changes occurring respectively on" its shape, and "No big changes occurring respectively on" its Surface, and having respectively considerable "decreasing" of its Speed.
	Object 2: "*Deformable*" object "2" moves, in "F" spot, on the "*Left Middle*" of the "*Inside*" area of the camera field of view, heading "*Up Left*" , "Toward" the object "1", "No big changes occurring respectively on" its shape, and "No big changes occurring respectively on" its Surface , and having respectively "stable" Speed.
	Object 1 and Object 2: The two objects are respectively "Merged", A "Physical" "Aggressive" Interaction occurs between them.
468	**Object 1:** "*Deformable*" object "1" moves, in "D" spot, on the "*Left Middle*" of the "*Outside*" area of the camera field of view, heading "*Left Middle*", "No big changes occurring respectively on" its shape, and "No big changes occurring respectively on" its Surface , and having respectively slight "decreasing" of its Speed.
	Object 2: "*Deformable*" object "2" moves, in "E" spot, on the "Up *Left*" of the "*Outside*" area of the camera field of view, heading "*Down Left*", "No big changes occurring respectively on" its shape, and "No big changes occurring respectively on" its Surface, and having respectively slight "increasing" of its Speed.
	Object 1 and Object 2: The two objects are respectively "Approaching", No Interaction occurs between them.
488	**Object 1:** "*Deformable*" object "1" exits the scene, in "D" spot, on the "*Left Middle*" of the "*Outside*" area of the camera field of view, heading "Toward" the object "2", "Occurring respectively irregularity in " its shape, and "No big changes occurring respectively on" its Surface.
	Object 2: "*Deformable*" object "2" moves, from "E" spot, on the "Up *Left*" of the "*Outside*" area of the camera field of view, heading "*Up Left*", "No big changes occurring respectively on" its shape, and "No big changes occurring respectively on" its Surface, and having respectively "stable" Speed.
	Object 1 and Object 2: The two objects are respectively "Approaching", No Interaction occurs between them.
491	**Object 2:** "*Deformable*" object "2" exits the scene, from "E" spot, on the "Up *Left*" of the "*Outside*" area of the camera field of view, "No big changes occurring respectively on" its shape, and "No big changes occurring respectively on" its Surface, and having respectively "stable" Speed.

IV. CONCLUSION

In this study, we proposed a generic ontology for video description, mainly for video surveillance, taking into consideration some shared concepts as context, object, sub_object, activities, etc. Also, it presents some entities with new concepts like deformable/non-deformable object, twelve scene types, close/far interaction, aggressiveness of interaction, etc. This ontology was used to fulfill an automatic textual description of video surveillance, focusing on interactions between two objects, while using deep neuron network algorithm for interaction classification, after extractions of many important and very useful features. An example of generated description is shown in table I above. The textual descriptions can be built on during live monitoring for alerting relevant observers to areas of concern, and during post incidents investigation for intelligently fetching the Big-data resting in the archives.

REFERENCES

[1] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, "A Survey on Ontologies for Human Behavior Recognition," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 43:1–43:33, Mar. 2014.

[2] A. N. Mohamed, "A Novice Guide towards Human Motion Analysis and Understanding - Semantic Scholar," *arXiv preprint arXiv:1509.01074*, 2015.

[3] C. Villalonga *et al.*, "Ontology-based high-level context inference for human behavior identification," *SENSORS*, vol. 16, no. 10, pp. 1–26, 2016.

[4] X. H. Wang, D. Q. Zhang, T. Gu, and H. K. Pung, "Ontology based context modeling and reasoning using OWL," in *IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second*, pp. 18–22, 2004.

[5] R. Hervás, J. Bravo, and J. Fontecha, "A Context Model based on Ontological Languages: A Proposal for Information Visualization," *J. UCS, 16*, pp. 1539–1555, 2010.

[6] A. Agostini, C. Bettini, and D. Riboni, "Hybrid Reasoning in the CARE Middleware for Context Awareness," *Int. J. Web Eng. Technol.*, vol. 5, no. 1, pp. 3–23, May 2009.

[7] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, "A Fuzzy Ontology for Semantic Modelling and Recognition of Human Behaviour," *Know.-Based Syst.*, vol. 66, no. 1, pp. 46–60, Aug. 2014.

[8] J. A. Morente-Molinera, I. J. Pérez, M. R. Ureña, and E. Herrera-Viedma, "Building and Managing Fuzzy Ontologies with Heterogeneous Linguistic Information," *Know.-Based Syst.*, vol. 88, no. C, pp. 154–164, Nov. 2015.

[9] R. Vezzani and R. Cucchiara, "Video Surveillance Online Repository (ViSOR): An Integrated Framework," *Multimedia Tools Appl.*, vol. 50, no. 2, pp. 359–380, Nov. 2010.

[10] N. Q. Ly, A. M. Truong, and H. V. Nguyen, "Specific Behavior Recognition Based on Behavior Ontology," in *Recent Developments in Intelligent Information and Database Systems*, D. Król, L. Madeyski, and N. T. Nguyen, Eds. Cham: Springer International Publishing, pp. 99–109, 2016.

[11] M. A. P. Alonso, P. H. Leal, H. J. Escalante, and E. S. Succar, "ViVA Project." .

[12] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata, "Nonrigid Motion Analysis," *Comput. Vis. Image Underst.*, vol. 70, no. 2, pp. 142–156, May 1998.

[13] W. F. Youssef, S. Haidar, and P. Joly, "Classifying deformable and non-deformable video objects," in *7th International Conference on Imaging for Crime Detection and Prevention (ICDP)*, pp. 1–6, 2016.

[14] A. N. L. Leontyev, "Activity, Consciousness, and Personality.," *Englewood Cliffs, NJ Prentice-Hall. - References - Scientific Research Publishing*, 1978.

[15] V. Kaptelinin and B. A. Nardi, "Acting with Technology: Activity Theory and Interaction Design". *MIT Press*, 2006.

[16] "CAVIAR: Context Aware Vision using Image-based Active Recognition" *[Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.* [Accessed:20-Nov-2018].