Video Representation and Suspicious Event Detection using Semantic Technologies

Ashish Singh Patel [a,*], Giovanni Merlino [b], Dario Bruneo [b], Antonio Puliafito [b], O.P. Vyas [c] and Muneendra Ojha [a]

[a] *Department of Computer Science and Engineering, DSPM International Institute of Information Technology Naya Raipur, Atal Nagar, Raipur, India*
E-mails: ashish@iiitnr.edu.in, muneendra@iiitnr.edu.in
[b] *Department of Engineering, University of Messina,Messina, Italy*
E-mails: gmerlino@unime.it, dbruneo@unime.it, apuliafito@unime.it
[c] *Department of Information Technology, Indian Institute of Information Technology-Allahabad, Prayagraj, India*
E-mail: dropvyas@gmail.com

Editor: Armin Haller, Australian National University, Australia
Solicited review: Three Anonymous Reviewers

Abstract. Storage and analysis of video surveillance data is a significant challenge, requiring video interpretation and event detection in the relevant context. To perform this task, the low-level features including shape, texture, and color information are extracted and represented in symbolic forms. In this work, a methodology is proposed, which extracts the salient features and properties using machine learning techniques and represent this information as Linked Data using a domain ontology that is explicitly tailored for detection of certain activities. An ontology is also developed to include concepts and properties which may be applicable in the domain of surveillance and its applications. The proposed approach is validated with actual implementation and is thus evaluated by recognizing suspicious activity in an open parking space. The suspicious activity detection is formalized through inference rules and SPARQL queries. Eventually, Semantic Web Technology has proven to be a remarkable toolchain to interpret videos, thus opening novel possibilities for video scene representation, and detection of complex events, without any human involvement. The proposed novel approach can thus have representation of frame-level information of a video in structured representation and perform event detection while reducing storage and enhancing semantically-aided retrieval of video data.

Keywords: Smart City, Data Integration, Data Modeling, Surveillance Video, Ontology, Video Semantics, Video Dataset, Object Tracking

1. Introduction

As surveillance systems are getting affordable, large-scale deployment of such systems are prevalent these days including in parking spaces. Thus, smart parking is becoming an integral part of smart city initiatives, deployment of surveillance systems in such places are resulting in the generation of massive surveillance video data. While, the most crucial task of surveillance systems is to identify unusual activities and events, the detection of these anomalous behavior poses a major challenge in the video data science research. Video data is considered as unstructured data: it is not quantitative but consist of information spread over highly correlated frames. It requires a concrete model to analyze and extract meaningful information. According to a survey [1], the video data being communicated on the global IP network per month is worth 5 million years of watch time. The survey esti-

*Corresponding author. E-mail: ashish@iiitnr.edu.in.

mates that video traffic will rise to 82% of total global data traffic by 2021 from 73% in 2016 and surveillance video traffic will increase by seven times of its current amount by 2021. These figures look imposing when we consider the fact that a majority of surveillance data is not shared over the Internet. Surveillance cameras are capturing an untold amount of video data that is merely stored in archives, remains unanalyzed and finally overwritten after a certain duration. Such waste has prompted an urgent requirement to develop technologies that are not only efficient in storage, retrieval, and processing of video data but are also able to draw meaningful information from the content. However, fetching meaningful knowledge from video data or automatic recognition of events happening in a video poses several technical and domain-specific challenges.

As humans can understand based on cognition, knowledge and experience, information present in a video needs to be extracted, transformed, and linked with domain knowledge to acquire interpretation capabilities through software agents [2]. This requires strong reasoning and analytical support to be able to detect an event, especially of anomalous nature and bridge the semantic gap between machine interpretation and human perception. Low-level features (such as shape, size, color, etc.) extracted using video processing are not enough to generate the information required for video scene understanding. Those complex events that are rare are hard to train using supervised machine learning due to lack of sufficient training examples and massive computational capability requirements. Formulating an event requires domain and context knowledge, however, most of the present techniques lack the ability to bridge the semantic gap between low-level and high-level features, do not support data integration as well. In such scenarios, machine learning algorithms may not be applicable due to the limited number of training examples and lack of formalism [3].

Semantic Web Technology, is, therefore leveraged to fill this gap by creating domain ontology, which is effective in representing high-level semantics present in the video. Semantic Web Technologies [4] facilitate data integration along with rule-based reasoning using Semantic Web Rule Language (SWRL) [5] and SPARQL, achieving widespread interoperability in a predefined domain by using same ontology. Ontology supports Description Logics (DL), which can be used to perform spatial and temporal reasoning [6]. The semantic information present in the video is represented in Resource Description Framework (RDF) [4] format, which is machine-readable triplet and describes the relational information in *"subject-object-predicate"* form. RDF statements are constructed using concepts defined in vocabularies written in Web Ontology Language (OWL) ontologies.

Our approach extracts the frame-level parameters using machine learning techniques to generate a higher-level semantics for detecting unusual and suspicious events from the surveillance video data. An ontology is developed which represents the object(s) and interactions between the object(s) present in video frame. The relationships between the objects in an image are generated by creating SWRL-based rules, and events are formulated using SPARQL queries. Our review of existing literature suggests that our approach, representing frame-level information of a video in the structured machine-interpretable format, while enabling event detection by means of Semantic Web Technology is not yet explored, and is thus proposed in this paper. The key contribution of our work are as follows:

- Frame level representation of video scene in RDF, which saves a lot of storage space, facilitates reasoning and efficient information retrieval.
- Deriving relationships between the objects in an image using SWRL, i. e., reasoning over video.
- Activity detection using SPARQL: once all the information is represented in RDF graphs, activity / events can be recognised and retrieved by formulating SPARQL queries.
- Opens up new opportunities for video data analysis research, where training examples are fewer or resources are computationally costly.
- Accuracy (relationship detection) is very good and performance is high.
- A video dataset, which consists of six different trimmed localized activities in smart parking scenarios, totaling 92 videos.
- A novel approach for object tracking is also proposed here, based on SWRL and Description Logics.

The results obtained using the proposed approach are promising, as the proposed methodology efficiently represents the frame-level information in RDF, then performs SWRL reasoning to extract spatio-temporal relationships between objects. The represented semantic information is retrieved to identify various scenar-

ios and use-cases (demonstrated by recognizing suspicious events in smart parking scenarios).

The remainder of this paper is organized as follows: Section 2 presents the related work. Section 3 demonstrates the proposed approach for representing the events in the smart parking domain. Section 4 shows the results of the proposed work, describing parameters for evaluation, and outlines current issues and limitations. Conclusions about the contribution of the work are drawn in Section 5 while Section 6 covers future work prospects.

2. Related Work

In this section, we first review previous work on the extraction of high-level semantic information present in the video. We then briefly summarize the existing work on utilizing ontology-based approaches.

2.1. Non semantic approaches

The non-semantic approach predominantly includes machine learning and feature based methodologies. You et al. [7] proposed a semantic framework for video genre classification based on the Hidden Markov Model (HMM) and Gaussian mixture model. The framework utilized the visual features by generating a semantic feature computation approach along with analysis on the relationship between such features and video semantics. The approach was complex and depended highly on the way the video features are computed. Si et al. [8] proposed an unsupervised learning approach for event detection by using the predefined set of atomic actions and relations (a combination of atomic actions) like touch, bend, sit, etc. These successive events were modeled in the learned grammar and made context-sensitive. The learned grammar could be used to improve the noisy bottom-up detection of atomic actions. It was also proposed to be used to infer the semantics of the scene. Zhu et al. [9] analyzed that low-level features alone, often are less significant for naive users, preferred to recognize using high-level semantic information (concepts). The shot was segmented using a color histogram. The identification of textual data in the video was performed in two ways; the first one involved the extraction of embedded text in the video like scores and another was the detection of text, which was already present in the scene. Text regions are recognized using edge detection techniques. Camera motion was identified using the mutual relationship between motion vectors in the P frame. Furthermore, the audio level was used to detect events with high noise, such as whistling, etc. Data were transformed to fit for association rule mining in item-set and temporal distance between two item-set, i.e., the video event was calculated. A deep hierarchical context model for event recognition was proposed by Wang et al. [10] which is effective in low image resolution and intra-class variation. The model could simultaneously learn and integrate context at all three levels, thus utilizes the context information efficiently. Context features (neighborhood of the event) used in the model to generate mid-level representations, and then combine the context information for recognizing events. The approach was evaluated on benchmarks dataset (VIRAT 1.0 ground dataset and UT-Interaction dataset), performed excellently.

Patterns determined through the machine-learning techniques applied to various feature descriptors of the video are very crucial for event analysis. Xie et al. [11] proposed a method for event detection, defined event by their dissimilarity among the discovered patterns, event description, event-modeling components, and current event mining systems. They have defined an event identification framework by identifying five W's and 1H (when, what, who, where, why, how). Also, they have classified metadata as intrinsic and extrinsic which contains event-related information. The segmentation involved identifying the part of the video where it happened (time, space, and duration), recognition involved identification of one or more W's described earlier, verification required test of the specific property, annotation, and adding labels to the data. The task of discovery is about finding the event without having prior knowledge of semantics. Hamid et al. [12] proposed an unsupervised activity analysis using n-grams suffix trees to mine motion patterns at different temporal scales. Activities were represented in the form of suffix trees. The class of the action was identified by mapping it to the problem of finding a maximal clique in a graph. The event is detected automatically by extracting the interaction of a person with the object using the Gaussian Mixture Model. An anomaly was detected if an activity does not appear in any of the sub-sequence. Baradel et al. [13] proposed a method for human activity detection from RGB data, without relying on pose information. They have defined a glimpse as a group of interest points relevant to classified activities. Due to the high correlation of events, visual point tracking is required, resulting in the collection of glimpses. However, the tracks whose

location is not continuous in the spatial and temporal domain can lead to a change in semantic information, being a significant challenge. This problem is solved by selecting a local as well as distributed representation of glimpse points based on the sequential attention model and tracking the set of glimpse points by integrating them in final recognition. Liao et al. [14] proposed a novel framework for analyzing the surveillance video and recognizing the event. In the first step, an object was detected using Convolution Neural Network (CNN), then the owners of the objects were identified and monitored in real-time. If any object was moved, it was verified whether the person who moved the object was the owner or not. In case the person who moved was not the owner, the scene is further analyzed to differentiate between the stealing and moving away. They have also proposed a dataset consisting of such scenarios to evaluate the proposed approach. The approach was compared with the state of the art results on existing benchmarks dataset related to luggage detection and management.

Snoek et al. [15] listed the issues and challenges in concept-based video retrieval. They have also emphasized the semantic-gap, thus come up with concept-based video search by primarily focusing on methods of information retrieval, machine learning, human computer interaction and computer vision. Also, they have explored the task of concept detection by fusing the feature and information from classifiers to model the relations along with tools and techniques for benchmarking as performed in the NIST TRECVID benchmark.

Cheng et al. [16] conducted a study by using TRECVID 2015 dataset to understand the importance of features that are more relevant for video hyper-linking like meta-data, subtitle, content-based features including (audio and visual) along with the context of the video. However, the major improvements in search quality resulted from textual features rather than content-based features.

Shen et al. [17] proposed a method for event detection by using a subspace selection technique that can identify various classes, also preserve intramodal geometry of samples within a class. The approach was divided into two major tasks; the first one involved the extraction of video features from the video segments while in second task Modality Mixture Projections (MMP) were used to generate the signature of video. The MMP is a dimensionality reduction technique based on linear discriminant analysis which preserves the geometric projections. The approach was demonstrated on soccer video and TRECVID news dataset. Chen et. al. [18] presented an approach to generate captions of the video scene for video understanding. They have isolated two significant challenges for the task of video captioning (broad domain and multimodal information) as compared to video indexing and retrieval. Thus, they have divided the task of captioning in two tasks; the first task was the latent topic generation and the second one was topic-guided caption generation. The topic generation task predicts the topic of the video based on an unsupervised learning method built using video contents and captions in the video. This reduces the overall complexity by narrowing the topics and cover the various modalities by topics. They have also proposed a topic-guided ensemble framework by correlated the two tasks to generate more precise video captions. But their approach cannot be employed for video event detection, however useful for video understanding.

Deep learning can effectively model human cognition and behavior, thus lead to bridge the semantic gap between machine-level interpretation and human understanding. However, it requires massive labeled data and computationally expensive, often suffers due to lack of training data. Caruccio et al. [19] proposed a layered knowledge representation framework for automatic video detection consisting of environment layer (including capturing devices like camera and sensors), frame layer (analyzes frame sequences), elements of context representations, general context descriptors and action representations. The activity was detected in the framework by forming logic based visual representations of the scenarios while combining a set of small actions. The approach was complex and very specific to the use-case. It could not completely represent the information present in a scene. Gan et al. [20] proposed a CNN based approach named Deep Event Network (DevNet) for Event detection. DevNet took key frames of the video as input and construct a saliency map by pack-passing, which was used to find the key frames. Events were formulated as a semantic abstraction of video rather than just concepts. Event of "Town-hall-meeting" was formulated by combining objects, a scene, actions, and acoustics. The objects may include person, podium, scene of a conference room while actions include talking, meeting and speech, clapping as acoustic concepts. DevNet localized key shreds of evidence and detected high-level events as well. The approach was evaluated for event detection and evidence recounting on TRECVID 2014 and MEDTest dataset and achieved promising

results. He et al. [21] proposed a multi-modal fusion model, which exploits spatial-temporal modeling for human activity recognition. The model named StNet (Spatial-temporal Network) based on ResNet for enhanced modeling of spatial and temporal characteristics leading to video understanding. Multi-modal information contained in the video was integrated using a temporal Xception network (iTXN). The other framework, such as Inception, Resnet-V2, ResNeXt and SENet, could also be used instead of ResNet based architecture. The results are promising due to the exploitation of multi-modal information. Furthermore, a model for spatio-temporal representation based on the residual network named pseudo-3D residual network (P3D) is proposed by Qui et al. [22]. As 3D ConvNet development from scratch requires a significant amount of computations. Various types of bottleneck building blocks were constructed in a residual simulating *3x3x3* convolutions from *1x3x3* convolutional filters equivalent to 2D-CNN and *3x1x1* convolution for creating temporal dimension on particular feature maps with time. A novel architecture named Pseudo-3D Residual Net (P3D ResNet) based on ResNet but having the different placement of blocks, having a philosophy that by increasing structural variation on the deeper layer will make the network more robust. P3D ResNet demonstrated to perform better by 5.3% and 1.8% on the classification of Sports-1M video dataset than 3D CNN and 2D CNN.

2.2. Semantic Approaches

Features extracted from multimedia contents are represented in symbolic or numerical form. The knowledge inferred from these features, is represented in terms of concepts, properties, sub-concepts, and their respective relationships can be individually identified and described. This knowledge can be interlinked with the known concepts for data integration and thus facilitates multi-modal analysis. Ram et. al. [23] proposed the Video Event Representation Language (VERL), a formal language for describing an ontology of events using objects and state. They described an event as a change of state of an object. Events in a state may lead to other state, but the scope of the ontology was limited and cannot be applied to other domains and concepts. Moreover, it does not follow OWL-DL syntax. Juan et al. [24] presented an ontology which can represent high-level semantic features and knowledge using a hierarchical framework for video event and annotations. However, the ontology is not integrated with other standards like MPEG7 and does not include domain related concepts. Bermejo et al. [25] discussed an ontology-based approach which detects complex events and abnormal situation by integrating the sensor data (e.g. acceleration, speed, distance, lane change, etc.). The integrated information was used to aid decision support system for traffic management. Fan et al. [26] proposed to incorporate concept ontology for hierarchical video classification. More specific semantics were represented in the deeper layers of the hierarchy. Concept ontology provided contextual and logical relationships. As a single ontology may not meet all requirements, multiple concept ontologies for video concept organization were needed. The specific semantics were represented in deeper layers of the hierarchy. Duong et al.[27] proposed an ontology-based approach to describe the content and allow sharing with a consensus-based algorithm for reconciliation of conflicts. The visual features were extracted using MPEG7 visual descriptors, which were then used to generate video-level summary. It was, however, not suitable for representing the frame-level information. Elleuch et al. [28] proposed a fuzzy ontology to enhance concept detection by using context information about concepts based on visual modality. The context modeling was performed in three steps, i.e., semantic knowledge representation, semantic concept categorization, and refinement. A context ontology was constructed first to model the relationships between concepts and then a deductive engine was built on fuzzy rules and optimized based on genetic algorithms. Grassie et al. [29] proposed a semantic model which enables annotation to create structured knowledge at multiple levels of granularity and complexity. Ontologies were built to support linking to LOD cloud at data level. The high level interpretation of the video was limited to brief textual comments and tags explaining the whole video. In most of the cases, videos were not labeled or annotated to encode all relevant information with tags, as their interpretation were often confusing. One use case was used to demonstrate the applicability of semantic representation and linking it to DBpedia [30] resource by using annotation tools. Patricio et al. [31] proposed a framework to construct a symbolic model which exploited contextual information and tracking data in a scene. Knowledge representation and reasoning was performed using OWL and DL. An ontology was developed based on the DL, which defined the concepts, roles, and relations, giving the basic idea of the domain. The framework consisted of a general tracking layer which generated trajec-

tory and context layer representing context and knowledge extracted from the scene. Domain ontologies had to be created manually or semi-automatically needing considerable effort and domain knowledge. Xu et al. [32] proposed a video structure description ontology, which parsed the video into text information using spatial and temporal segmentation, feature extraction, object recognition, and semantic web technologies. The extracted video content was represented in RDF using domain-specific ontology created for traffic domain surveillance videos. However, the data mining and inference rule generation for various events are still unexplored. Vallet et. al. [33] proposed a content retrieval method using ontological knowledge (a semantic distance of the concepts) considering user preferences. Ontologies provided a formal framework for representing semantic definition and facilitated the generation of new knowledge-base through inference rules. The model was deficient, in that it only captured long-term preferences, without considering short term preferences. Naphade et al. [34] constructed 834 semantic concepts based on the properties of multimedia content, but many terms were not suitable for automated tagging. LSCOM produced an ontology consisting 1000 concepts of broadcast news domain. Apart from ontology design, binary relations to hold higher relations (rule) by relating target concepts and also includes explicit rules were created. Hauptmann et al. [35] proposed high-level semantics by providing descriptors of visual content and experimentally demonstrated that video retrieval improves by increasing the number of semantic concepts, used concepts from MediaMill and LSCOM to evaluate TRECVID 2005 collection. Video retrieval efficiency was shown to be proportional to the relevance of concepts. Mutual information was used to determine the helpfulness of concepts. Mahmood et al. [36] proposed a method to extract the semantic content from the sports video. They highlighted the variety aspect of the data, which consisted of semi-structured and unstructured format. The proposed model was based on speech processing, Natural Language Processing (NLP), and Semantic Web Technologies to predict the best combination of players for next $'n'$ minutes. Text from the video was extracted and then converted to RDF using semantic web technologies and NLP. Best performance for the next few minutes was identified based on factors like weather and their past performance in a match but no details were provided on methodology and evaluation of the proposed approach. Tani et al. [37] proposed a rule based approach using SWRL for event detection, but handled only spatial events like walking and running. They could not detect temporal events which happened over the course of time. Additionally, the proposed methodology could not represent frame-level interactions between the objects.

According to Sikos [38], video contents are challenging to parse due to lack of semantics in software systems. Most of the annotation formats provide metadata about the title, creator, time, comments, and lyrics in XML format. However, this information is not machine interpretable, making it unsuitable for access, sharing and reuse. Existing vocabularies such as Dublin Core and Schema.org only provide de-facto standard for annotating video objects, while semantic interoperability requires explicit descriptors to represent information, should be unique and defined in entity. Sikos [6] proposed a DL based knowledge representation, which can be used for multimedia analysis, event detection, and interpretation of high-level media descriptors. High-level video-semantics requires comprehensive reasoning along with suitable ontologies. Most of the existing ontology do not supports all constructs of DL which could efficiently model complex reasoning using atomic concepts by implying assertion, conjunction, disjunction, etc. and follow SROIQ DL like role restrictions, concepts, etc. Sikos [39] demonstrated that ontologies for representing video events require spatial and temporal features including specific motion events in video scenes.

Smart parking is an integral part of smart city initiative. Denizens of the city face massive problem in finding a proper parking space. As surveillance systems are getting affordable, large-scale deployment of such systems in parking spaces is resulting in generation of massive surveillance video data. This data is beneficial in order to analyze the trajectory, driver behavior of the vehicles along with safety and security of the car [40]. Most of the work done in literature is focused towards assisting a driver to the parking system, i.e., identifying the nearest appropriate parking location. However, parking lot itself needs to be monitored to ensure the safety and security of the vehicle when parked inside the parking lot [41][42]. In this paper, we demonstrate the applicability of our methodology by identifying unusual activities to ensure safety of the parked vehicle.

3. Proposed Work

In this paper, we propose a method to detect suspicious events occurring inside the parking lot to mon-

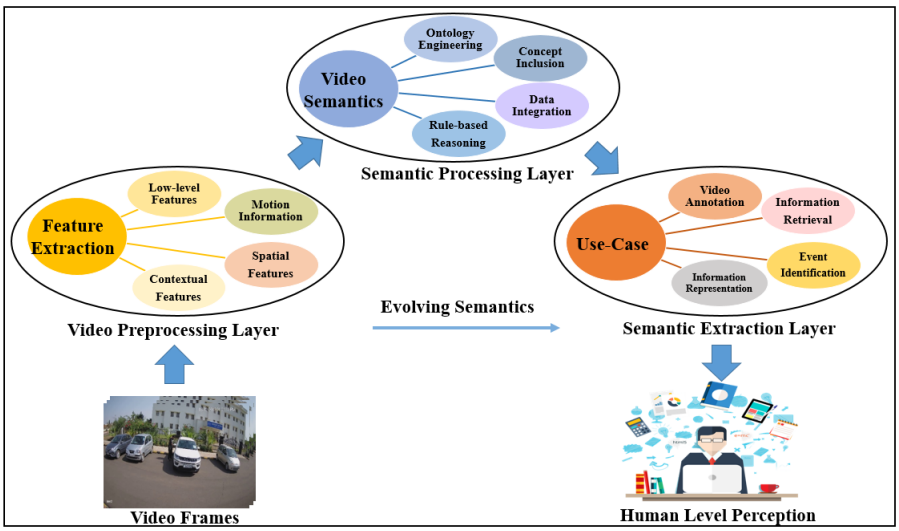


Fig. 1. Workflow and Framework of the Approach

itor the safety of parked vehicles. The description logic expressivity of this ontology is ALHI+(D). Inference rules are created to perform the reasoning over the video data. We then formulate the SPARQL and DL queries to extract high-level semantic information. This formulation further requires the extraction of low-level features and representation using video ontology to create high-level semantics, matching human-level perception. The overall workflow is shown in Figure 1. The workflow contains three processing layers: Video Preprocessing Layer, Semantic Processing Layer and Semantic Extraction Layer. The video preprocessing layer performs the feature extraction through a series of steps involving low-level features extraction, motion information, spatial features, and contextual information. These features are then passed onto the semantic processing layer. The semantic processing layer then generates the structured semantic information by processing the video which is represented in machine-readable format after performing ontology engineering, data integration, concept identification, and rule-based reasoning. Finally, the semantic extraction layer transforms the processed semantic information for various use-cases.

3.1. Definitions

A set of terms are defined which are found in literature and used in a standard context:

- ***Scene*** - A sequence of video frames having same background and objects.
- ***Activity*** [43] - Something which captures user attention, consists of interaction between multiple objects, can be usual or abnormal.
- ***Events*** [23] -Activity which captures user attention, requires modeling of temporal and multimodal characteristics. It involves an understanding of object behaviors and recognition of motion patterns.
- ***Sub-event*** - A uniquely identifiable activity (such as someone sitting on the car) of a complex event (such as a person damaging the car).
- ***Suspicious Activity***-An activity that is rare and potentially dangerous and may lead to unfavorable consequences [43].
- ***Ontology*** [4] - An ontology is a machine-readable semantic description of data. It also documents a particular domain to develop a common understanding of concepts.
- ***Concepts*** - A concept can be an object property or data property or a class defined in the ontology.

3.2. Events and Concepts

Concepts and sub-events are formulated to represent the information present in a video scene. Events that need to be detected are characterized by combining sub-events, properties, and relations among them. Frame-level information is used to construct the sub-events having a temporal attribute, which in turn is used to identify complex events. The scene is represented by developing an ontology that represents an

object along with its position in every frame. For each frame, a date and timestamp attribute is associated with every identified object.

3.3. Feature Extraction and Selection

Low-level features are extracted using edge detection, color detection, hough line detection, contour detection. Since a frame consists of multiple objects, the properties of these objects such as length, width, dominant color, type, location, and time-stamp in a video are also considered as low-level features. We used YOLO [44] for extracting type and location of the object. It returns a bounding box of the object with which the size and location of the object can also be calculated. Mid-level features consist of class hierarchy of the objects and relationships between the objects. Since not every feature is relevant to every event scenario, the low-level features are further picked on the basis of ontology and domain. A detailed workflow is shown in Figure 2 wherein part (a), depicts the extraction of low-level features, ontology development, and frame representation, while part (b) shows the generation of relationship and detection of an activity. The constructed ontology represents the spatio-temporal relations between objects in a video scene. The features of a frame are represented using ontology as data properties. In contrast to data properties, the object properties are referred to as mid-level features. As shown in Figure 2 (a), each object of the frame is represented as an individual. In Fig. 2 (b), the mid-level features, i.e., the relations between the objects present in same frame and temporal frames are inferred using SWRL rules between the individuals. Relation between objects *isInTheVicinityOf* is shown using same color line. The green-colored connecting line shows the existence of *overlaps* relation between *Car-1* and *Car-2*. The *isInTheVicinityOf* relation exists between *Person-1* and *Person-2* shown by blue-colored connecting lines while the same relation between *Person-1* and *Car-1* is shown through orange-colored connecting lines. In final step, abnormal events are formulated by reasoning over the behavior (reason for suspicion) and using SPARQL from RDF database.

3.4. Parking Lot Ontology

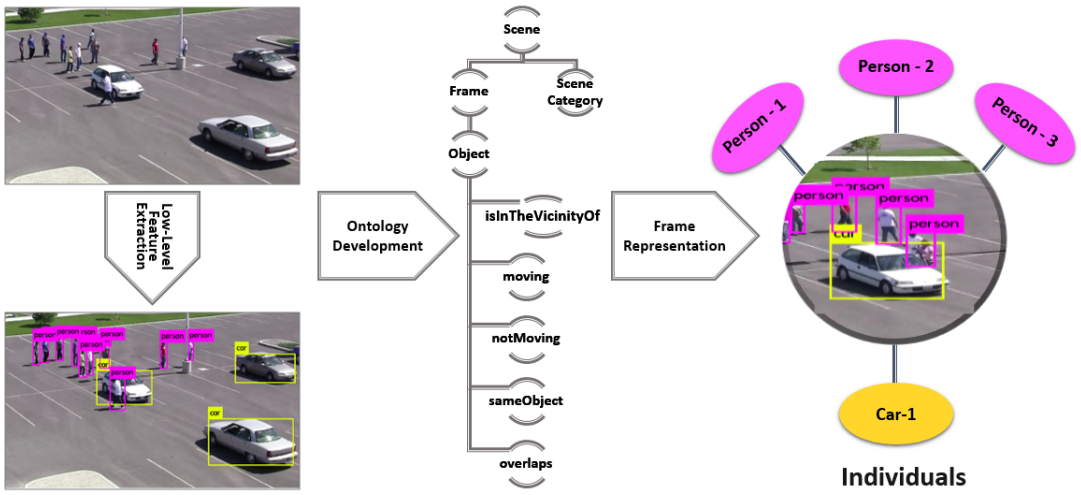
We develop an ontology that includes the concepts and parameters of the parking domain. The extracted data from the video is represented in the RDF format using proposed domain ontology. The ontology[1] follows OWL 2 DL constructs. OWL 2 adds new functionality with respect to OWL 1 . Some of the new features are syntactic (e.g., disjoint union of classes) while others offer new expressibility, including richer datatypes, data ranges, qualified cardinality restrictions, asymmetric, reflexive, and disjoint properties along with enhanced annotation capabilities. The pellet incremental reasoner is used to test the consistency of the proposed ontology. The ontology contains classes to represent frame level information in a video scene. Figure 3 shows an ontograph of the video frames represented in RDF format using constructed ontology. The proposed work is carried out in protege tool [45].

The top four boxes in Figure 3 represents the classes (*thing, scene, frame, and object*) followed by different boxes representing individuals of the class object. The relationships between classes, subclasses, individuals, and object properties are represented using respective colored arcs. Each individual represents one unique object in a frame number. Individual name *Person 1-1* is first object (type person) of first frame of a video scene. Green-colored arcs represent the *hasIndividual* relationship between class and the individual. Arcs with blue color represent the *hasSubclass* relationship between a class and a subclass. Red-colored arc represents object property *isInTheVicinityOf* between two individuals. Black-colored arcs represent object property *sameObject* between two individuals. However, owl:sameAs essentially means that two individual have same properties and instance, but in this case data properties and object properties of the individual are different, also the object belong to different frame. Therefore, *sameObject* is defined in our ontology to be used instead of owl:sameAs.

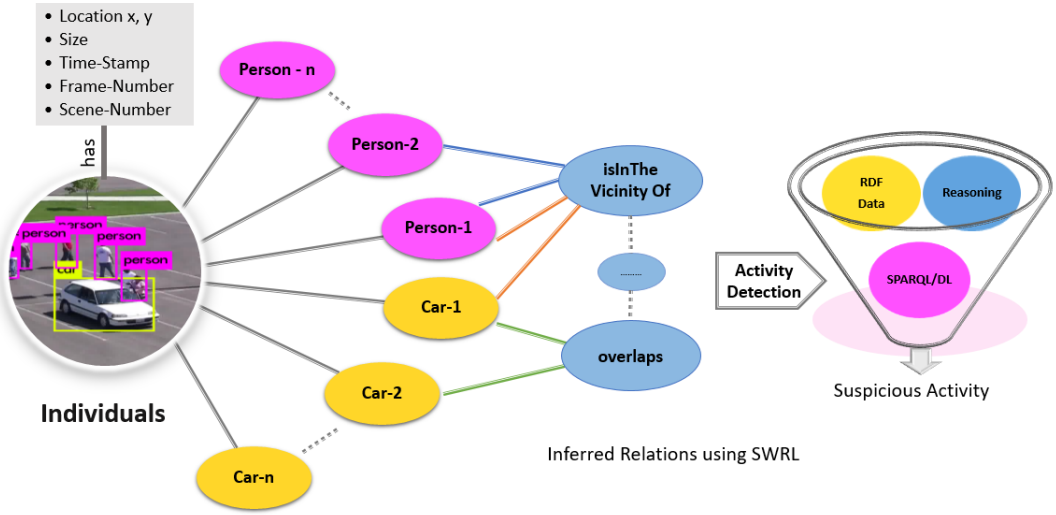
The data properties used are

- *ObjectLocationLT* - Coordinate of the center of the object along x axis.
- *ObjectLocationYC* - Coordinate of the center of the object along y axis.
- *ObjectType* - Type of the object like car, person.
- *ObjectWidth* - Width of the object.
- *ObjectLength* - Length of the object.
- *hasTime* - Date Time Stamp upto milliseconds of a video frame.

[1] https://github.com/aspdr/Video-Representation-and-Suspicious-Event-Detection-using-Semantic-Technologies/blob/master/smart-parking-sparql.owl



(a) Feature Extraction, Ontology development and Frame Representation



(b) Relationship generation and Activity detection

Fig. 2. Detailed Workflow of the Proposed Approach

- *hasSceneNumber* - Scene number of the video.
- *Frame-Number* - Frame number of the scene.
- *vicinityDuration* - Duration for which two objects are in vicinity.

The object properties used are

- *isInTheVicinityOf* - This relation holds between two objects when they are nearby.
- *sameObject* - holds when two objects are same in different frame.
- *moving* - holds when object is moving.
- *notMoving* - holds when object in not moving.
- *overlaps* - holds when two objects overlaps.

3.5. Rule-based Reasoning for Event Detection

We perform reasoning and scene interpretation using inference rules. Based on domain knowledge, the rules are formulated using DL to detect suspicious activity near a parked car in a parking lot using SWRL [5]. The rules are described with car as a reference, modified accordingly for other type of vehicles. SWRL supports new inferences based on the reasoning over

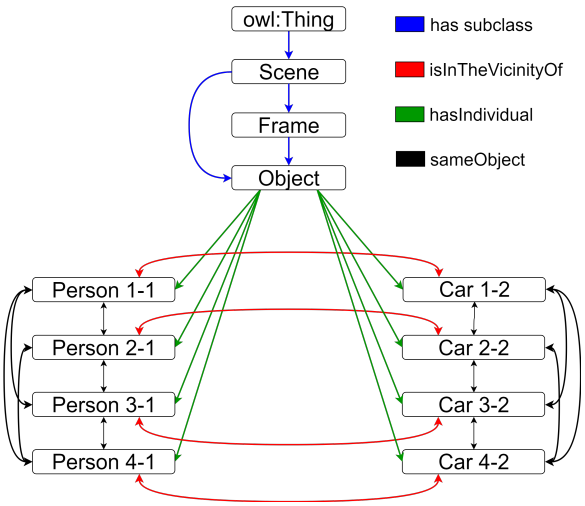


Fig. 3. Video information extracted and represented in RDF using Ontology

existing classes, objects, and data properties. SWRL is expressed in the form of an implication between an antecedent (body) and consequent (head). The intended meaning can be read as: whenever the conditions specified in the antecedent hold, then the conditions specified in the consequent must also hold. All the statements on the left side of the implication operator (⇒) are known as antecedent (body) and on the right side are known as consequent. Antecedent consists of statements connected through conjunction operator (∧). Antecedent can also include built-in's for e.g. *swrlb:add(), swrlb:subtract()* etc., for mathematical operation. Variables used in the rules are described in Table 1. The threshold values for *yaTh, ysTh, xaTh and xsTh* are obtained using several independent experiments and were found to be most suitable for the task. In this work, *yaTh, ysTh, xaTh and xsTh* were set to 100, 170, 215 and 100 respectively.

3.5.1. *isInTheVicinityOf*

Rule 1 sets the object property *isInTheVicinityOf* between two individuals if they are nearby each other.

Table 1
Description of the variables

Variable Name	Description
o1, o2	Refers to an individual of type *object*
v1, v2	Refers to frame number of *o1* and *o2*
x1, x2	Refers to the horizontal pixel of *o1* and *o2*
y1, y2	Refers to the vertical pixel *o1* and *o2*
t1, t2	Refers to the type of objects *o1* and *o2*
xs, ys	Value after subtracting threshold from *x1* and *y1*
xa, ya	Refers to the threshold added to *x1* and *y1*
yaTh, ysTh	Threshold values for calculating *ya* and *ys*
xaTh, xsTh	Threshold values for calculating *xa* and *xs*
w1, w2	Refers to width of *o1* and *o2*
l1, l2	Refers to length of *o1* and *o2*
ws, ls	Width and length of *o1* subtracted from *o2*

Rule 1

$$Object(?o1) \wedge Object(?o2) \wedge Frame\text{-}number(?o1,?v1) \wedge Frame\text{-}number(?o2,?v2) \wedge swrlb{:}equal(?v1,?v2) \wedge objectLocationLT(?o1,?x1) \wedge objectLocationLT(?o2,?x2) \wedge swrlb{:}subtract(?xs,?x1,xsTh) \wedge swrlb{:}add(?xa,?x1,xaTh) \wedge swrlb{:}lessThan(?x2,?xa) \wedge swrlb{:}greaterThan(?x2,?xs) \wedge objectLocationYC(?o2,?y2) \wedge objectLocationYC(?o1,?y1) \wedge swrlb{:}add(?ya,?y1,yaTh) \wedge swrlb{:}subtract(?ys,?y1,ysTh) \wedge swrlb{:}greaterThan(?y2,?ys) \wedge swrlb{:}lessThan(?y2,?ya) \wedge objectType(?o1,?t1) \wedge objectType(?o2,?t2) \wedge swrlb{:}notEqual(?t1,?t2) \Rightarrow isInTheVicinityOf(?o1,?o2).$$

- *Object (?o1)* - Represents a variable o1 which is of type Object(Class in the ontology, it can be of type individual, class, data property or an object property).
- *Frame-number (?o1,?v1)* - Evaluates to true whenever the data property value (*Frame-Number*) of object *o1* is stored in variable *v1*.
- *swrlb:equal (?v1,?v2), swrlb:notEqual (?v1,?v2)* - SWRL built-in function to compare values if they are equal or unequal.
- *swrlb:add(?v1,?v2,?v3), swrlb:subtract(?v1,?v2, ?v3)* - SWRL built-in function to perform mathematical operation and store the result in v1.

- *swrlb:greaterThan (?y2,?ys), swrlb:lessThan (?y2, ?ys)*- Compares to values y2 and ys and returns true when y2 satisfies the condition.
- *objectLocationYC (?o1,?y1), objectLocationLT (?o1,?x1)* - sets the y coordinate and x coordinate of object o1 in variable y1 and x1.
- *ObjectType (?o1,?t1)* - Sets the class of the object o1 in t1.
- *isInTheVicinityOf (?o1,?o2)* - Sets the *isInTheVicinityOf* property between object o1 and o2, states objects are in vicinity of each other when antecedent on left side of the above rule evaluates to true.

3.5.2. sameObject

Rule 2 sets the object property *sameObject*. Two objects are said to be *sameObject* when objects present in subsequent frames have the same shape, size, same dominant color, and are locationally nearby. However, they have been identified to be two different individuals belonging to the same object class (e.g. two persons or two cars).

Rule 2

Object (?o1) ∧ Object (?o2) ∧ Frame-number (?o2, ?v2) ∧ Frame-number (?o1, ?v) ∧ swrlb:add (?v3, ?v, 1) ∧ swrlb:equal(?v3, ?v2) ∧ objectLocationLT(?o1, ?x1) ∧ objectLocationLT (?o2, ?x2) ∧ swrlb:subtract (?xs, ?x1, xsTh) ∧ swrlb:add (?xa, ?x1, xaTh) ∧ swrlb:greaterThan (?x2, ?xs) ∧ swrlb:lessThan (?x2, ?xa) ∧ objectType (?o1, ?t) ∧ objectType (?o2, ?t2) ∧ swrlb:equal (?t, ?t2) ∧ objectLocationYC (?o1, ?y1) ∧ objectLocationYC (?o2,?y2) ∧ swrlb:add (?ya, ?y1, yaTh) ∧ swrlb:subtract (?ys, ?y1, ysTh) ∧ swrlb:greaterThan (?y2, ?ys) ∧ swrlb:lessThan (?y2, ?ya) ⇒ sameObject(?o1, ?o2)

- *sameObject (?o1,?o2)* - Sets the *sameObject* relation between object *o1* and *o2*, states that both object are same when antecedent on left side of the above rule evaluates to true.
- Rest of the antecedents used, are already defined earlier in Section 3.5.1.

3.5.3. moving

Rule 3 sets the object property *moving* when object present in subsequent frames have the same shape, size, same dominant color, and are locationally nearby, but is continuously changing over the certain number of frames. However, they have been identified to be two different individuals belonging to the same object class (e.g. two persons or two cars).

Rule 3

Object (?o1) ∧ Object (?o2) ∧ Frame-number (?o2,?v2) ∧ Frame-number (?o1,?v) ∧ swrlb:add (?v3, ?v, 1) ∧ swrlb:equal (?v3,?v2) ∧ objectLocationLT (?o1,?x1) ∧ objectLocationLT (?o2,?x2) ∧ swrlb:subtract (?xs,?x1,xsTh) ∧ swrlb:add (?xa, ?x1, xaTh) ∧ swrlb:greaterThan (?x2, ?xs) ∧ swrlb:lessThan (?x2,?xa) ∧ objectType (?o1,?t) ∧ objectType (?o2,?t2) ∧ swrlb:equal (?t,?t2) ∧ objectLocationYC (?o1,?y1) ∧ objectLocationYC (?o2,?y2) ∧ swrlb:add (?ya,?y1,yaTh) ∧ swrlb:subtract (?ys,?y1,ysTh) ∧ swrlb:greaterThan (?y2,?ys) ∧ swrlb:lessThan (?y2,?ya) ∧ swrlb:notEqual (?y1,?y2) ∧ swrlb:notEqual (?x1,?x2) ⇒ moving(?o1,?o2)

- *moving (?o1,?o2)* - Sets the *moving* relation between object *o1* and *o2*, states that both object are moving when antecedent on left side of the above rule evaluates to true.
- Rest of the antecedents used, are already defined earlier in Section 3.5.1.

3.5.4. notMoving

Rule 4 sets the object property *notMoving* when object present in subsequent frames have the same shape, size, same dominant color, and are locationally same, but location is fixed over certain frames. However, they have been identified to be two different individuals belonging to the same object class (e.g. two persons or two cars).

Rule 4

Object (?o1) ∧Object (?o2) ∧ Frame-number (?o2,?v2) ∧ Frame-number (?o1, ?v) ∧ swrlb:add (?v3,?v,1) ∧ swrlb:equal (?v3, ?v2) ∧ objectLocationLT (?o1, ?x1) ∧ objectLocationLT (?o2,?x2) ∧ swrlb:subtract(?xs,?x1,xsTh) ∧ swrlb:add (?xa,?x1,xaTh) ∧ swrlb:greaterThan (?x2,?xs) ∧ swrlb:lessThan (?x2, ?xa) ∧ object-Type (?o1,?t) ∧ objectType (?o2,?t2) ∧ swrlb:equal (?t,?t2) ∧ objectLocationYC (?o1,?y1) ∧ objectLocationYC (?o2,?y2) ∧ swrlb:add (?ya, ?y1, yaTh) ∧ swrlb:subtract (?ys,?y1,ysTh) ∧ swrlb:greaterThan (?y2,?ys) ∧ swrlb:lessThan (?y2,?ya) ∧ swrlb:Equa l(?y1,?y2) ∧swrlb:Equal (?x1,?x2) ⇒ not-Moving (?o1,?o2)

- *notMoving (?o1, ?o2)* - Sets the *notMoving* relation between object *o1* and *o2*, states that both object are *notMoving* when antecedent on left side of the above rule evaluates to true.
- Rest of the antecedents used, are already defined earlier in Section 3.5.1.

3.5.5. overlaps

Rule 5 sets the object property *overlaps* when the boundaries (bounding box of object) of the two different object are overlapping in current frame.

Rule 5

Object (?o1) ∧ Object(?o2) ∧ Frame-number (?o2, ?v2) ∧ Frame-number(?o1, ?v1) ∧ ObjectWidth (?o1,?w1) ∧ Ob-jectWidth (?o2,?w2)∧ ObjectLength (?o1,?l1) ∧ ObjectLength (?o2,?l2) ∧ objectLo-cationLT (?o1,?x1) ∧ objectLocationLT (?o2,?x2) ∧ swrlb:add(?wa,?w2,?x1) ∧ swrlb:subtract (?ws,?x1,?w2) ∧ swrlb:add (?la,?l2,?y2) ∧ swrlb:subtract (?ls,?y2,?y2) ∧ swrlb:equal (?v1,?v2) ∧ swrlb:greaterThan (?x1,?ws) ∧ swrlb:lessThan(?x1,?ws) ∧ swrlb:greaterThan(?y1,?ls) ∧ swrlb:lessThan (?y1,?ls) ⇒ overlaps(?o1,?o2)

- *ObjectWidth (?o1,?w1)* - Evaluates to true whenever the data property value (ObjetWidth) of object o1 is not null.
- *ObjectLength (?o1,?l1)* - Evaluates to true when data property value (ObjetLength) of object o1 is not null.
- *overlaps (?o1,?o2)* - Sets the *overlaps* relation between object *o1* and *o2*, states that both object are *overlaps* when antecedent on left side of the above rule evaluates to true.
- Rest of the antecedents used, are already defined earlier in Section 3.5.1.

3.6. Suspicious Activity Detection around a parkied car using SPARQL query

SPARQL query is used to identify suspicious activities. It queries the RDF data using existing object properties and data properties. It can match against graph patterns and has a rich set of operators and functions on numbers, strings, date/time, and terms. A SPARQL query is triggered to extract meaningful information about occurrence of the event with time. A SPARQL query is created on the basis of logical description used to define an event. The queries are described with car as a reference, modified accordingly for other type of vehicles. For example, the activity of loitering around a parked car can be suspected by identifying the presence of a person in the vicinity of a car for more than certain duration. Figure 4 provides the complete SPARQL query created to identify loitering around the vehicle.

In the first part of the query given in Figure 4, prefixes are defined. Prefix owl defines the schema of the OWL-DL construct. The rdfs define the resource description format schema, xsd defines the XML schema. The time defines the time ontology. The date defines the schema of date type and built-inâĂŹs. The *select* statement in the query returns four distinct instances of type object at two specified intervals. Instances *inst1*, *inst3* are returned at time instant *t1*, and instances *inst2*, *inst4* are returned at time instant *t2*, such that *inst1 isInTheVicinityOf inst3* and *inst2* isInTheVicinityOf *inst4* as shown in lines 9 and 10 of Figure 4. *inst1* is *sameObject* with *inst2*, and *inst3* is *sameObject* with *inst4* as shown in lines 8 and 11 in Figure 4. The relation *isInTheVicinityOf* describes the spatial relationship while the relation *sameObject* establishes the temporal relationship over different frames, resulting in representation and retrieval of spatio-temporal information present in the video scene.

Similarly, other activities such as âĂIJa person touching the carâĂİ, âĂIJa person looking inside the carâĂİ etc., have also been defined with a suitable log-

ical description and their respective SPARQL queries are created. There are seven unusual or abnormal activities identified for this study. These are: âĂŞ

- *General Loitering*
- *Person walking around the car*
- *Person touching the car*
- *Person looking inside the car*
- *Person attempting to damage the car*
- *Group of people passing nearby car*
- *Trying to open the door of the car*

The above activities are formulated to test the applicability of the proposed approach and identified using SPARQL queries. Similarly, many more activities can be identified by defining more object properties, relations, and reasoning over those object properties.

3.7. Object Tracking

Object property *sameObject* established between two objects (one object of the current frame another of the subsequent frame) by comparing it with the all the objects in next frame. This *sameObject* property is extended over entire scene by applying transitive property on all the objects. If an object is same in first and second frame, and same in second, third frame, then the object is also same is first and third frame using transitive property as shown in Figure 5. The tracking results are listed in Table 5 as relation *sameObject*.

4. Results and Discussion

Low-level features are extracted from each frame of the video. The extracted features are then populated using the ontology for scene interpretation. Relations between the objects are inferred using SWRL rules. The temporal relations are represented by comparing the information present in current frame with the next frame, and then the transitivity of object properties is applied to identify relationships over the entire scene thereby, reducing the computation and making the framework very agile and scalable as compared to other machine learning and video processing techniques. An activity dataset[2] that contains six unusual activities, with a total of 92 videos is created. Details of the activity dataset is shown in Table 2. A person trying

[2]https://github.com/aspdr/Video-Representation-and-Suspicious-Event-Detection-using-Semantic-Technologies/tree/master/Trimmed%20Activity%20Dataset

SPARQL Query

```
1.  PREFIX owl:<http://
    www.w3.org/2002/07/owl#>
2.  PREFIX rdfs:<http://www.
    w3.org/2000/01/rdf-schema#>
3.  PREFIX xsd:<http://www.w3.
    org/2001/XMLSchema#>
4.  PREFIX time:<http://www.w3.
    org/2006/time#>
5.  PREFIX date:<https://www.w3.
    org/TR/owl-time/#>
6.  PREFIX owlpark:<http://www.
    semanticweb.org/
    myontology#>
7.  SELECT distinct ?inst1 ?inst2
    ?inst3 ?inst4 ?t1 ?t2
8.  where {SELECT ?inst1 ?inst2{
    ?inst1 owlpark:sameObject
    ?inst2.}
9.  ?inst1 owlpark:
    isInTheVicinityOf ?inst3.
10. ?inst2 owlpark:
    isInTheVicinityOf ?inst4.
11. ?inst3 owlpark:
    sameObject ?inst4.
12. ?inst1 owlpark:hasTime ?t1.
13. ?inst2 owlpark:hasTime ?t2.
14. FILTER
15. (?t1 = "2018-07-06T03:00:00"
    xsd:dateTime &&
16. ?t2="2018-07-06T03:00:03.6"
    xsd:dateTime)}
```

Fig. 4. SPARQL Query for moving around the Car

to open the car door or a person touching the car may not be suspicious in case the person is the owner or a genuine driver of the car. Therefore, a level of attention is assigned to each activity and classified as low, medium and high, as shown in Table 2. Experiments are performed for the activities listed in Table 2. The proposed approach is also applied to the crowd scene sequence dataset of the University of Florida (PNNL2) [47]. This dataset consists of a crowd movement near a parked car in an open parking space as shown in Figure 12 along with objects with respective bounding boxes.

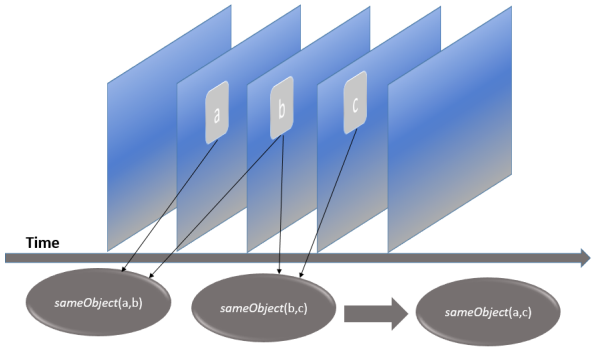


Fig. 5. Object Tracking using transitive property

Table 2
Description of Activity Dataset

Activity Description	Video Count	Suspicious	Alertness Level
Person walking around the car (Figure. 6)[46] [43]	10	Yes	High
Person touching the car (Figure. 7)	15	Potentially	Low
Person looking inside the car (Fig. 8)	24	Yes	High
Person attempting to damage the car (Figure. 9)	17	Yes	High
Group of people passing nearby car (Figure. 10)	11	Yes	Low
Trying to open the door of the car (Figure. 11)	15	Potentially	High

Fig. 6. Person walking around the car (Loitering)

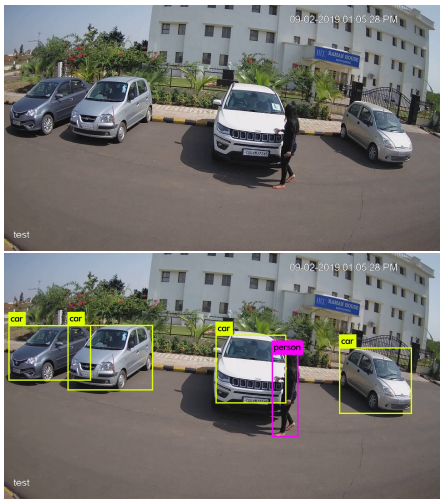


Fig. 7. Person touching the car

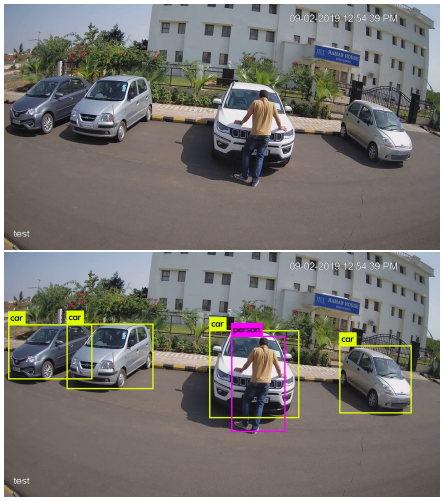


Fig. 8. Person looking inside the car

Performance and accuracy measures of various steps are given in Tables 3, 4, 5, and 6 .

4.1. Performance

The performance of the presented approach is tested on three different types of systems as shown in Table 3. The table lists execution time for processing ten frames in three environments. The first two systems are normal specification PCâĂŹs with easy to find system configuration. The third system is workstation with a relatively higher computing power consisting of Intel Xeon CPU, 64 GB RAM, and Nvidia 1080Ti GPU

Fig. 9. Person attempting to damage the car

Fig. 11. Person trying to open the door of the car.

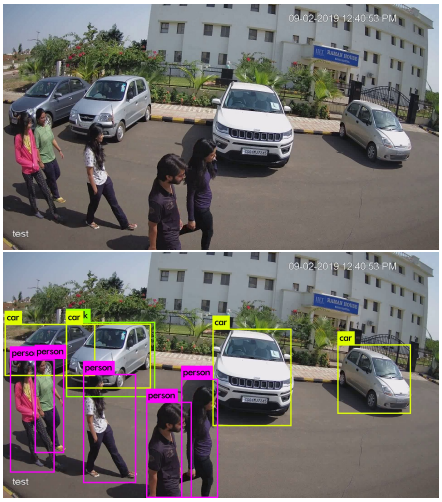


Fig. 10. A group passing by parking lot

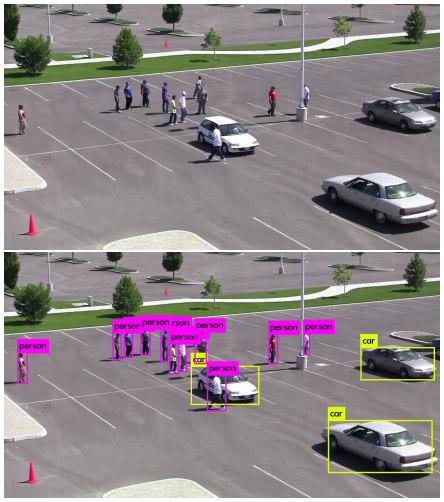


Fig. 12. Person walking around the vehicle in PNNL2 dataset [47].

with 11 GB of dedicated graphics memory. Total execution time consists of two components, i.e., time for extraction of low-level features and time for inference rule generation. The results demonstrate that the low-level feature extraction is highly dependent on presence of GPU as in the absence of GPU, the performance of the system is substantially low. However, it is observed that the time taken for generation of inference and reasoning does not depend on the presence of GPU. The inference and reasoning part performs reasonably well even with 8 GB of RAM and Intel i7 processor. This is a prominent finding of this work which promises a possibility of implementing event detection using smart devices present on the edge of the surveillance systems with pre-trained deep-net models (to reduce the extraction of low-level features) and meaningful knowledge inference through semantic technologies. The presented framework is computationally efficient, and can be easily deployed in any of the listed system configurations.

The presented framework is computationally efficient, and can be easily deployed in any of the listed system configurations. There is no notable difference between the time taken by each rule as the time complexity depends on the number of atoms used in rules [2] shown in Table 4. Inference time for the relation

Table 3
Execution time of the framework in different environments

CPU	Memory (GB)		Time taken to perform (seconds)	
	System	GPU	Low-Level Features	Inference & Reasoning
Intel I7	8	1	240	1.65
Intel I7	16	0	360	1.3
Intel Xeon	64	11	0.39	1.06

Table 4
Relation wise execution time for different environment

CPU	RAM (GB)	GPU (GB)	Relation	Inference Time (secs)	Variables
Intel I7	8	1	sameObject	1.384	10
			isInTheVicinityOf	1.318	10
			overlaps	1.306	8
			notMoving	1.252	8
			moving	1.246	8
Intel I7	16	1	sameObject	1.187	10
			isInTheVicinityOf	1.128	10
			overlaps	1.135	8
			moving	1.147	8
			notMoving	1.151	8
Intel Xeon	64	11	sameObject	0.987	10
			isInTheVicinityOf	0.920	10
			overlaps	0.971	8
			moving	0.987	8
			notMoving	0.957	8

sameOb ject is highest as it has most number of atomic antecedents. It is also evident from the table that inference time is least in the server which is having the most powerful CPU.

4.2. Accuracy

The proposed work is evaluated by measuring the accuracy of the inferred relations and activities. Accuracy of SPARQL query directly depends on inferred relationship as it extracts presence of relationship in the temporal domain. We measured the accuracy of inferred relations, on open parking dataset [47] and the accuracy of activities on our dataset shown in Table 2.

4.2.1. Relationship Inference

The accuracy of the inferred relations is measured by calculating precision and recall of relations by forming a confusion matrix shown in Table 5. The matrix is calculated by considering relations among 7325 objects of open parking dataset [47]. Each frame has 14 objects on an average, and the number of frames per second (fps) of the video is 15. In total 70616 relations (*overlaps, isInTheVicinityOf, sameObject, moving and notMoving*) exists between the objects, of which the inferred *sameObject* relation has more number of false positives because of close objects and object occlusion. Thus, proposed framework performs quite well in inferring the complex spatio-temporal relations and interactions between the objects.

Table 5
Accuracy of inferring the relations in a video scene

Relation	Total Relations	Precision (%)	Recall (%)
overlaps	2436	97.63	98.59
isInTheVicinityOf	380	98.12	99.43
sameObject	60475	81.72	97.45
moving	5429	98.36	99.37
notMoving	1896	96.03	98.71

4.2.2. Activity Detection

Activity recognition is performed by executing SPARQL queries tailored explicitly for a particular activity as listed in Table 2. The number of true alarm, false alarm and missed alarm are recorded to calculate the accuracy of the approach. The results are shown in Table 5.

- **True Alarm** âĂŞ The number of correct recognition, when system identified it and it also happened.
- **False Alarm** - The activity does not happen in real but system recognized it as occurred. These are the negative examples in the dataset.
- **Misses** - The activity which goes unrecognized by the system, but occurred.

As per Table 6, the following observation can be made for respective activities:

- ***Loitering*** - There are few misses due to the movement of person behind the car, due to which it becomes an occluded object.
- ***Person looking inside the car*** - Misses occur due to occlusion of the person behind another car and false alarms are reported because the person is not actually looking inside the car but just standing in front and looking at something else.
- ***Person attempting to damage the car*** - Few misses are observed because the person attempts

Table 6
Accuracy of the various activities by the framework

Activity Description	Total	True Alarm	False Alarm	Misses	True Alarm (%)
Person walking around the car	10	8	0	2	80
Person touching the car	17	13	2	2	76.67
Person looking inside the car	26	20	2	4	76.92
Person attempting to damage the car	17	15	0	2	88.23
Group of people passing nearby car	11	6	0	5	54.55
Trying to open the door of the car	15	12	0	3	80

to damage the car, but took lesser time than expected.

– ***Group of people passing nearby car*** - Misses are observed because group was moving at a significant distance from the parked car.
– ***Trying to open the door of the car*** - At times, misses are observed as the time taken is less than the time assigned for the activity.

4.2.3. Activity: General loitering

The proposed methodology is utilized to identify the abnormal activity of loitering which is identified as one of the most common suspicious behaviors in the literature [43]. Loitering is defined as a person who enters the scene and remains within the scene for more than a certain duration. For reference, the duration of 60 seconds was mentioned in PETS2007[48]. The SPARQL query is formulated for defining loitering activity. It is said to be loitering when an individual is of type *person* and present in scene for more than t seconds. The âĂŸtâĂŹ here is kept 20 seconds. Our methodology performs significantly better in detecting loitering activity as compared to the previous approach [49] in literature as shown in Table 7. The performance of the approach is compared with the versatile loitering [49] on PETS 2006 and PETS 2016 datasets. Following sequences of PETS2006 [50] and PETS2016 [51] dataset are used:

– ***PETS2006-S1-T1-C*** - This is a left luggage scenario listed in PETS2006, in which a person loiters before leaving the unattended luggage. The activity predominantly occurs between duration of 40 to 80 seconds. Thus, the sequence from 40 seconds onwards is used to classify the activity. The snapshot of the sequence is shown in Figure 13 and results are shown in Table 7.

Fig. 13. Sequence S1-T1-C containing movement of a person with luggage in PETS2006 dataset

– ***PETS2006 - S2-T3-C*** - This is a left luggage scenario listed in PETS2006, in which two persons enters the video scene from front of each other. The first person carries the suitcase, making it unattended in the ground, loiters, then leave afterwards. The activity predominantly occurs between duration of 40 to 80 seconds. Thus, the sequence from 40 seconds onwards is used to classify the activity. The snapshot of the sequence is shown in Figure 14 and results are shown in Table 7.
– ***PETS2006 - S3-T7-A*** - This is also a left luggage scenario listed in PETS2006, in which a per-

Fig. 14. Sequence S2-T3-C containing the movement of a person with luggage in PETS2006 dataset

Fig. 15. Sequence S3-T7-A containing the movement of a person with luggage in PETS2006 dataset

son is waiting for a train. While waiting, the person keeps his suitcase in the ground and picks it up again after some time, making it unattended for a certain duration. The person loiters before picking up the unattended luggage. The activity occurs between a duration of 20 to 60 seconds. Thus, the sequence from 20 seconds onwards is used to classify the activity, containing at least 20 seconds of loitering. The snapshot of the sequence is shown in Figure 15 and results are shown in Table 7.

– ***PETS2016 - 03_06*** - This sequence is labeled as something is wrong scenario in PETS2016 dataset. It is part of the ARENA multi-camera dataset containing various activities around the parked vehicle in a parking lot. The sequence contains suspicious behavior of the loitering of a person near a truck. The total length of the sequence is 78 seconds. It involves the movement of a person near the truck. For this study data from two cameras, TRK_RGB_1 and TRK_RGB_2 are processed. The TRK_RGB_1 captures the initial movement of the loiter as normal, while TRK_RGB_2 captures the loitering activity which occurs for more than 20 seconds. Our approach performs better than the existing approach in detecting this activity, as listed in Table 7. The snapshot of the sequence is also shown in Figure 16 and Figure 17.

– ***PETS2016 - 14_05*** - This sequence is also part of PETS2006 ARENA dataset and 100 seconds long. It is labeled as a criminal scenario as one person is loitering and other person steals. We have only processed and classified the video containing loitering activity captured in TRK_RGB_3 clip in the latter part of the

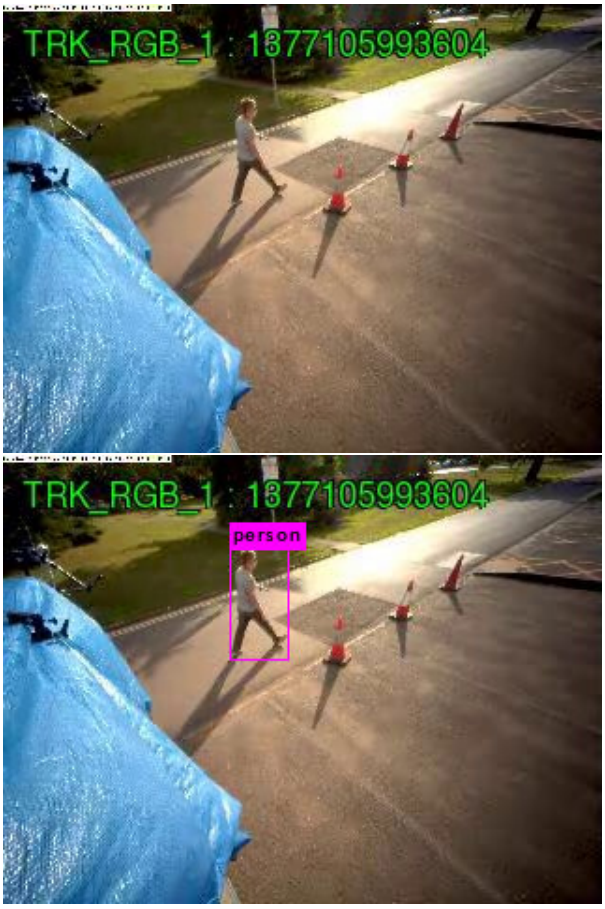


Fig. 16. Sequence 03_06 TRK_RGB_1 containing the movement of person near parked truck in PETS2016 dataset

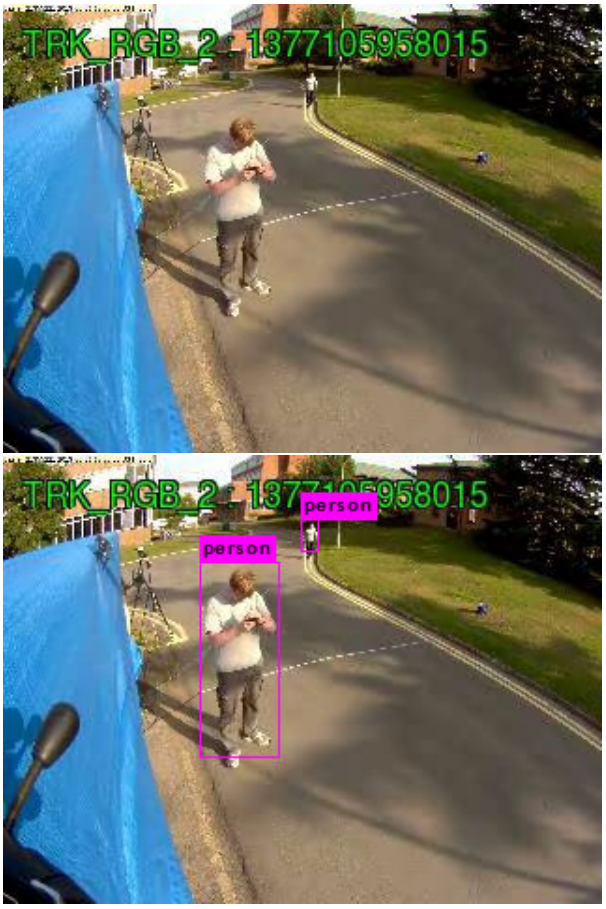


Fig. 17. Sequence 03_06 TRK_RGB_2 containing loitering in PETS2016 dataset

video clip. Our approach successfully detected the activity of loitering but the existing approach identified it a normal activity in Table 7, demonstrating the robustness of our proposed approach and framework. The snapshot of the sequence is shown in Figure 18.

4.2.4. Activity: Person walking around the the vehicle

Suspicious activity of moving around the vehicle is evaluated on sequence of PETS2014 Arena dataset [52] containing the series of multiple camera recordings for understanding human behavior around the vehicle with the intention pro-actively identifying the potential threats. Therefore, we have evaluated accuracy of our framework on the various sequences of PETS2014 dataset. If the relationship *isInTheVicinityOf* holds between truck and person for 20 seconds then it is classified as suspicious activity of moving around the vehicle. Furthermore, the sequences are bifurcated to generate multiple scenarios for evaluating the robustness of the proposed approach:

- **PETS2014 - 06_01** - The sequence (ENV_RGB_3) consists of security personnel moving around the truck and labeled in something is wrong scenario in the dataset. The total length of the sequence is 3 minutes 16 seconds, which involves the movement of a security guard around the truck. The sequence is further divided into three sub-parts as listed in Table 8 and the presence of a person is detected as suspicious behavior (something is wrong as labeled in the dataset) in all the subsequences. The snapshot of the sequence is shown in Figure 19.
- **PETS2014 - 06_04** - The sequence (ENV_RGB_3) consist of two security personnel moving around the truck and labeled as something is wrong scenario in the dataset. The total length of the se-

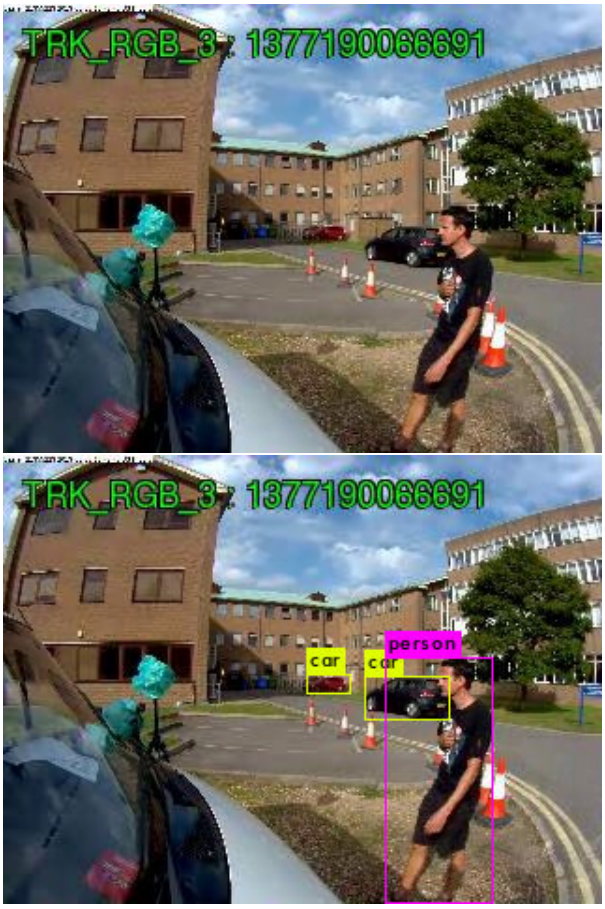


Fig. 18. Sequence 14_05 TRK_RGB_2 containing loitering in PETS2016 dataset

quence is 100 seconds, involves movement of security guard around the truck. For sometime, the security guards are not captured as they are on the other side of the truck. The duration was set to 20 seconds. The sequence is further divided in two sub parts as listed in Table 8 and the activity is successfully identified in both of the sequences using our SPARQL query. The snapshot of the sequence is shown in Figure 20.

– **PETS2014 - 10_03 ENV_RGB_3** - The sequence (ENV_RGB_3) consists of a person walking around the parked vehicle. For some-time person is not captured as he is behind the truck. Later on, come back towards the back of the truck and then walks away from the front. The scene also contains few other people also walking through the pathway. The scenario is labeled as something is wrong scenario in the dataset. The total length of the sequence is 78 seconds. The sequence is further divided into two subparts, as listed in Table 8. The snapshot of the sequence is shown in Figure 21.

Table 7
Comparison of the Loitering detection with the existing approach

Dataset	Sequence Name	Time frame (secs)	Actual	Approach [49]	Ours
PETS 2006	S1-T1-C-3	41-80	Loiter	Loiter	Loiter
	S2-T3-C-3	41-80	Loiter	Loiter	Loiter
	S3-T7-A-3	21-60	Loiter	Loiter	Loiter
PETS 2016	03_06 TRK RGB_2	11-50	Loiter	No Loiter	Loiter
	03_06 TRK RGB_1	11-50	No Loiter	No Loiter	No Loiter
	14_05 TRK RGB_2	11-50	Loiter	No Loiter	Loiter

– **PETS2014 - 10_04 ENV_RGB_3** - The sequence (ENV_RGB_3) consists of a person walking around the parked vehicle. The person enters from the back of the truck, then takes a round from the front and goes on the other side of the truck. The person when on the other side of the truck is not captured, leading to not detection of the activity. The scene also contains two other person walking through the pathway and a car passing by. The scenario is labeled as something is wrong scenario in the dataset. The total length of the sequence is 78 seconds. The sequence is further divided into two subparts, as listed in Table 8. The snapshot of the sequence is shown in Figure 22.

– **PETS2014 - 10_05 ENV_RGB_3** - The sequence (ENV_RGB_3) consist of a person walking around the parked vehicle. The person enters the scene from the front of the vehicle, then stops near the door of the vehicle and again turns backward and walks towards another side of the vehicle. The person could not be captured in the second half of the clip as he is behind the truck (in invisible region of the camera). The total length of the sequence is 60 seconds, which involves the movement of a security guard around the truck. The sequence is further divided into three sub-

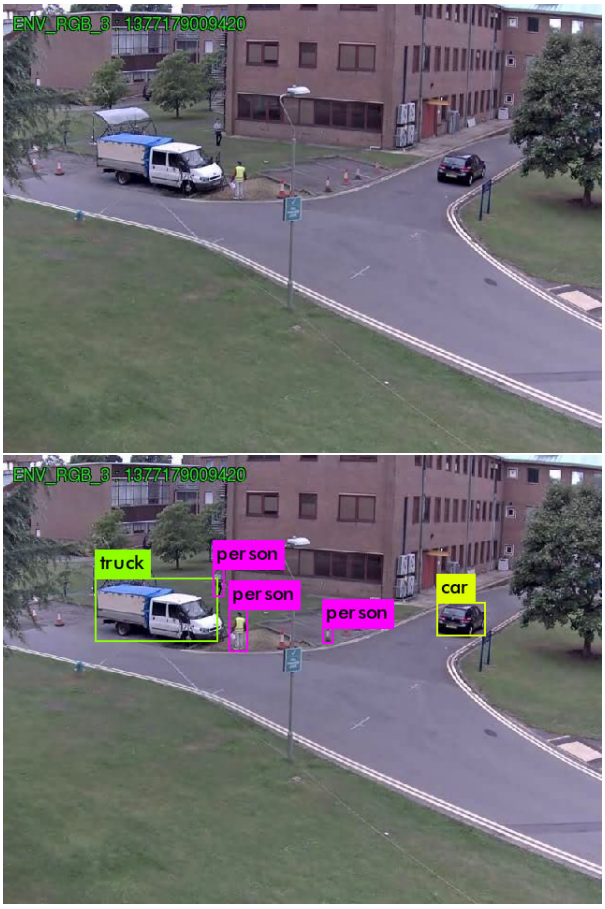


Fig. 19. Sequence 06_01 ENV_RGB_3 containing movement of person around the vehicle in PETS2004 dataset

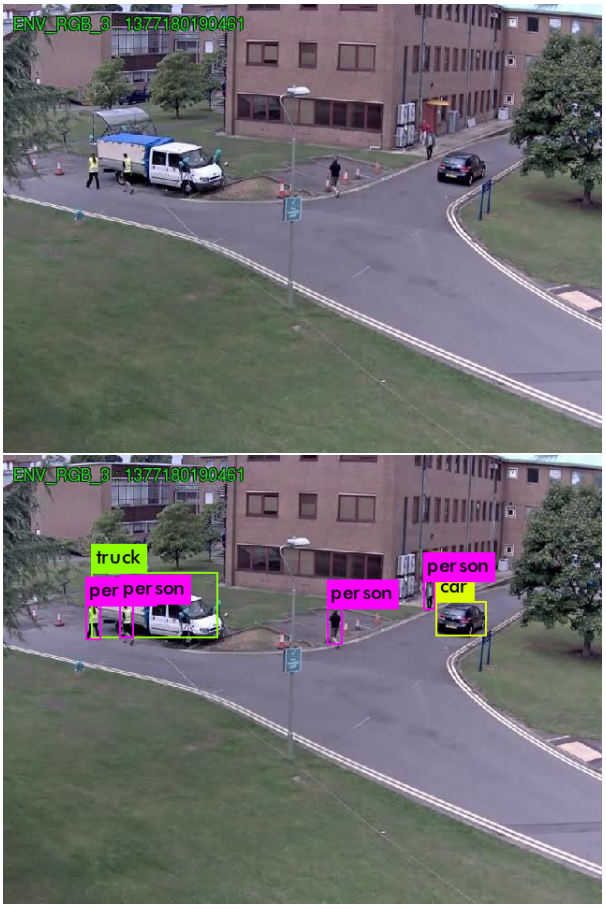


Fig. 20. Sequence 06_04 ENV_RGB_3 containing the movement of person around the vehicle in PETS2004 dataset

parts, as listed in Table 8. The snapshot of the sequence is shown in Figure 23.

The performance of the proposed approach is shown in Table 8 demonstrating the effectiveness to detect potential criminal scenario of unintentional movement around the parked vehicle on PETS 2014 dataset. However, in few scenario the activity could not be detected as the object (person) got occluded behind the bigger object (truck). Results demonstrate the accuracy of the proposed approach and create an exceptional foundation to categorize an event as suspicious or normal on top of which further decisions can be made. The proposed work can also be readily applied to several such scenarios for activity detection and scene representation.

5. Conclusion

In this paper, a novel approach of event detection and video analytics is presented, by creating a framework for event detection, and video understanding with high-level semantics. Firstly, frame-level features are extracted from the video. Secondly, an ontology is developed, which can represent the frame-level information of the use-case video data, i.e., parking lot footage. Selected frame-level information is mapped to data properties of the developed ontology. Relationships between objects in a video footage are identified using ad-hoc SWRL rules, including rules for object tracking. A labeled dataset of suspicious activities is created to test the applicability of the proposed approach. This dataset can be highly useful beyond the scope of this article, to aid developers in providing solutions to parking-related challenges, as there is no comparable preexisting dataset, to the best of our knowl-

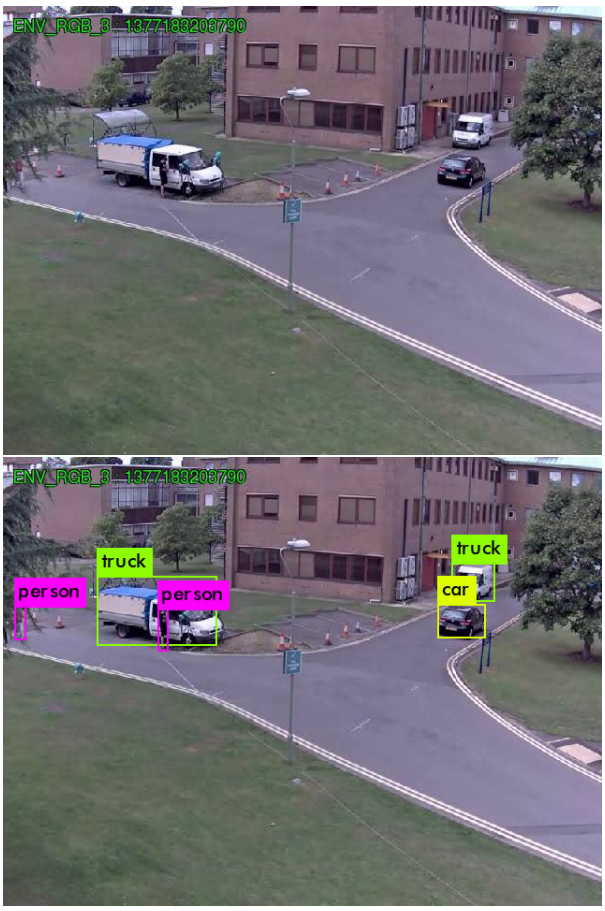


Fig. 21. Sequence 10_03 ENV_RGB_3 containing the movement of person around the vehicle in PETS2004 dataset

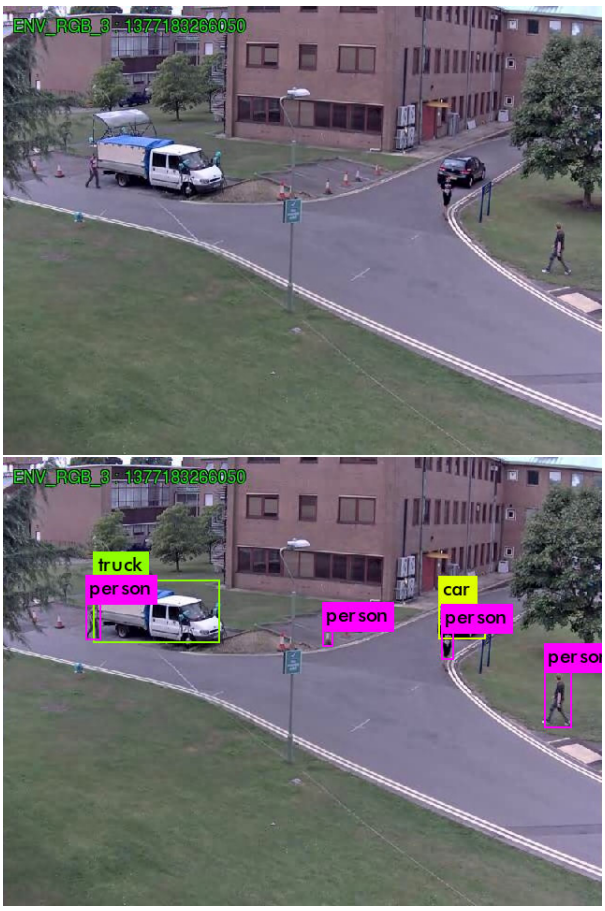


Fig. 22. Sequence 10_04 ENV_RGB_3 containing the movement of person around the vehicle in PETS2004 dataset

edge. As a use case, six suspicious events are identified in surveillance footage of a parking lot, thus filling a well-known semantic gap between low-level features of a video and high-level (hidden) semantics. Furthermore, the approach is also validated on the PETS2006, PETS2014 and PETS2016 datasets. Our approach can help in representing most of the spatial and temporal information present in the video, which could be useful for object tracking and generating high-level semantic information. Our work demonstrated the potential in leveraging semantic web technology for activity detection, especially useful in scenarios featuring lack of training data and limited computing resources. The proposed approach also covers extracting and representing salient information present in video frames in machine-readable and machine-interpretable format, which improves ease of retrieval, processing, and storage.

6. Future Work

Video representation and event detection opens up a plethora of use-cases, and are applicable to various domains. In the future, an extensive comparison shall be done with other machine learning approaches. One of the vital requirements for smart city initiatives lies in the applicability of low-power edge devices having limited computing resources. Therefore, the ability of edge/fog computing devices to handle such workloads, deriving from the effectiveness of the current approach, will also be investigated. Most of the activities which can be categorized as alarming, also include some common activities like walking, running, standing, touching, sitting, opening the door, etc. These are sub-activities that do not require interactions among multiple objects in the time domain but focus only on a few objects. By detecting these sub-activities, more complex (and atypical) activities can be derived. Ob-

Fig. 23. Sequence 10_05 ENV_RGB_3 containing movement of person around the vehicle in PETS2004 dataset

ject tracking is one of the most challenging problems in the field of computer vision. The object tracking approach presented in this paper will be further explored to benchmark tracking and evaluate results. Another integral part of this approach is ontology engineering, which requires substantial manpower, skilled in domain-specific concepts. Although a significant amount of work in literature is available with regard to automatic concept detection from the perspective of ontology construction, most work still remains tedious, error-prone, and time-consuming. A video ontology may work well enough to represent all the interactions between objects, both spatially and temporally. An ontology thus built, can actually be used to represent a complete video in a RDF graph, for ease of storage, access, and retrieval. Therefore, methods and data-driven techniques to generate ontologies automatically will also be investigated.

Table 8

Performance of framework in identifying potentially criminal activity of moving around the vehicle on PETS2014 dataset

Sequence	Time frame (secs)	Actual	Outcome	Observation
06_01 ENV RGB_3	1-50	Suspicious	Suspicious	Detected in all clips
	51-100	Suspicious	Suspicious	
	101-150	Suspicious	Suspicious	
06_04 ENV RGB_3	0-50	Suspicious	Suspicious	Person is not visible in second clip
	51-100	Suspicious	Normal	
10_03 ENV RGB_3	1-40	Suspicious	Normal	Person not captured in the first clip
	41-78	Suspicious	Suspicious	
10_04 ENV RGB_3	1-40	Suspicious	Normal	Person not captured in the second clip
	41-79	Suspicious	Suspicious	
10_05 ENV RGB_3	1-40	Suspicious	Suspicious	Person not captured in the second clip
	41-59	Suspicious	Normal	

References

[1] Cisco Visual Networking Index: Forecast and Methodology 2016-2021, 2017. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html.

[2] L.F. Sikos, *Description logics in multimedia reasoning*, Springer, 2017, pp. 1–205. ISBN 978-3-319-54066-5. doi:10.1007/978-3-319-54066-5.

[3] A.S. Patel, M. Ojha, M. Rani, A. Khare, O.P. Vyas and R. Vyas, Ontology-Based Multi-agent Smart Bike Sharing System (SBSS), in: *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*, 2018, pp. 417–422. doi:10.1109/SMARTCOMP.2018.00039.

[4] C. Bizer, T. Heath and T. Berners-Lee, Linked data-the story so far, *International journal on Semantic Web and Information Systems* **5**(3) (2009), 1–22. ISBN 1552-6283. doi:10.4018/jswis.2009081901. http://eprints.soton.ac.uk/271285/.

[5] I. Horrocks, P.F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean et al., SWRL: A semantic web rule language combining OWL and RuleML, *W3C Member submission* **21**(79) (2004), 1–31.

[6] L.F. Sikos, VidOnt : a core reference ontology for reasoning over video scenes scenes *, *Journal of Information and Telecommunication* **2**(2) (2018), 1–13. doi:10.1080/24751839.2018.1437696.

[7] J. You, G. Liu and A. Perkis, A semantic framework for video genre classification and event analysis, *Signal Processing: Image Communication* **25**(4) (2010), 287–302. doi:10.1016/j.image.2010.02.001.

[8] Z. Si, M. Pei, B. Yao and S.C. Zhu, Unsupervised learning of event AND-OR grammar and semantics from video, *Proceedings of the IEEE International Conference on Computer Vision* (2011), 41–48. doi:10.1109/ICCV.2011.6126223.

[9] X. Zhu, X. Wu, A.K. Elmagarmid, Z. Feng and L. Wu, Video data mining: Semantic indexing and event detection from the association perspective, *IEEE Transactions on Knowledge and Data Engineering* **17**(5) (2005), 665–677. doi:10.1109/TKDE.2005.83.

[10] X. Wang and Q. Ji, Video event recognition with deep hierarchical context model, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **07-12-June** (2015), 4418–4427. doi:10.1109/CVPR.2015.7299071.

[11] L. Xie, H. Sundaram and M. Campbell, Event mining in multimedia streams, *Proceedings of the IEEE* **96**(4) (2008), 623–647. doi:10.1109/JPROC.2008.916362.

[12] R. Hamid, S. Maddi, A. Bobick and I. Essa, Structure from statistics - Unsupervised activity analysis using suffix trees, *Proceedings of the IEEE International Conference on Computer Vision* (2007). ISBN 978-1-4244-1630-1. doi:10.1109/ICCV.2007.4408894.

[13] F. Baradel, C. Wolf, J. Mille and G.W. Taylor, Glimpse Clouds: Human Activity Recognition from Unstructured Feature Points, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2018), 469–478. ISBN 9781538664209. doi:10.1109/CVPR.2018.00056.

[14] W. Liao, C. Yang, M. Ying Yang and B. Rosenhahn, SECURITY EVENT RECOGNITION for VISUAL SURVEILLANCE, *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **4**(1W1) (2017), 19–26. doi:10.5194/isprs-annals-IV-1-W1-19-2017.

[15] C.G. Snoek and M. Worring, Concept-based video retrieval, *Foundations and Trends in Information Retrieval* **2**(4) (2008), 215–322. doi:10.1561/1500000014.

[16] Z. Cheng, X. Li, J. Shen and A.G. Hauptmann, Which information sources are more effective and reliable in video search, *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2016), 1069–1072. doi:10.1145/2911451.2914765.

[17] J. Shen, D. Tao and X. Li, Modality mixture projections for semantic video event detection, *IEEE Transactions on Circuits and Systems for Video Technology* **18**(11) (2008), 1587–1596. doi:10.1109/TCSVT.2008.2005607.

[18] Generating Video Descriptions with Latent Topic Guidance, *IEEE Transactions on Multimedia* **21**(9) (2019), 2407–2418. doi:10.1109/TMM.2019.2896515.

[19] L. Caruccio, G. Polese, G. Tortora and D. Iannone, EDCAR: A knowledge representation framework to enhance automatic video surveillance, *Expert Systems with Applications* (2019). doi:10.1016/j.eswa.2019.04.031.

[20] C. Gan, N. Wang, Y. Yang, D.Y. Yeung and A.G. Hauptmann, DevNet: A Deep Event Network for multimedia event detection and evidence recounting, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **07-12-June** (2015), 2568–2577. ISBN 9781467369640. doi:10.1109/CVPR.2015.7298872.

[21] D. He, F. Li, Q. Zhao, X. Long, Y. Fu and S. Wen, Exploiting Spatial-Temporal Modelling and Multi-Modal Fusion for Human Action Recognition (2018). http://arxiv.org/abs/1806.10319.

[22] Z. Qiu, T. Yao and T. Mei, Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks, in: *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5534–5542. doi:10.1109/ICCV.2017.590.

[23] R. Nevatia, J. Hobbs and B. Bolles, An ontology for video event representation, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2004). doi:10.1109/CVPR.2004.301.

[24] A.R.J. François, R. Nevatia, J. Hobbs and R.C. Bolles, VERL: An ontology framework for representing and annotating video events, *IEEE Multimedia* **12**(4) (2005), 76–86. doi:10.1109/MMUL.2005.87.

[25] A.J. Bermejo, J. Villadangos, J.J. Astrain, A. Córdoba, L. Azpilicueta, U. Gárate and F. Falcone, Ontology based road traffic management in emergency situations, *Ad-Hoc and Sensor Wireless Networks* **20**(1–2) (2013), 47–69.

[26] J. Fan, H. Luo, Y. Gao and R. Jain, Incorporating concept ontology for hierarchical video classification, annotation, and visualization, *IEEE Transactions on Multimedia* **9**(5) (2007), 939–957. doi:10.1109/TMM.2007.900143.

[27] T.H. Duong, N.T. Nguyen, H.B. Truong and V.H. Nguyen, A collaborative algorithm for semantic video annotation using a consensus-based social network analysis, *Expert Systems with Applications* **42**(1) (2015), 246–258. doi:10.1016/j.eswa.2014.07.046.

[28] N. Elleuch, M. Zarka, A. Ben Ammar and M.A. Alimi, A fuzzy ontology: based framework for reasoning in visual video content analysis and indexing, *Proceedings of the Eleventh International Workshop on Multimedia Data Mining* (2011), 1. ISBN 978-1-4503-0841-0. doi:10.1145/2237827.2237828.

[29] M. Grassi, C. Morbidoni and M. Nucci, A Collaborative Video Annotation System Based on Semantic Web Technologies, *Cognitive Computation* **4**(4) (2012), 497–514. doi:10.1007/s12559-012-9172-1.

[30] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann and Z. Ives, DBpedia: A Nucleus for a Web of Open Data, in: *In 6th IntâĂŹl Semantic Web Conference, Busan, Korea*, Springer, 2007, pp. 11–15.

[31] J. Gómez-Romero, M.A. Patricio, J. García and J.M. Molina, Ontology-based context representation and reasoning for object tracking and scene interpretation in video, *Expert Systems with Applications* **38**(6) (2011), 7494–7510. doi:10.1016/j.eswa.2010.12.118.

[32] Z. Xu, Y. Liu, L. Mei, C. Hu and L. Chen, Semantic based representing and organizing surveillance big data using video structural description technology, *Journal of Systems and Software* **102** (2015), 217–225. doi:10.1016/j.jss.2014.07.024.

[33] D. Vallet, P. Castells, M. Fernández, P. Mylonas and Y. Avrithis, Personalized content retrieval in context using ontological knowledge, *IEEE Transactions on Circuits and Systems for Video Technology* **17**(3) (2007), 336–345. doi:10.1109/TCSVT.2007.890633.

[34] M. Naphade, J.R. Smith, J. Tesic, S.F. Chang, W. Hsu, L. Kennedy, A. Hauptmann and J. Curtis, Large-scale concept ontology for multimedia, *IEEE Multimedia* **13**(3) (2006), 86–91. doi:10.1109/MMUL.2006.63.

[35] A. Hauptmann, R. Yan, W.H. Lin, M. Christel and H. Wactlar, Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news, *IEEE Transactions on Multimedia* **9**(5) (2007), 958–966. doi:10.1109/TMM.2007.900150.

[36] K. Mahmood, Cloud Based Sports Analytics Using Semantic Web Tools and Technologies (2015), 431–433. doi:10.1109/GCCE.2015.7398708.
[37] M.Y.K. Tani, A. Lablack, A. Ghomari and I.M. Bilasco, Events detection using a video-surveillance ontology and a rule-based approach, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **8926** (2015), 299–308. ISBN 9783319161808.
[38] L.F. Sikos and D.M.W. Powers, Knowledge-Driven Video Information Retrieval with LOD: From Semi-Structured to Structured Video Metadata, *Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval* (2015), 35–37. doi:10.1145/2810133.2810141.
[39] L.F. Sikos, A Novel Approach to Multimedia Ontology Engineering for Automated Reasoning over Audiovisual LOD Datasets, in: *ACIIDS*, 2016.
[40] M. Jangid, V.K. Verma and V.G. Shankar, Counting and Classification of Vehicle Through Virtual Region for Private Parking Solution, in: *Proceedings of First International Conference on Smart System, Innovations and Computing*, Springer Singapore, Singapore, 2018, pp. 761–770.
[41] M. Tschentscher, B. PruÃ§ and D. Horn, A simulated car-park environment for the evaluation of video-based on-site parking guidance systems, in: *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 1571–1576. doi:10.1109/IVS.2017.7995933.
[42] B. Zhang, V. Appia, I. Pekkucuksen, Y. Liu, A.U. Batur, P. Shastry, S. Liu, S. Sivasankaran and K. Chitnis, A Surround View Camera Solution for Embedded Systems, in: *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 676–681. ISSN 2160-7508. doi:10.1109/CVPRW.2014.103.
[43] A. Ben Mabrouk and E. Zagrouba, Abnormal behavior recognition for intelligent video surveillance systems: A review, *Expert Systems with Applications* **91** (2018), 480–491. doi:10.1016/j.eswa.2017.09.029.
[44] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, in: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. doi:10.1109/CVPR.2016.91.
[45] M.A. Musen, The protégé project: a look back and a look forward, *AI Matters* **1**(4) (2015), 4–12. doi:10.1145/2757001.2757003.
[46] G.L. Foresti, C. Micheloni and L. Snidaro, Event classification for automatic visual-based surveillance of parking lots, *Proceedings - International Conference on Pattern Recognition* **3**(Dimi) (2004), 314–317. ISBN 0769521282. doi:10.1109/ICPR.2004.1334530.
[47] A. Dehghan, S.M. Assari and M. Shah, GMMCP tracker: Globally optimal Generalized Maximum Multi Clique problem for multiple object tracking, in: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4091–4099. ISSN 1063-6919. doi:10.1109/CVPR.2015.7299036.
[48] J. Ferryman, PETS 2007 Benchmark Data, 2007. http://www.cvg.reading.ac.uk/PETS2007/data.html.
[49] S. Arivazhagan and R. Newlin Shebiah, Versatile loitering detection based on non-verbal cues using dense trajectory descriptors, *Multimedia Tools and Applications* **78**(8) (2019), 10933–10963. doi:10.1007/s11042-018-6618-9.
[50] J. Ferryman, PETS 2006 Benchmark Data, 2006. http://www.cvg.reading.ac.uk/PETS2006/data.html.
[51] L. Patino, T. Cane, A. Vallee and J. Ferryman, PETS 2016: Dataset and Challenge, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2016), 1240–1247. ISBN 9781467388504. doi:10.1109/CVPRW.2016.157.
[52] L. Patino and J. Ferryman, PETS 2014: Dataset and challenge, in: *11th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2014*, 2014. ISBN 9781479948710. doi:10.1109/AVSS.2014.6918694.