




Automatic textual description of interactions between two objects in surveillance videos

Wael F. Youssef¹ · Siba Haidar²  · Philippe Joly³

Received: 28 June 2020 / Accepted: 24 March 2021

Published online: 10 June 2021

© The Author(s) 2021 

Abstract

The purpose of our work is to automatically generate textual video description schemas from surveillance video scenes compatible with police incidents reports. Our proposed approach is based on a generic and flexible context-free ontology. The general schema is of the form [actuator] [action] [over/with] [actuated object] [+ descriptors: distance, speed, etc.]. We focus on scenes containing exactly two objects. Through elaborated steps, we generate a formatted textual description. We try to identify the existence of an interaction between the two objects, including remote interaction which does not involve physical contact and we point out when aggressivity took place in these cases. We use supervised deep learning to classify scenes into interaction or no-interaction classes and then into subclasses. The chosen descriptors used to represent subclasses are keys in surveillance systems that help generate live alerts and facilitate offline investigation.

Keywords Textual description · Video surveillance · Aggressive interaction · Scene analysis · Video understanding · Natural language

1 Introduction

Among the most modern public safety and law enforcement tools are video surveillance systems. They provide a significant source of data, becoming the strong point of most investigations. A fundamental professional need is to extract useful information from the massive quantity of visual data generated by these surveillance systems.

Most existing video surveillance systems provide only the infrastructure to capture, transmit, store and distribute video images. Tedious tasks such as detecting an incident in a live stream or searching the archives for a specific scene depend on scarce and expensive human resources. The automatic scene description is essential in video surveillance. It facilitates the post-processing of incidents,

like dispatching of patrols. Moreover, video search can be promoted to another level by introducing object detection and tracking, as well as some more specific features such as direction, shape, deformability, or interaction. In particular, motion is a cue feature; the key is to focus on a non-linear motion like the one we may observe during an interaction between objects.

This work proposes a new approach for a generic context-independent textual description of video surveillance scenes from the real-world. In this paper, we present new representations for the sentences based on well-structured templates, which can be applied to generate a scene description, similar to those used in police reports. Our approach is based on our ontology described in [1]

✉ Siba Haidar, siba.haidar@ul.edu.lb; Wael F. Youssef, waelfyoussef@gmail.com | ¹Beirut CCTV Control Room, Police of Beirut, Internal Security Forces, Beirut, Lebanon. ²Lebanese University, Beirut, Lebanon. ³IRIT, Université de Toulouse, CNRS, UT3 Toulouse, France.



and explained in detail in the third paragraph, "Proposed Approach".

This paper mainly consists of seven paragraphs, in addition to the introduction; In the next section, we present the works related to the automatic video scene description.

In paragraph 3, we present our proposed approach. We introduce how we produce activity matrices of useful characteristics that can be used for generating alerts and querying the scenes. As well, we present how to generate textual descriptions from these matrices. Next, in paragraph 4, we explain the experiments and results.

Then, in paragraph 5, we discuss the drawbacks and difficulties of our approach and its advantages.

Finally, in paragraph 6, a general conclusion, primary contributions and future works of this work are presented.

2 State of the art

Automatic video scene description includes understanding and differentiating between the diversity of backgrounds, objects, interactions and scenes types. Moreover, it requires a translation of the information into a comprehensible textual description or what is known as natural language. In the last decade, researchers have studied multiple strategies and ontologies to bridge the gap between visual content and textual description. For that, computer vision and natural language processing (NLP) fields are addressing such an issue separately or jointly [2].

In the state of art, two main types of approaches can be noticed; behaviour understanding and sentence generation approaches on one side and sequence learning approaches.

Sequence learning approaches directly learn how to match video content and sentences. This approach can be divided into video encoding and decoding stages. In the encoding stage, the visual features are directly extracted and learnt using different types of deep neural network algorithms, like CNN, RNN or LSTM. The produced result composes a fixed or dynamic real-valued vector. In the video decoding stage, the resulting vector from the first stage is used for text generation. The main techniques may involve speech recognition, language modelling, image captioning, translation and more. These approaches are considered domain-specific, suitable for short video clips with limited vocabularies of objects and activities [3]. Some interesting works following the sequence learning approach for video description can be seen in [4, 5].

Behaviour understanding mainly relies on extracting features used to train individual classifiers to identify background, objects and actions in the scenes. Sentence generation generally requires a template with some syntactical structures, like Subject-Verb-Object SVO tuples. It may use

a probabilistic model to map the essential visual content results from the video with each template element. However, these templates are mostly dedicated to some non-generic variety of scenes [6].

Working specifically on video surveillance scenes, most of the research works targeting the description of these scenes follow the behaviour understanding and sentence generation approaches [7, 8]. Different templates have been proposed, like:

- (Object) (Action) in (Place) [at (high/low/middle) speed].
- A (colour) (size) (speed) (object type) coming from (entry zone) toward (exit zone).

The complicity here is due to the diversity of scenes in terms of location, object types and number, and various actions and interactions.

To simplify this problem, researchers proposed to add more assumptions. These propositions focused on particular types of objects, specific contexts and some specific actions. But these assumptions added more restrictions to the applicability in the real world. As an example, we mention [9] where their system assumes a surveillance scene with some prior region information and only four object classes: pedestrian, car, bike and cycle, while no interaction was considered. Another example is [10], where the concepts include vehicle, people and traffic signs, to allow users to annotate traffic events. Some other researches [11, 12] focus only on one type of interaction; the aggressivity.

Our goal will be to design a description system for video surveillance, taking advantage of behaviour understanding and sentence generation approaches, integrating some improvements by applying machine and deep learning methods.

3 Proposed approach

Our approach's key idea is to leverage meaningful content features from the scene for better understanding and appropriate description. We compute a dense set of spatiotemporal feature vectors to provide a localised description of the action. These features are well selected to satisfy the need of police operators and investigators. They are considered useful in generating alerts, enquiring the video surveillance footage and inferring textual sentences. We input those features into a supervised learning method for interaction classification. The proposed framework is conceptually simple and has low storage and computational cost, making it attractive for real-time implementation.

Our approach, named Video Surveillance Scene Description (VSSD), includes the following stages; First, we segment the video based on the number of objects in the

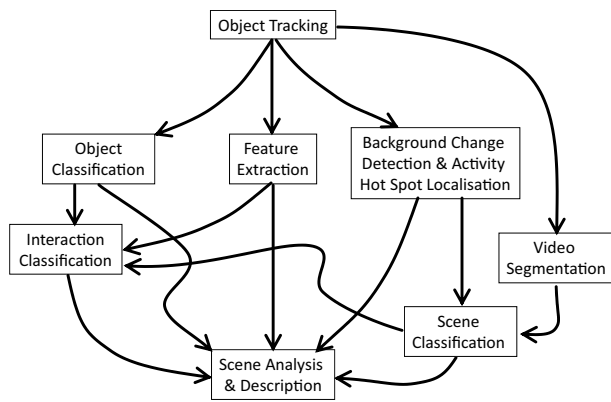


Fig. 1 Proposed approach diagram, where each phase connected to an arrow starting point feeds the phase connected to the terminal point of the corresponding arrow

scenes. We care to distinguish between deformable and non-deformable objects at this stage because there is a direct connection between the deformability of an object and its behaviour in an interaction. Second, we generate an activity map, highlighting the routes and the hot spots in the field of view (FOV) and indicating the background changes. Next, we extract several features and feed them to a DNN to serve for interaction classification. Then, we identify the critical moments in the scenes. And finally, we fill an activity matrix from which we generate the textual description which follows our templates.

Only two presumptions exist: 1- fixed camera, 2- only two moving objects in the scene. Although it is of high interest, there is a lack of scientific studies targeting this particular case.

In the diagram shown in Fig. 1, we present our approach’s workflow.

3.1 Segmentation and tracking

Many tests were done on available segmentation and multi-object tracking algorithms. A search for functional segmentation and tracking methods led to algorithms and methods shown in Table 1 and others. After comparative tests, we selected the method in [19], called "Motion-Based

Multiple Object Tracking", provided by Matlab. This algorithm is based on two main steps; 1- Detecting moving objects in each frame based on Gaussian Mixture Models and 2- Tracking the moving objects from frame to frame, based on Kalman filter.

The next step is to cut the video into scenes according to the number of present objects; zero, one, two or more.

3.2 Object classification

From the surveillance point of view, non-deformable object actions are easy to analyse. While deformable parts of an object move freely in an unpredicted way, making the interaction more complicated and harder to interpret.

Our main goal is to have an abstract description. We could not find, for our knowledge, a generic algorithm that can segment any object type into its semantic sub-components under different external factors proof, i.e., invariant regarding the scene visibility and the lighting. Therefore, we limit the classification and restrain from the analysis of sub-object segments in the presence of interacting deformable objects. We check the blobs encasing each object to determine if it is deformable or not. The applied method is ours and has been published in [20].

3.3 Background model and activity localisation

Once the object segmentation task is accomplished, we now know whether a pixel belongs to an object or the background for each video frame. Using this information, we build the coefficient matrix C_t . C_t is a cumulative matrix, where each value represents the number of times the corresponding pixel belonged to an object, from frame 0 until frame t . The matrix is obtained by cumulating individual binary matrices M_t , each for a given frame at time t , in (1) and (2):

$$M_t(x, y) = \begin{cases} 0, & p(x, y) \in \text{an object at time } t \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

where $p(x, y)$ is the pixel of coordinates (x, y) .

Table 1 List of some tested algorithms for objects segmentation and tracking

#	name	reference
1	Tracking Interacting Objects	[13]
2	Discrete–Continuous Energy Minimization for Multi-Target Tracking	[14]
3	Continuous Energy Minimization for Multi-Target Tracking	[15]
4	GMCP-Tracker	[16]
5	Moving-Target-tracking-with-openCV	[17]
6	Online Multi-Object Tracking by Decision Making	[18]
7	Motion-Based Multiple Object Tracking	[19]

$$C_t = \sum_{k=0}^t M_k \tag{2}$$

Based on this coefficient matrix, next, we will calculate the background model.

3.3.1 Background change detection

Since the camera is fixed, we can locate the background parts that have changed due to the objects' movement, the routes followed by them and the activity's spots.

The temporary background model B_t is a matrix representing the temporary cumulative background, where all pixels from the frame are taken into consideration except the one belonging to an object in the current frame. It is calculated as in (3):

$$B_0 = I_0 \cdot C_0 \tag{3}$$

Then for each frame of the sequence, we update the value of the matrix as follows in (4):

$$B_t = (B_{t-1} \cdot C_{t-1} + I_t \cdot M_t) / C_t \tag{4}$$

The dot symbol (.) represents the dot product in matrix operation, i.e., point to point multiplication and the division symbol (/) also represents the point to point division, unless the denominator is null; in this case, the result is set to zero. I_t is the frame at time t .

In a video sequence with n frames $S = \{I_t | t = 0..n\}$, then B_n represents the background model. It is merely the average image without the moving objects.

In reality, more than one background model may be needed, especially in long scenes cases. This is due to possible permanent changes in the background, light changes or inert objects displacement. Therefore, the algorithm generates and uses a new background model every time a large percentage of pixels significantly change their values for a relatively long period.

3.3.2 Routes and activity spots plan

In the coefficient matrix C_t , low values represent the flows of the moving objects. To find out the routes and the map of activities, first, we normalise C_n : $C_n = C_n/n$. Then, we apply on C_n a simple image processing method in three steps:

1. Pixel quantification: to reduce the pixel intensity interval into four values, respectively: no activity, marginal areas, regular routes and hot spots representing highly frequented locations.
2. Morphological filtering: we perform opening and closing to smooth the map.

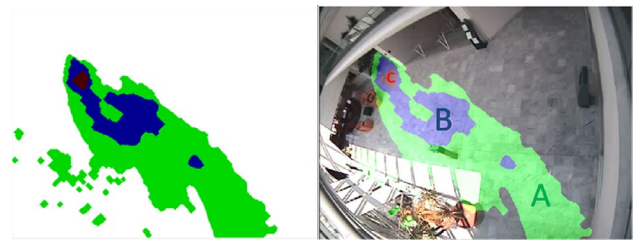


Fig. 2 Scene model (right) obtained by adding the processed coefficient matrix to the background model. Marginal areas **A**, routes **B** and activity hot spots **C** are shown in different colours. Scene "Left-Box" from the dataset "CAVIAR" 2004

3. Background addition: the resulting matrix is added to the background model B_n .

We obtain a scene model which can be used for detecting anomalies and in the description phase. Figure 2 shows a scene model example.

3.4 Feature extraction

The choice of the features to extract from the videos' scenes impacts the efficiency and accuracy of the method.

Five types of features were extracted:

1. Object spatial features: mainly dimensions (width, height, surface, perimeter), position, shape (bounding box, intensity, RGB, Hu moments [21]) and type (deformable/non-deformable) of each of the objects.
2. Object temporal features: variations of spatial features between frames, including displacements such as distance, speed and angle.
3. Inter-objects features: the difference of spatial features between the two objects.
4. Inter-frames features: for most of the above features f , we extract the derivative f' and the second derivative f'' . Then for each f, f' and f'' , we find seven global inter-frames features in a fixed-size window: the minimum, the first, the last, the middle, the average, the median and the standard deviation, normalised by the maximum value.
5. Trajectory features: we consider three trajectories, one for each object centroid and one for their middle point, to which we apply three smoothing filters: average (5), first-order (6) and second-order prediction (7) according to Taylor development.

$$x_p(t + 1) = (x(t) + x(t + 2))/2 \tag{5}$$

$$x_p(t + 1) = x(t) + x'(t) \tag{6}$$

$$x_p(t + 1) = x(t) + x'(t) + (x''(t))/2 \tag{7}$$

For each of the nine trajectories, we calculate two features; (a) the standard deviation of the distance between the filtered position and the one corresponding to the centroid and (b) the largest distance. Those values give information about the smoothness of the trajectories.

3.5 Scene classification

In our ontology [1], we proposed classifying the video scenes into 15 types according to the number of objects before and after any interaction combined with other factors like background changes and feature changes. We mainly monitor the changes of two characteristics, the Hu moments and the surface of the object. Table 2 summarises the expected evolutions of those features in different cases. We show in Fig. 3 an example of scene type 7, corresponding to an object left in the scene; the scene is from the dataset VISOR [22].

3.6 Interaction classification

An essential task for the safety observers of public places is to identify irregular actions. Police officers want to observe any unusual interaction between humans or vehicles. An interaction between two objects is called distant or remote if it does not involve physical contact. Most of the physical interactions are preceded by remote ones. Note, as an example, that people shout before they start fighting. Hence, it is crucial to detect the moment a distant interaction starts.

Furthermore and for the benefit of law enforcement, the interaction’s aggressiveness is also to be identified.

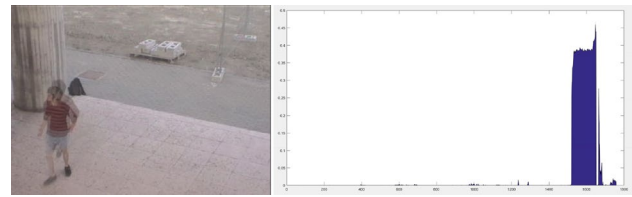


Fig. 3 (left) A scene where a person leaves a bag and then another person comes and takes it. (right) The graph of comparison between the frames and the corresponding temporary background model. X axis is the frame number, Y axis is the percentage of change between the frames and the temporary background model. The graph indicates critical background changes near the frame 1500 and 1700, the moments of bag deposit and retrieval. VISOR Dataset 2017

We add accordingly two more subclasses: aggressive and peaceful interaction.

Interaction classification is accomplished using three different multi-layered DNNs. First, we apply a binary interaction classification over fixed-size windows of the scenes to identify the existence of interaction. Next, scenes with interaction undergo two more in-depth analysis adding the sub-classes: distant or physical and aggressive or peaceful. Post-processing, each window classification result is smoothed regarding the temporal-consistency [23] to label the complete scenes.

3.7 Scene analysis and description

We inspire our description schema from the real incidents’ cases. Incident police reports should contain the five Ws; *who, what, where, when* and *why*. For objectivity purpose,

Table 2 Scene classification into 15 types, according to number of objects before and after the action, to the background changes and to objects’ features

#scene type	#obj before	#obj after	bckg changes	Feature changes
1	NA	NA	NA	NA
2	1	0	No	
3	1	0	Yes	
4	1	1	No	
5	1	1	Yes	No Hu moments changes
6	1	1	Yes	Hu moments changes & surface changes
7	1	1	Yes	All features change
8	1	2	Yes	
9	1	2	No	
10	2	1	Yes	
11	2	1	No	
12	2	0	Yes	
13	2	2		
14	2	2		
15	Many	Many		

we replace the *why* with *how*. The latter key points draw a frame to our textual description.

At relevant moments in the scene, two kinds of descriptions are generated. The first description concerns each moving object's initial state separately: position, deformability, speed and movement direction. The second description focus on the interaction between the two objects. It reports the type of interaction, its aggressiveness and its influence on each of the objects after-state.

3.7.1 Scene key moments

To choose the key moments suitable for description, we rely on irregularities in the objects' characteristics or the interaction. By irregularities, we mean sudden changes, local maxima or minima and so forth. Several characteristics determine the behaviour of an object:

1. Object characteristics: deformability, shape (invariant Hu moments, surface), relative position by reference to labelled areas or an area of interest, see Fig. 4, and displacement related (speed, direction).
2. Inter-object characteristics: the distance between the objects, and relative position or direction.
3. Interaction features: existence, distance, and aggressiveness.

These characteristics are represented using graph-based pattern discovery, from which we extract key moments to generate the description. An example of the keyframes for the scene "LeftBox" from the database [24] is shown in Fig. 4.

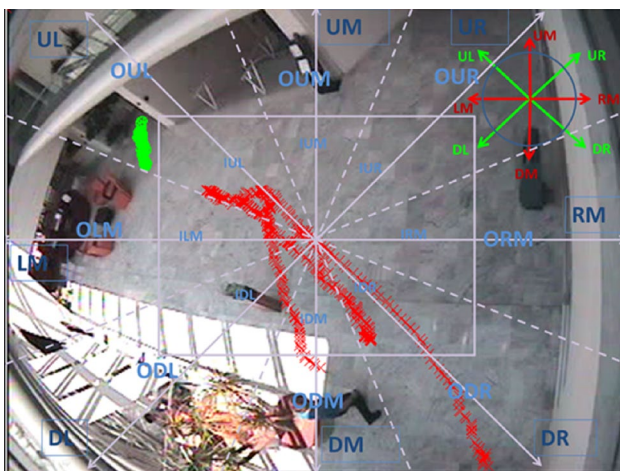


Fig. 4 An example of a key frame, over which we show the eight directions (in red and green) and the sixteen areas. We can also see the trajectories (object1 in red, object2 in green). Scene "LeftBox", CAVIAR dataset [24]

Key moments should be defined according to essential variations in the extracted features. The importance of the variations differs from a scene type to another and from one user needs to another. Considering that and as we want to keep our system generic, the description density control is given to the user as a hyper-parameter. Thresholds can be set as well, as hyper-parameters, to control the amount of the generated data. Triggered by the key moments, an activity matrix is generated.

3.7.2 Activity matrix

For each key moment $k_i, i \in \{1, 2, \dots, m\}$, corresponding characteristics are generated and filled in a vector V_i , see (8).

$$V_i = \{F, O1, O2, IO, IN\} \tag{8}$$

where:

- F is the frame number of the key moment k_i ,
 - $O1$ and $O2$ are the object's characteristics vectors,
 - IO contains the inter-object characteristics and.
 - IN has interaction features values.
- Now, consider:

$$A = \{V_i, i = 1 \dots m\} \tag{9}$$

The vectors V_i form the activity matrix A , Eq. (9), which then saves, for chosen key moments of the scene, a full set of features' values, describing it. This matrix can be used to build sophisticated scenario models based on context-dependent thresholds.

Finally, A is mapped into textual descriptions by simply applying logical rules to fill the blanks in our structured templates.

3.7.3 Scene description

We introduce template models for textual description. We propose two types of templates:

1. Object templates: for each of the objects, three templates were introduced. In these templates, words between quotes denote values, clauses between curly brackets surround options and square brackets indicate facultativity. At the key moment where an object enters or exits the scene, the description follows two slightly different templates, called "object entrance template" and "object exit template". In both cases, we use absolute description (ex: big) instead of comparative description (ex: bigger) of the current moment k_i relatively to the previous one k_{i-1} . At each of the key

moments of an object, other than the *entrance* and the *exit*, an "object template" is structured as follows:

```
"Type" object "ID" {
    moves in "Area of Interest" spot
    |
    leaves "Area of Interest" spot
    [ and moves in "Area of Interest" spot ]
    |
    moves
}
on the "Frame Area Symbol"
of the { inside | outside } area of the camera field of view,
heading [immediately] "Direction",
[
  { toward | away from } the object "ID",
]
{ no big changes occurring respectively on
  |
  occurring respectively irregularity in
  } its shape, and
{ no big | big } changes occurring respectively on its surface,
[ having now { smaller | bigger } one, ]
and having [respectively { considerable | slight } ]
{ increasing of its | decreasing of its | stable } speed.
```

2. Inter-objects templates: two templates were introduced.

a. "inter-objects entrance template": when the second object enters the scene, at this key moment, the description is activated and follows:

```
The two objects are respectively
{ far | close },
{ no interaction
  |
  a { distant | physical}
    { aggressive | peaceful } interaction
} occurs between them.
```

b. at each of the key moments, a proposed "inter-objects general template" is structured as follow:

```
The two objects are respectively
{ approaching | receding | merged },
{ no interaction
  |
  a { distant | physical}
    { aggressive | peaceful } interaction
} occurs between them.
```

In order to establish the mapping between the activity matrix A and the templates, we use logical threshold-based rules [1]. An example of these rules, a sample rule considering the position and the direction comparison between the two objects is presented below. Hereby two conditions were applied.

For object 1, at a key moment k_i , in a given scene, let α be the angle between the vector direction of object 1 and the vector direction formed by the centroid of object 1 and the centroid of object2 (10):

$$\alpha = \left(\overrightarrow{(c_{1,i-1}, c_{1,i})}, \overrightarrow{(c_{1,i}, c_{2,i})} \right) \quad (10)$$

Then, if $|\alpha| \leq 45$ add to description {toward the object 2}, if $|\alpha| \geq 135$, add to description {away from the object 2}.

Finally, the system generates two kinds of descriptions; a full one and a short one. The full description involves each key moment and includes all the features values. Only the features associated with the most important key moments appear in the short description, the ones showing irregularities. Tables 7 and 8 show an example of describing the scene "LeftBox" taken from the database "CAVIAR" [24]. We can see the activity matrix in Table 7 and the full description in Table 8.

"At frame 392: Object 2: Deformable object 2 moves, in F spot, on the right middle of the inside area of the camera field of view, heading immediately down right, toward the object 1, no big change occurring respectively on its shape and big changes occurring respectively on its surface having now bigger one, and having respectively slight increasing of its speed. Object 1 & Object 2: The two objects are respectively receding; a physical peaceful interaction occurs between them".

Example of the results we were able to generate for scene description at a single moment.

4 Experiments and results

Our proposed method consists of many stages, conceptual, modelling, then learning for classification. We did experiments to evaluate the learning stage of the process.

4.1 Dataset selection, preparation and pre-processing

There is no dataset dedicated for two-objects interaction in surveillance video. We examined many available general video surveillance datasets; some are mentioned in Table 3. Among these datasets, we extracted 323 scenes suitable for our experiments and we manually annotated them. We got a total of 1903s of videos. A small number was later discarded due to tracking failure. We pre-processed this crafted dataset by cutting it into fixed-size scenes, each of 25 frames. We obtained 6029 windows, from which only 2208 windows of interaction. Among the

Table 3 Used datasets

#	name	reference
1	BEHAVE Interactions Test Case	[25]
2	CAVIAR: Context Aware Vision using Image-based Active Recognition	[24]
3	“EPFL” data set: Multi-camera Pedestrian Videos	[26]
4	UT-Interaction dataset	[27]
5	Advanced Video and Signal based Surveillance	[28]
6	VISOR Video surveillance online repository	[22]
7	VIRAT Video Dataset	[29]

2208 interaction windows, were 5962 distant interactions and 67 physical ones. Then, we extracted the five types of features discussed earlier, which gave us 2498 features.

In the pre-processing data phase, we augmented the dataset volume by adding artificial scenes, to note: reverse footage. We duplicated, as well, the number of positive cases to balance the negative and positive inputs for the three classifications. The three classifications were trained and tested according to the dataset mentioned in Table 4.

4.2 Classification training and results

To implement our classification models, after many tests on classical ML and NN algorithms, we chose a multi-layered DNN; Feedforward fully connected networks called Pattern recognition networks [30] in MATLAB and Simulink environment were trained by backpropagation of error. To achieve the desired outputs, several tests were made. For each of the three classifications, one classification DNN model, showing best results, was selected. Table 5 lists the used parameters for the three DNN classifications and Table 6 shows the results for the three classifications.

5 Discussion

In our approach, we faced many challenges; however, important outcomes were delivered.

While the size of the dataset we crafted remains insufficient for a very sharp evaluation benchmark, the absence

Table 4 Balanced dataset input characteristics

classif	input records	feature / record	input records classifica-tion	
Inter. No	14,902	2305	Inter. 7657	No 7245
Dist. Phys.	7079	2303	Phys. 3640	Dist. 3439
Aggr. Peac.	6703	2303	Aggr. 3439	Peac. 3264

Table 5 Used parameters for the 3 classification DNNs

#scene type	# hidden layers	# neurons	# of epochs	Test set results
Inter. No	7	586	325	87.5%
Dist. Phys.	4	426	102	93.7%
Aggr. Peac.	4	546	88	93.8%
classif	train set	val set	test set	
Inter. No	80%	10%	10%	
Dist. Phys.	70%	15%	15%	
Aggr. Peac.	70%	15%	15%	

Other parameters: chosen algorithm: Pattern recognition network: feedforward network composed of fully connected layers—Sigma $\sigma: 5.0 e^{-7}$ —Lambda $\lambda: 5.0 e^{-5}$ —Activation function: *Logsig* for the hidden layers & *Softmax* for the output layer.

of a specific annotated dataset for two-objects interaction confirms that the problem is actual and still not resolved.

The segmentation and tracking algorithm suffered from:

1. One major issue: the traditional occlusion when two moving objects are physically close. Using the Kalman filter to estimate the location did not work well with moving occluded objects. The occluded object location and boundary box were estimated for some number of frames, while the foreground object is missed. Consequently, the physical interaction in a scene was detected for only some frames per scene. And then, analysing and describing the interaction had no more effect until the two objects separate.
2. One marginal issue: the false segmentation of the object when it is moving in a complex background. This issue can trigger a description declaring a significant change in the object Hu moments or surface. This can be overpassed at the level of thresholding.

A simple solution could be by replacing this algorithm. Having lately good results with detecting objects using deep learning, like YOLOv5 and Faster R-CNN, a good plan could be by testing these algorithms; then, if one delivers

Table 6 Results for the three classification DNN algorithms

Classif	True Positive	False Positive	False Negative	True Negative
Inter. No	568 43.1%	19 1.4%	146 11.1%	585 44.4%
Dist. Phys.	400 45.5%	24 2.7%	31 3.5%	424 48.2%
Aggr. Peac.	320 41.9%	8 1%	39 5.1%	397 52.0%

Table 8 The full description of the scene "LeftBox"[24], results of mapping matrix A into our proposed templates. At each key moment, corresponding to one of the irregularities mentioned in the description

of Table 7 and marked in red, the full state of the objects and the interaction are described. Notice: near frame 145 the two objects start a distant peaceful interaction, which ends near frame 190.

Info	Frame #	Desc target	Textual description	
450	11	Object 1	"Deformable" object "1" enters the scene, from "A" spot, on the "Down Middle" of the "Outside" area of the camera field of view, heading "Up Middle", having respectively "Regular" shape, "Small" surface, and "High" speed.	
		Object 2	"Deformable" object "2" enters the scene, from "B" spot, on the "Up Left" of the "Outside" area of the camera field of view, heading "Down Middle", having respectively "regular" shape, "small" surface, and "low" speed.	
Death Frame	101	Object 1	"Deformable" object "1" leaves, "A" spot, and moves, in "B" spot, on the "Up Left" of the "Inside" area of the camera field of view, heading immediately "Right 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" enters the scene, from "B" spot, on the "Up Left" of the "Outside" area of the camera field of view, heading "Down Middle", having respectively "regular" shape, "small" surface, and "low" speed.	
1		Object 1 & Object 2	The two objects are respectively "Far", No Interaction occurs between them.	
Birth Frame	110	Object 1	"Deformable" object "1" moves, in "B" spot, on the "Up Left" of the "Inside" area of the camera field of view, heading immediately "Down Right", "Away from" the object "2", "No big changes occurring respectively on" its shape, and "No big changes occurring respectively on" its surface, and having respectively "stable" speed.	
		Object 2	"Deformable" object "2" moves, in "B" spot, on the "Up Left" of the "Outside" area of the camera field of view, heading "Down Middle", "No big changes occurring respectively on" its shape, and "No big changes occurring respectively on" its surface, and having respectively "stable" Speed.	
		Object 1 & Object 2	The two objects are respectively "Receding", no Interaction occurs between them.	
177	145	Object 1	"Deformable" object "1" leaves, "B" spot, and moves, in "A" spot, on the "Down Middle" of the "Inside" area of the camera field of view, heading "Down Right", "Away from" the object "2", "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" moves, in "B" spot, on the "Up Left" of the "Outside" area of the camera field of view, heading "Down 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.	
ACTIVITY NUMBER	160	Object 1 & Object 2	The two objects are respectively "Receding", A "Distant" "Peaceful" Interaction occurs between them.	
		Object 1	"Deformable" object "1" leaves, "A" spot, and moves, in "B" spot, on the "Down Right" of the "Inside" area of the camera field of view, heading "Down Right", "Away from" the object "2", "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" leaves, "B" spot, and moves, in "C" spot, on the "Up Left" of the "Outside" area of the camera field of view, heading "Down Middle", "Toward" the object "1", "No big changes occurring respectively on" its shape, and "No big changes occurring respectively on" its surface, and having respectively considerable "increasing" of its Speed.	
48	185	Object 1 & Object 2	The two objects are respectively "Receding", A "Distant" "Peaceful" Interaction occurs between them.	
		Object 1	"Deformable" object "1" moves, in "B" spot, on the "Down Right" of the "Inside" area of the camera field of view, heading immediately "Up Middle", "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" moves, in "C" spot, on the "Up Left" of the "Outside" area of the camera field of view, heading immediately "Up 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.	
190	190	Object 1 & Object 2	The two objects are respectively "Approaching", No Interaction occurs between them.	
		Object 1	"Deformable" object "1" moves, in "B" spot, on the "Down Right" of the "Inside" area of the camera field of view, heading "Up Middle", "Toward" the object "2", "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" moves, in "C" spot, on the "Up Left" of the "Outside" area of the camera field of view, heading "Up Middle", "Away from" 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.	
VIDEO ID	230	Object 1 & Object 2	The two objects are respectively "Approaching", No Interaction occurs between them.	
		Object 1	"Deformable" object "1" moves, in "B" spot, on the "Left Middle" of the "Inside" area of the camera field of view, heading immediately "Up Left", "Toward" the object "2", "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" leaves, "C" spot, and moves, in "B" spot, on the "Up Left" of the "Outside" area of the camera field of view, heading "Up Middle", "Occurring respectively irregularity in" its shape, and "No big changes occurring respectively on" its surface, and having respectively considerable "increasing" of its speed.	
	238	238	Object 1 & Object 2	The two objects are respectively "Approaching", No Interaction occurs between them.
			Object 1	"Deformable" object "1" moves, in "B" spot, on the "Left Middle" of the "Inside" area of the camera field of view, heading "Up Left", "Toward" the object "2", "No big changes occurring respectively on" its shape, and "No big changes occurring respectively on" its surface, and having slight "decreasing" of its speed.
	412	412	Object 2	"Deformable" object "2" exits the scene, from "B" spot, on the "Up Left" of the "Outside" area of the camera field of view, heading "Up 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 1			"Deformable" object "1" leaves "B" spot and exits the scene, from "A" spot, on the "Down Right" of the "outside" area of the camera field of view, heading immediately "Down Middle", "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.	

better results and satisfies all the conditions, our selected tracking and segmentation algorithm can be replaced in our overall approach.

On the other hand, we claim that our VSSD approach differs positively in the following points. The input features are dedicated to the interaction classification process, where many of the other methods do not export appropriate features from the videos. VSSD takes into consideration the diversity of scenes and does not apply any restrictions.

VSSD implements the original classification idea of distant versus physical interaction, while other works focus only on the physical one. Detecting distant aggressive interaction can alert the observers, in a surveillance control room, at early stages, giving them precious time to act.

The object features are expressed and stored, forming a higher semantical set of metadata. Such annotation is immensely helpful to the end-user when querying the archives for an incident with a specific description. We tried to encounter, in those features, most of the queries used in practice to search for an incident in the archives.

Finally, the description alert flags depend on a set of hyperparameters whose values are surely context-dependent. This leaves the main handle to the user. The grammar is expandable; the system can be considered as a toolset. The new structured templates contain the primary information reported by the police in real case incident description. Consequently, the textual description can be generated automatically as draft reports.

6 Conclusion and Perspectives

In this work, we look at the fundamental problem of generating textual descriptions of important contents in video surveillance scenes. We based on our new generic context-free ontology focused on objects interactions. We point out some basic undercover problematic needs and provide many solutions.

While analysing and understanding a wide variety of video scenes, our approach introduces new concepts and highlights important features to classify video objects' interactions. We propose a new classification of scenes based on background changes, the number of objects and the study of specific object characteristics. We classify interactions using deep learning. We introduce a very useful activity matrix, highlighting appropriate key moments and serves for description based on graph pattern discovery.

Finally, we propose very effective rule-based templates to structure the textual descriptions. Our templates support CCTV reports for real incidents description since one of the authors is a Major at the head of a central CCTV Control Room in the Lebanese Internal Security Forces. And

therefore, we know that, when properly implemented and used, visual surveillance systems, supported by intelligent video analysis, can become a very effective weapon in law enforcement agencies' hands. We intend to learn the classifiers on a real dataset, but this project is highly critical and needs special permissions. From a practical point of view, we can imagine the generation of specific scene descriptions, integrating labelled contextual information.

We consider our work as a leading project that can be extended by analysing deeper levels of classifications. Also, we may add more semantic depth to the description model. To state as an example, deciding whether a specific interaction has "bad" or "good" overall influence.

Furthermore, to obtain better tracking and segmentation results, the current tracking and segmentation algorithm can be replaced by a more recent algorithm based on deep learning, potential YOLO v5 or Faster R-CNN.

In a more classic extension, it is interesting to apply our approach on more complex scenes, showing interactions between more than two objects.

Finally, as working on thousands of hours of videos surveillance footage, all the output data from our system, when applying, can form big data. This big data or metadata collected and accumulated can be then learned through clustering and regression to model and predict the objects' behaviours and interactions.

Authors' contributions This work was based mainly on Wael F. Youssef thesis, the primary author of this manuscript. The thesis is entitled "*Instantiation of a textual description schema of video surveillance scenes*", available online on the server of the University of Toulouse III—Paul Sabatier "<http://thesesups.ups-tlse.fr/4586/1/2019TOU30249.pdf>". All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by W. F. Y. and S. H. The first draft of the manuscript was written by W. F. Y. and S. H. and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the research project "Analysis and description of interaction between two moving objects in a video scene captured by a fixed camera" and was partially funded by the Lebanese University under the reference number "21948".

Availability of data and material Part of the dataset used in the experiments was derived from the referenced mentioned datasets. Derived data supporting this study's findings are available from the corresponding author W. F. Y. on request.

Another part is subject to confidentiality and needs a particular permission procedure to be communicated.

Authors confirm that all data and materials as well as the software application or custom code, support their published claims and comply with field standards.

Code availability All the approach custom code, including the video pre-processing, the features extractions, the tracking and segmentation, and description generation code, as well as the code for data

cleaning and analysis associated with the current submission, are available upon request.

Declarations

Conflicts of interest The authors have no conflicts of interest to declare that they are relevant to this article's content.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Reference

1. Youssef WF, Haidar S, Joly P (2018) Generic video surveillance description ontology. In: 1st international conference on big data and cyber-security intelligence (BDCSIntell 2018). BDCSIntell 2018. 6
2. Andrei Georgios E, Daniel H et al (2021): Language and vision workshop - 2018. <http://languageandvision.com/2018.html>. Accessed Feb 2021
3. Aafaq N, Mian A, Liu W, Gilani SZ, Shah M (2018) Video description: a survey of methods, datasets and evaluation metrics
4. Venugopalan S, Xu H, Donahue J, Rohrbach M, Mooney R, Saeenko, K (2014) Translating videos to natural language using deep recurrent neural networks
5. Pan Y, Mei T, Yao T, Li H, Rui Y (2016) Jointly modeling embedding and translation to bridge video and language. In: proceedings of the IEEE conference on computer vision and pattern recognition. pp 4594–4602
6. Xu J, Mei T, Yao T, Rui Y (2016) MSR-VTT: a large video description dataset for bridging video and language. In: proceedings of the IEEE conference on computer vision and pattern recognition. pp 5288–5296
7. Lou J, Liu Q, Tan T, Hu W (2002) Semantic interpretation of object activities in a surveillance system. In: 'Object recognition supported by user interaction for service robots' 16th International Conference on Pattern Recognition (IEEE Comput. Soc 2002). pp 777–780.
8. Ahmed SkA, Dogra DP, Kar S, Roy PP (2019) Natural language description of surveillance events. In: Chandra P, Giri D, Li F, Kar S, Jana DK (eds) Information technology and applied mathematics. Springer, Singapore, pp 141–151
9. Kojima A, Tamura T, Fukunaga K (2002) Natural language description of human activities from video images based on concept hierarchy of actions. *Int J Comput Vis* 50(2):171–184
10. Xu Z, Hu, C, Mei L (2016) Video structured description technology based intelligence analysis of surveillance videos for public security applications. *Multimed Tools Appl* 75(19):12155–12172
11. Zizi TKT, Zizi T, Ramli S et al (2017) Aggressive movement detection using optical flow features base on digital & thermal camera. (193):6
12. Zulkifley MA, Samanu NS, Zulkepli NAAN, Kadim Z, Woon HH (2016) Kalman filter-based aggressive behaviour detection for indoor environment. In: Kim KJ, Joukov N (eds) Information science and applications (ICISA) 2016, Springer, pp 829–837
13. Wang X, Türetken E, Fleuret F, Fua P (2016) Tracking interacting objects using intertwined flows. *IEEE Trans Pattern Anal MachIntell* 38(11):2312–2326
14. Andriyenko A, Schindler K, Roth S (2012) Discrete-continuous optimisation for multi-target tracking. In: '2012 IEEE Conference on Computer Vision and Pattern Recognition' 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (IEEE). pp 1926–1933
15. Milan A, Roth S, Schindler K (2014) Continuous energy minimization for multitarget tracking. *IEEE Trans Pattern Anal MachIntell* 36:58–72
16. Roshan Zamir A, Dehghan A, Shah M. GMCP-tracker: Global multi-object tracking using generalised minimum clique graphs. <http://crcv.ucf.edu/projects/GMCP-Tracker/>
17. Son KD (2019) Contribute to son-oh-yeah/Moving-Target-Tracking-with-OpenCV development by creating an account on GitHub.
18. Xiang Y, Alahi A., Savarese S (2015) Learning to track: online multi-object tracking by decision making. In: '2015 IEEE International Conference on computer vision (ICCV)' 2015 IEEE International Conference on computer vision (ICCV), (IEEE). pp 4705–4713
19. Motion-based multiple object tracking - MATLAB & Simulink. <https://www.mathworks.com/help/vision/examples/motion-based-multiple-object-tracking.html>. Accessed Feb 2021
20. Youssef Wael F, Haidar S, Joly P (2016) Classifying deformable and non-deformable video objects. In: 7th International Conference on Imaging for Crime Detection and Prevention (ICDP 2016) (Institution of Engineering and Technology). pp 1–6.
21. Ming-Kuei Hu (1962) Visual pattern recognition by moment invariants. *IRE Trans Inf Theory* 8(2):179–187
22. Vezzani R, Cucchiara R (2010) ViSOR: video surveillance online repository. 2010(2):13
23. Jaffré G, Joly P (2005) Improvement of a temporal video index produced by an object detector. In: Gagalowicz A, Philips W (eds) *Computer Analysis of Images and Patterns*, Springer, Berlin Heidelberg, pp 472–479
24. CAVIAR: Context Aware Vision using Image-based Active Recognition. <https://homepages.inf.ed.ac.uk/rbf/CAVIAR/>. Accessed Feb 2021
25. Blunsden S, Fisher RB (2010) The BEHAVE video dataset: ground truthed video for multi-person behavior classification. 4:1–12
26. Multi-camera pedestrians video – CVLAB. <https://cvlab.epfl.ch/data/data-pom-index-php/>. Accessed Feb 2021
27. Ryoo MS, Chen CC, Aggarwal JK, Roy-Chowdhury A. (2010) An overview of contest on semantic description of human activities (SDHA) 2010. In: Ünay D, Çataltepe Z, Aksoy S (eds) *Recognising patterns in signals, speech, images and videos*. Springer, Berlin, Heidelberg, pp 270–285
28. I-Lids Dataset for AVSS 2007 (2007)
29. Oh S, Hoogs A, Perera A et al (2011) A large-scale benchmark dataset for event recognition in surveillance video. In: CVPR 2011. CVPR 2011, pp 3153–3160
30. Pattern recognition network - MATLAB patternnet. <https://www.mathworks.com/help/deeplearning/ref/patternnet.html>. Accessed Feb 2021

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.