Ontology-Based Analysis of CCTV Data

N.F. Kahar, E. Izquierdo
Multimedia and Vision Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
Mile End Road, London E1 4NS
United Kingdom
{n.f.kahar, e.izquierdo}@qmul.ac.uk

Keywords: Ontology-based video analysis, Knowledge representation, Semantic reasoning, Rules, SPARQL Query.

Abstract

The availability of vast amount of closed-circuit television (CCTV) videos and the critical need for a fast and accurate analysis of all its content especially in security applications, has led to an equally urgent need for its automated processing. Indeed, identifying suspected acts of vandalism or crime from a massive collection of videos are extremely tedious using manual approach. Therefore, with application for security in mind, this paper presents a new approach for automated analysis of CCTV using ontology-based knowledge representation and reasoning. The proposed approach exploits event semantics for the identification of possible crime and perpetrators. The framework consists of salient feature extraction from CCTV footage, inter-level data parsing, high-level conceptual model event representation, rules inferencing, knowledge reasoning and queries. Although a complete (end-to-end) framework was implemented, the main contribution of this work is primarily on the high-level analysis and proposed underpinning Ontology. The proposed approach has been validated with a very large database of CCTV footage from the 2011 London riots. The results validate the strength of the approach.

1 Introduction

Today, CCTV surveillance has become a prominent aspect in day-to-day life as an approach to crime prevention, traffic monitoring and home security. According to IHS, there are 245 million professionally installed video surveillance cameras active and operational globally in 2014 [1] and the number has increased to 350 million in 2016 [2]. The British Security Industry Authority (BSIA) estimated that there are up to 5.9 million CCTV cameras in the UK alone [3]. However, the problem arises due to the fact that although there is a huge amount of CCTV data gathered, currently there is no efficient way to analyse it. The enormous collection of data has also led to issues especially those affecting the timely process of justice. This is a major drawback of security-related domain. Consequently, there is a growing need for an automated event understanding that focuses on security, crime detection and prevention.

Using ontologies for understanding events promises to significantly improve automated video analysis [4]–[7]. Some research have proposed ontologies for specific domains of visual surveillance [8]–[10]. This paper presents an ontology-based analysis of CCTV data consisting of features extraction, parsing module for inter-level data transformation, higher-level event representation, knowledge reasoning and queries.

The contribution of this work lays in the implementation of ontology-based knowledge representation and reasoning in visual surveillance application by utilizing data sources obtained from lower-level video analysis. Information such as objects and actions obtained from the video analysis module are used to deduce new knowledge in the system, which leads to high-level interpretation of video scenes. The rest of this paper is organized as follows. First, Section 2 presents the system framework and describe the modules’ functionality. Section 3 focuses on the conceptual model for event representation and properties of concepts in the ontology. Section 4 discuss on the formal representation of situations and section 5 presents rules implementation, reasoning and knowledge base queries. Section 6 reports on the experimental results of the implemented framework and finally in section 7 conclusions are outlined.

2 System Framework

The implemented system is divided into four main modules: sensing, video analysis, parsing and semantic reasoning as shown in Fig. 1.

![Fig. 1. System framework](image-url)
2.1 Sensing Module

The sensing module consists of smart surveillance cameras and other sensors. It provides the input data to be analysed. CCTV is used to monitor events that take place at specific locations in real time. There are various types of CCTV cameras today, with many different features and options. Cameras may be focused on a fixed location, set to scan a particular area, or they can be operated remotely by specially trained operators. However, the approach presented in this paper does not rely on any specific CCTV or sensors. We process footage “from the wild” as captured by any conventional CCTV system available to the security forces.

2.2 Video Analysis Module

The video analysis module performs analysis to detect objects, recognize actions and other salient features from the video footage to represent the scene model in the ontology. We implemented Histogram of Oriented Gradient (HOG) feature descriptor as discussed by Dalal and Triggs in [11] for person detection task, and for face detection, we adopted Haar feature-based cascade classifier approach based on the work of Paul Viola and Michael Jones in [12]. Note that the low-level processing for object detection is not our focus of discussion in this paper. The methods implemented in the video analysis module are chosen because they give a good detection results with lower computational cost despite both methods are not among the recent ones.

For action recognition, manual annotations are implemented to represent actions such as running, throwing and kicking in the video footage. The annotation process is being done using ViPER-GT. It provides the process of authoring ground truth through frame-by-frame mark up of video metadata stored in the Viper format. To create descriptors for every action, we use spatial attribute type bbox to represent actions in the video. The bbox gives the coordinates of the top-left corner of the box, the height, the width and the frame number of the annotations. The descriptor composed of type of actions, name and attributes. The descriptor file is a simple XML based format.

2.3 Parsing Module

The information concerning detected objects and actions are transformed into formal semantic data by the parsing module. The parsing module utilizes the syntactic data from the video analysis module and transforms it to high-level semantic concepts in an ontology represented in OWL/RDF. They achieve three main objectives, which are the enrichment, generation and population of an OWL ontology. There are a number of transformation approaches from eXtensible Markup Language (XML) to Web Ontology Language (OWL) that have been developed as discussed in these survey [13], [14]. However, studies on various mapping approaches has led us to adopt a method called XML2OWL proposed by Bohring and Auer [15] as it aligns with our goal which focuses on populating new instances to the ontology.

2.4 Semantic Reasoning Module

Following parsing process, the semantic data is handed to the semantic reasoning module which manage all the semantic operations. Semantic reasoning module consists of RDF store and domain ontology for high-level representations; rules, rule engine and reasoner for rule-based classifications and inferences; and query engine for knowledge queries. The queries are used to retrieve the in-memory triples in the newly asserted knowledge. The retrieved information will help users to interpret the semantic content in the video footage.

3 Ontology

The main contribution of this work is on the high-level analysis and proposed underpinning Ontology. The purpose of the ontology is to formalize the basic concepts, attributes of concepts and the relationships between concepts in a domain of discourse. A generic concept ontology could support a broad range of applications in the similar domain, thus reduce developmental efforts and realize gains through reuse. For this reason, the common event model for multimedia applications proposed in [16] is implemented as a basis of our ontology. It consists of six basic aspects of event description which encompass temporal, spatial, informational, structural, experiential and causal aspect. It provides a base model for representing events in a broad variety of contexts and allowing the integration of possibly discrete and continuous events from different applications. In our ontology, the ForensicDomain concept stands for the notion of the domain of discourse. It is the top concept from which the remaining of our ontology is derived. The subsequent elementary concepts of event description are defined as time, location, documenting media, event interpretation, event type, action and object. In this paper, the focus of the discussion is given to two elementary concepts; object and action.

3.1 Object and Action Conceptual Model

Object conceptual model provides information about the entities involved in the event. It is characterized by two main concepts; living and material. All living entities are classified under the living concept and non-living entities are classified under the material concept. The living concept features human and animal, and human is further divided into two subclasses; person and group. Person refers to a single individual, characterized by face and body features, whereas group represents a cluster of people. Material concept is divided into three concepts, fixed object, mobile object and portable object. Fixed object represents object permanently located, build or installed such as building and road while portable object are mainly small objects such as bin, stone and bottle that are possibly spotted at the scene. Mobile object is any movable object largely vehicles.

Action conceptual model describes three types of action granularity levels known as atomic action, composite action and interaction. Atomic action e.g. walking, running, waving, etc. is the lowest granularity degree of action representation.
A composite action represents an intermediate granularity level of representation and are composed of multiple atomic actions, where all actions are carried out by the same person. For instance, action ‘smashing’ is defined as ‘walking’ and ‘hitting’ where as ‘running’ and ‘throwing’ represent ‘attacking’. Interaction is sub event that involves more than one person and consist of atomic actions and/or composite actions. This design allows us to work in different degrees of granularity and to decompose complex activity and behaviour recognition stages into simpler procedures.

### 3.2 Properties of Concepts

To address relationships between concepts in the domain, we create several concept properties to describe these relationships. We define person-to-framenumber relationship with property hasFrameNumber, person-to-action with hasAction, person-to-face with hasFace, face-to-person with isPerson and person-kicking-portableobject relationship with property isKickingA. The property hasFrameNumber is used to describe the corresponding frame number of detected objects and actions in the video footage during video analysis stage while hasAction property is used to describe an action and the person who performs it. The property hasFace and isPerson has an inverse functional characteristic and is used to link the concept of person and face. Property such as hasFrameNumber are automatically asserted in the ontology during the parsing process, whereas properties like hasAction, hasFace, isPerson and isKickingA are defined during rules specification, to describe relationships between concepts.

## 4 Event Scenario

### 4.1 Representation of Situations

The focus of our discussion is based on several examples of scenario during a riot event that give rise to a number of situations. These situations are mainly described by objects and actions involved in the scene. From various situations that occur during the event, we will focus on three situations and presents how the situations are captured in a formal language. In our representation, we make the assumption that complex events can be decomposed into simpler events and vice versa. Primitive events are the simplest type of events inferred directly from the observables in the video data.

i. **Representation of ‘PersonhasActionRunning’**

   Running is classified in atomic action class which represents the lowest granularity degree of action representation. In this statement, a relation is created between two concepts; object in the first parameter followed by action in the second, linked by ‘hasAction’ object property. Attribute of this situation may include the location of the situation and the time associated with the occurrence of the action.

ii. **Representation of ‘PersonhasActionAttacking’**

   Attacking is a composite action derived from two consecutive atomic actions, running and throwing. Similar to the first statement, the first parameter in the statement is represented by an object and the second is action, linked by ‘hasAction’ object property. Attacking action is derived through inference process. Attribute of this situation may include the location and time associated with the occurrence of the action and material object involved in the action. The sub-action may have attributes of their own, for example speed and velocity of running action while performing the attack.

iii. **Representation of ‘PersonisKickingACar’**

   In this statement, three concepts are interconnected in a single statement. Living object (person), action (kicking) and material object (car) are linked by ‘isKickingA’ object property. The objects appeared and actions performed on the scene are formulated using rules to deduce this event. Attribute for this situation may include the time and a relative location in which this event takes place.

## 5 Inference and Query

### 5.1 Inference using rules

In addition to annotations of specific events and definitions of composite properties, relations and events, we also specify inference rules that allow us to draw conclusions from what we recognize in the data. A rule-based system is used to store and manipulate knowledge to interpret information in a useful way. One of the great advantages of having situations represented in a formal language is that facts of knowledge that are not explicitly stated can be derived using an inference engine. As discussed before, every instance that represents person, face, actions and material objects detected in the video analysis module has ‘hasFrameNumber’ property to represent the frame number in the video footage. Thus, new relations between face, person, action and material objects can be derived using this asserted knowledge through rules inferencing process. For instance, we may classify that person, face and action detected on the same frame represents a ‘person’ having the face ‘face’ doing a certain ‘action’. The following simple Horn clause rule expresses this condition:

If

\[
\text{hasFrameNumber (Person, X) and hasFrameNumber (Running, X)}
\]

Then

\[
\text{hasAction (Person, Running)}
\]

If

\[
\text{hasAction (Person, Running) and hasAction (Person, Throwing)}
\]

Then

\[
\text{hasAction (Person, Attacking)}
\]

These rules have two advantages. First, they are more general, applying to arbitrary person, face and action instances. Second, they are much more succinct. The second rule gives an example of the power of rules where event attacking is inferred from existing atomic actions that was asserted in the knowledge base.

Both Horn clause rules have been implemented using Semantic Web Rule Language (SWRL) supported by the Pellet reasoner [15]. When the rule is submitted to the Pellet reasoner
along with the facts from the asserted axioms, the newly inferred fact is automatically derived and added to the collection of facts about the situation. In this case, a new logical consequence ‘PersonhasActionAttacking’ is inferred and added into the knowledge base.

5.2 Querying Formal Representations

The difference between queries to a database and queries to an OWL knowledge base is that the answer to a knowledge base query may include facts that are inferred as well as facts that have been explicitly asserted. We use the SPARQL Protocol and RDF Query Language (SPARQL) language to express these queries. The purpose of SPARQL is to provide a formal language to ask meaning-driven questions in the knowledge base.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX untitled-ontology-14: <http://www.semanticweb.org/farhan/ontologies/2017/untitled-ontology-14#>

SELECT *
WHERE {
  ?ind rdf:type untitled-ontology-14:Attacking .
  ?ind untitled-ontology-14:hasFrameNumber ?hasFrameNumber
}
ORDER BY ?ind ?hasFrameNumber
```

In this SPARQL query example, we intended to acquire details of individuals who perform the attacking action in the video. After the reasoning process in the ontology is executed, a list of individuals which are related to attacking action defined in the rules are obtained. The query results were then correlated with a database of images created during video analysis process to identify the person’s face and body image. This new information will assist users to identify possible crime and perpetrators and understand the event semantics.

6 Experimental Results

The sample of event sources used in this paper is obtained from CCTV footage of 2011 London riot event provided by Scotland Yard. The video footage represents violent public disturbance behaviour, vandalism and major destruction of properties in various locations. 2.5GB of video consisting of 10,855 frames were carefully analysed and utilized to validate the developed system framework. The experimentation stage is geared towards the identification of suspect and event that have a high level of abstraction and semantics in the video footage. Through a precise combination of data and by utilizing rules and semantic reasoning, we demonstrated the outcome of our approach. Fig. 2 shows examples of person, face and car detection, and throwing, kicking and running action recognition performed on the video footage of riot event.

Our ontology consists of seven elementary classes, 106 subclasses and 10 object properties. Additionally, we populate 56,734 individuals into the ontology, which represent instances of objects and actions obtained from the video analysis module. We also create seven rules to define and deduce actions and events that happens in the video footage. In this paper, we demonstrate three event situations, ‘PersonhasActionRunning’, ‘PersonhasActionAttacking’ and ‘PersonisKickingACar’. The first situation demonstrates the use of SWRL rules to link two asserted concepts from the knowledge base. The second situation presents the implementation of rule-based reasoning to deduce new knowledge into the knowledge base and the third situation exemplify three asserted concepts being linked together to represent an important event semantic. The SPARQL query is used to acquire all the details.

6.1 Person and Face Detection and Action Recognition

Both person and face detections were implemented using the OpenCV image processing library. For person detection, the HOG person detector classifies 34,125 instances of person from 10,855 frames in the video. On the other hand, the Haar cascade classifier successfully classifies 12,408 instances of face using the face detection algorithm. Both methods
demonstrate a good real-time detection results with lower computational cost for this application regardless of its outmodedness. For action recognition, we annotated 36 instances of running, 34 instances of throwing and three instances of kicking consisting of tenth of frames representing each action instances. We also annotated 45 instances of cars, 59 frames representing fire and two groups of police during the event.

During the detection and annotation process, a text file is created to record the corresponding frame number of the detected objects and actions and their video capture time stamp. The video capture time stamp helps to validate frame synchronizations especially when different methods of object detection and annotation are applied to the same video.

6.2 Results for Inference and Query

The suspects in the event were identified through a ‘matching process’ which associates the body and facial image of a person with the performed action. The knowledge integration is made using semantic rules that chain together asserted data that have the same frame number. By utilizing these data, information about the suspect’s face and body features, clothing and performed actions could be established. We perform an experiment to demonstrate suspect identification through actions executed in the video.

6.2.1 Query Result for ‘PersonhasActionRunning’

As previously mentioned, we annotated 36 instances of running performed by different individuals in the video footage. The instances are label as Running_1, Running_2, Running_3 and so forth, where each label represents one running action consisting of a number of consecutive frames. Note that at this point, we do not have any clue about the person who execute these actions. We invoke Pallet reasoner to perform reasoning process on our ontology, make inference and assert new knowledge in the ontology based on the rules that we have designed. We then perform queries to extract inferred facts from the knowledge base. Based on queries of Running_1 action to the inferred OWL ontology, 133 persons and 7 faces were detected from 22 frames. We observed and analysed related images and obtained a sequence of ‘a running person’ from the identified frames. Based on the observation, there was one individual repeatedly appeared in the collection of images and is identified to be performing Running_1. The same process is being made to identify individual who performs other actions. Fig. 3 shows the selected frame sequence of Running_1 action, showing the person who performs it.

6.2.2 Query Result for ‘PersonhasActionAttacking’

Attacking is defined as a person who performs running and throwing actions. Attacking action was inferred from asserted input through rule-base inference process. The SPARQL query returns 667 inferred results related to attacking action representing person and faces involved during the riot event. We analyse the data and identified 12 occasions of attack happening during the riot. We present three sequences of attack performed by different individual in Fig. 4.

6.2.3 Query Result for ‘PersonisKickingACar’

There are three instances of kicking and 45 instances of car being annotated from the video footage. We are interested to know the person who performs kicking and which car that has been kicked. In order to do that, we create a rule which link the kicking action with the person who appeared during the action execution and the car that was present during that particular time. Based on the query results, we identify that Car_11 has been kicked by a man who was wearing blue and white stripe hoodie as shown in Fig. 5. Snippet of query result is also shown in the same figure. Note that the car has been cropped from the image by the person detector algorithm.

Fig. 3. Sequence of Running_1 actions extracted from image database based on query results

Fig. 4. Three attack action sequences during the riot obtained from rule-based inference process

Fig. 5. Query result for ‘PersonisKickingACar’ and the sequence of images to illustrate the event
7 Conclusions

The results of the experimentation stage demonstrate that ontology-based video analysis offers a promising result in analysing long content surveillance videos. By representing information from low-level video analysis in ontologies and incorporating high-level reasoning, rules and queries, benefits of using ontology-based knowledge representation and reasoning approach for video analysis are limitless.

The need for understanding event semantics is crucial, especially in safety and security domain where there are a vast number of video surveillance recordings that require a thorough implementation process, thus preventing an effective investigation from being carried out. To fulfil this need, we have designed a framework which supports both lower level feature extraction and high-level semantic representation. This framework enables prominent features in a CCTV video to be represented in an ontology, while the process of understanding event semantics is achieved through the implementation of rule-based reasoning and queries. We have demonstrated and validated the approach using sample data of 2011 London riot provided by Scotland Yard.

References

[1] N. Jenkins, “Video Surveillance Camera Installed Base Report - 2015,” 2015.
[2] SDM, “Rise of surveillance camera installed base slows,” SDM Magazine, 2016.
[3] British Security Industry Association, “The Picture is Not Clear: How many CCTV surveillance cameras in the UK,” no. 195, pp. 1–7, 2013.
[4] E. M. Tapia, T. Choudhury, and M. Philipose, “Building reliable activity models using hierarchical shrinkage and mined ontology,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3968 LNCS, pp. 17–32, 2006.
[5] U. Akdemir, P. Turaga, and R. Chellappa, “An ontology based approach for activity recognition from video,” Proc. 16th ACM Int. Conf. Multimed., pp. 709–712, 2008.
[6] Y. Zhang, W. Liu, N. Ding, X. Wang, and Y. Tan, “An Event Ontology Description Framework Based on SKOS,” 2015 IEEE 12th Intl Conf Ubiquitous Intell. Comput. 2015 IEEE 12th Intl Conf Auton. Trust. Comput. 2015 IEEE 15th Intl Conf Scalable Comput. Commun. Its Assoc. Work., pp. 1774–1779, 2015.
[7] R. Nevatia, J. Hobbs, and B. Bolles, “An Ontology for Video Event Representation,” 2004 Conf. Comput. Vis. Pattern Recognit. Work., pp. 1–10, 2004.
[8] J. C. SanMiguel and J. M. Martinez, “Dynamic video surveillance systems guided by domain ontologies,” Crime Detect. Prev. (ICDP 2009), 3rd Int. Conf., pp. 1–6, 2009.
[9] M. Y. K. Tani, A. Lablack, A. Ghomari, and I. M. Bilasco, “Events detection using a video-surveillance ontology and a rule-based approach,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8926, pp. 299–308, 2015.
[10] J. C. SanMiguel, J. M. Martinez, and A. Garcia, “An Ontology for Event Detection and its Application in Surveillance Video,” Adv. Video Signal Based Surveillance, 2009. AVSS ’09. Sixth IEEE Int. Conf., pp. 220–225, 2009.
[11] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” Proc. - 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005, vol. I, pp. 886–893, 2005.
[12] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” Comput. Vis. Pattern Recognit., vol. 1, pp. 511–518, 2001.
[13] M. Hacherouf, S. N. Bahloul, and C. Cruz, “Transforming XML documents to OWL ontologies: A survey,” Artic. J. Inf. Sci., vol. 41, no. 2, pp. 242–259, 2015.
[14] N. Bikakis, C. Tsinaraki, N. Gioldasis, I. Stavrakantonakis, and S. Christodoulakis, “The XML and semantic web worlds: Technologies, interoperability and integration: A survey of the state of the art,” Studies in Computational Intelligence, vol. 418. pp. 319–360, 2013.
[15] H. Bohring and S. Auer, “Mapping XML to OWL Ontologies,” Leipziger Inform. Vol. 72 LNI, pp. 147–156, 2005.
[16] U. Westermann and R. Jain, “Toward a common event model for multimedia applications,” IEEE Multimed., vol. 14, no. 1, pp. 19–29, 2007.