Semantic Web 0 (2017) 1
IOS Press

On the use of semantic technologies for video analysis

Luca Greco, Pierluigi Ritrovato, Mario Vento
Dept. of Information and Electrical Eng., and Applied Mathematics (DIEM), University of Salerno
Via Giovanni Paolo II, 132, 84084 Fisciano (SA) - Italy

Abstract. The rapid proliferation of video recording devices has led to a huge explosion of contents, determining an ever increasing interest towards the development of methods and tools for automatic video analysis and interpretation. Through the years, the availability of contextual knowledge has proven to improve video analysis algorithms' performances in several ways, although the formal representation of semantic content in a shareable and fusion oriented manner is still an open problem. In this context, an interesting answer has come from Semantic technologies, that opened a new interesting perspective for the so called Knowledge Based Computer Vision (KBCV), adding new functionality, improving accuracy, and facilitating data exchange between video analysis systems in an open extensible manner. In this work, we propose a survey of the papers from the last fifteen years, back when first applications of semantic technologies to video analysis have appeared. The papers have been analyzed under different perspectives leading to the definition of a taxonomy of the different approaches and the semantic web technology stack adoption. As a result, some insights about current trends and future challenges are provided too.

Keywords: Computer Vision, Semantic Web, Video Analysis, OWL

1. Introduction

Nowadays, with the rapid proliferation of video recording devices and the related explosion of generated contents, we assist to an increasing interest in the development of methods and tools for automatic video analysis and interpretation [72]. Actually, several video based applications - such as video surveillance, road traffic control, sports events detection - still require a strong human intervention when a semantic understanding of contents is needed to detect (and eventually retrieve) objects, actions or events within a video stream. Manual analysis of video sequences is a very time consuming task and it often leads to inaccurate results due to the "video blindness" phenomenon that affects human operators when watching video for an extended period of time. In the video surveillance domain, for example, it has been stimated that an operator can miss up to 95% of scene activities after only 22 minutes of analysis [1][2].

In the last years, great efforts by the computer vision community have been devoted to the development of robust and reliable algorithms that, starting from raw pixels (classic bottom-up approach), aim at accomplishing video analysis tasks at different levels:

- *Low-level* video analysis methods address the ability to find the image regions corresponding to objects of interest *(detection)* and then track them across different frames while maintaining the correct identities *(tracking)* [84,22,24].
- *Mid-level* video analysis methods face the problem of recognizing simple or "atomic" events or activities such as loitering, falls, direction changes, group formations and separations [14].
- *High-level* video analysis methods concentrate on the detection of "complex" events or activities such as aggressions, fights, pickpocketing, thefts, general "suspicious events", activities of daily living (especially in the healthcare domain), vehicle stealing and so on [46][77].

While low-level processing aims at generating feature descriptions to summarize characteristics of data in a quantitative way, high-level processing is more

related to the interpretation and reasoning with visual data: it takes features descriptors as input and generates abstract, qualitative descriptions about contents, addressing the so called *semantic gap* problem [68]. Through the years, high-level methods have been implemented using various forms of artificial intelligence, from symbolic knowledge representation, based on hierarchical structures using semantic networks [41] or more traditional object oriented data models [35], up to rule-based systems [4]. Many works have been focused on improving the ability to search and retrieve specific contents from large repositories. Anyway, all the proposed approach share a common denominator: employing some kind of *a priori* domain knowledge, that can be also regarded as *context*. As demonstrated in [47,7,60,28] the availability of contextual knowledge for image and video analysis can improve algorithms performance in several ways: for example information about the scene environment (structures, static objects, illumination, positions) could help preventing detection errors (spurious blobs), tracking errors (occlusion splits, id-switches, ...) or recognize objects in a scene (a car on a road, a boat in the sea). Successful examples of improvement provided by the intelligent use of contextual information are those in speech and optical character recognition. Indeed, in both cases, by recognizing only basic elements like phonemes, syllables and by reducing the space of the possible interpretation hypotheses according to a dictionary, a breakthrough of the recognition capabilities in both systems has been achieved. An answer that is gaining increasing interest in the scientific community comes from the Semantic Web (SW) technologies, that address the problem of formally representing semantic content and supporting automated reasoning, sharing and integration of different sources. SW opened a new interesting perspective for Knowledge Based Computer Vision (KBCV)[78], facilitating data exchange between video analysis systems in an open extensible manner. In this work, we aim to survey the papers from the last fifteen years, back when first applications of semantic technologies to video analysis have appeared.

1.1. Aims and Methodology

The survey is aimed to increase the attention of researchers and practitioners working on video analysis towards the potentiality offered by semantic web technologies for improving the performance of existing solutions and enabling advanced video analytic functionalities. However scholars active in SW research could also gain useful hints about technology usage and applications for further development.

With this aim in mind, we surveyed a selection of the most relevant papers in the field, highlighting *which* semantic technologies are used, *how* they are applied and *what* results have been achieved. This allowed us to compose a picture of semantic web technology usage in the different video analysis domains and to identify open challenges. For each paper, a careful analysis of the referenced works has been carried out to identify new potential candidates and retrieve surveys already available in literature. Several papers using knowledge based approaches and SW to address only traditional image analysis tasks (segmentation, region labeling, object recognition, etc.) have not been included in this survey. The selection process lead to the identification of 77 papers: 31 conference papers, 41 journal papers, 4 book chapters, a W3C specification. The remainder of this work is organized as follows: in the next section an overview of the Semantic Web and the related technological stack is presented. Section 3 provides a review of past surveys about knowledge based approaches in the research areas of computer vision, image understanding, multimedia content analysis. Section 4 is the core of the paper: here a detailed review of the selected papers is provided. In Section 5 an analysis of the works and some considerations about the adoption of semantic web technologies for video analysis are provided. Section 6 draws the conclusions.

2. A quick overview of Semantic Web

The Semantic Web [8] has been introduced in 2001 as an extension of classic Web (the web of data) with the aim of representing resources in a *machine-understandable* way by means of formal descriptions and annotations that improve sharing and reuse across different systems and applications. In this context, ontology has become a standard for the “description (like a formal specification of a program) of the concepts and relationships that can formally exist for an agent or a community of agents" [34]. W3C recommended the Ontology Web Language (OWL) for ontology definition. OWL provides powerful and expressive constructs and Description Logic reasoning services. First specifications for the description of resources and the representation of properties were provided by the *Resource Description Framework* that use Uniform Resource Identifiers (URIs) for the identification of de-

scribed resources. RDF semantics are basically defined by: *RDF Schema (RDFS)*, that provides constructs to divide resources into classes and properties allowing to define subclasses and sub properties; *OWL* has a more expressive power with respect to RDFS and supports the definition of transitive, functional and inverse properties, equivalent classes and properties and cardinality restrictions. RDF contents are represented and stored as triples and can be effectively managed using *SPARQL*[1] (recoursive acronym of SPARQL Protocol and RDF Query Language) language, that allows to perform complex queries including union, optional query parts, and filters over RDF graphs. SPARQL 1.1[2] adds a number of new features to the query language, including subqueries, value assignment, path expressions, or aggregates - such as COUNT, etc..; it also includes a specification of *SPARQL Update* that supports, in addition, the conditional insertion and removal of triples from an RDF store. RDF concepts can be also manipulated programmatically by using different tools (OWL API [39], Jena [51]) available for different programming languages.

2.1. From OWL to rule support

In OWL the representation involves sets of objects (*concepts*) that can be related to themselves (or to datatypes) through relationships (*roles*), each object being an *individual*. Essentially based on Description Logics (DL), OWL has been proposed with 3 variants of different expressive power: the first two, named OWL Lite and OWL DL, are adaptation of DL $\mathcal{SHIF}(\mathcal{D})$ and $\mathcal{SHOIN}(\mathcal{D})$ respectively; OWL Full has full compatibility with RDFS but it is rarely implemented since it is undecidable. OWL2 [37] tries to compensate some drawbacks of the first release of OWL presenting insufficient capabilities to be applied to several practical applications [30]. OWL2 maintains backward compatibility with previous version but also extends OWL both in terms of syntax and expressiveness[3]. As for OWL, different profiles have been proposed with a trade-off between expressiveness and reasoning efficiency. In order to avoid adoption failures as for OWL Lite, the OWL2 profiles are built identifying maximal OWL2 sublanguages that are implementable in a polynomial time. In order to achieve this objective the main sources of intractability (i.e. disjunction, or negation together with conjunction, restrictions on maximal criminality, etc.) has been removed from the OWL2 profiles. As described in [37] the three OWL2 profiles are:

- OWL2 EL, that is particularly useful in applications employing ontologies that contain very large numbers of properties and/or classes. This profile captures the expressive power used by many such ontologies and is a subset of OWL 2 for which the basic reasoning problems can be performed in time that is polynomial with respect to the size of the ontology.
- OWL2 QL, that is meant for applications that use very large volumes of instance data, and where query answering is the most important reasoning task. In OWL 2 QL, conjunctive query answering can be implemented using conventional relational database systems. Using a suitable reasoning technique, sound and complete conjunctive query answering can be performed in LOGSPACE with respect to the size of the data (assertions).
- OWL2 RL, that is meant for applications that require scalable reasoning without sacrificing too much expressive power. It is designed to accommodate OWL 2 applications that can trade the full expressivity of the language for efficiency, as well as RDF(S) applications that need some added expressivity. The ontology consistency, class expression satisfiability, class expression subsumption, instance checking, and conjunctive query answering problems can be solved in time that is polynomial with respect to the size of the ontology.

Reasoning services are based on DL and adopt the well known *open world assumption*, where statements about knowledge that are not included in or inferred from the knowledge explicitly recorded in the system may be considered unknown, rather than wrong or false. Different tools and algorithms are available to perform reasoning (Pellet, Fact++, Hermit [53]) and services like *instance checking, subsumption, consistency, realisation* and *satisfiability* are typically provided. The integration of OWL with rules allows to overcome some limitations on expressiveness (for example the tree model property) especially in problems where structures cannot be effectively captured in atemporal manner (activity recognition). The *Seman-*

[1] SPARQL 1.0 recommendation was released by W3C on January 2008.

[2] Officially released as W3C recommendation on March 2013.

[3] see https://www.w3.org/TR/2012/REC-owl2-new-features-20121211/ for more details

tic Web Rule Language (SWRL) [40] has been introduced to this aim, allowing to define rules to be interpreted under the first logic semantics. In particular, DL-safe rules [54] were defined to maintain decidability by acting only over known individuals (most reasoners only implement a subset of SWRL to preserve DL-safety). In 2011, *SPIN* (SPARQL Inferencing Notation) has been introduced as a W3C Member Submission[4] providing meta-modeling capabilities that allow users to define their own SPARQL functions and query templates and includes a ready to use library of common functions. SPIN allows to specify constrains using the property `spin:constraint` (to link a class with constraint checks that all instances of the class need to fulfill) and add rules through the property `spin:rule` (to link a class with inferencing rules that construct new information from the statements about the instances). Since rules are expressed in SPARQL, they can run directly on RDF data without a need for "materialization". SPIN Templates also make it possible to define such rules in higher-level domain specific languages so that rule designers do not need to work with SPARQL directly. Thanks to the availability of development tools, SPIN has achieved a large adoption in industrial applications to represent SPARQL rules and constraints on Semantic Web models.

The relevance to define constrains on RDF data, needed in many practical applications, lead the W3C to create the RDF Data Shapes Working Group having in charge to produce W3C recommendation for describing structural constraints and validate RDF instance data against those constrains. In July 2017 The Working group released the recommendation for a new language for expressing constraints named Shapes Constraint Language (SHACL). The design philosophy behind SHACL is to provide a high-level vocabulary called SHACL Core and an extension mechanism that allows to associate SPARQL queries with classes and other resources, similar to how SPIN worked.

3. Previous surveys

In the past years, different interesting surveys have been proposed to discuss knowledge based computer vision in a broader perspective or to review the use of ontologies for image annotation and retrieval, context modeling, event and human activity recognition and pervasive computing. For example, Fiorini *et al.* [23] review the area of Knowledge-Based Computer Vision focusing on knowledge representation aspects. Their analysis highlights the *what* part (knowledge necessary for KBCV) and the *how* part (representation formalism, model structure and symbol grounding), both indicating ontologies as the most pervasive concept in recent trends. The works of Hanbury [36] and Dasiopoulou *et al.* [20] present an overview of the state of the art in image and video annotation methods and tools. They show how domain specific ontologies can support tools for describing contents or representing low-level descriptors. Sjekavica *et al.* [67] present an overview of the most common ontologies in multimedia domain and for annotation of multimedia content. In particular, their work compares four ontologies for semantic annotation - COMM (Core Ontology for Multimedia Annotation), Ontology for Media Resources 1.0, M3O (Multimedia Metadata Ontology), LSCOM (Large Scale Concept Ontology for Multimedia) - evaluating supported types of multimedia content, language, used design patterns, number of classes and properties. A survey of content based image retrieval with high level semantics is proposed by Liu *et al.* [48] that emphasize the use of object ontologies to provide a qualitative definition of high-level query concepts (color, position, size, shape) while the work of Kannan *et al.* in [43] exploits approaches for automatic video annotation with higher level concepts, complex events retrieval, automatic representation of image/video based on semantics and annotation of video based on rules. The work of Bettini *et al.* [12] describe ontologies and semantic web technologies as good candidates to meet the requirements of *context modeling* and reasoning techniques, being capable to handle effectively information types and their relationships with the uncertainty of context information. A comparison of available methods for context modeling is also provided. Turaga *et al.* [74] discuss methods to model actions with simple and complex dynamics and, referring to knowledge and logic-based approaches, emphasizes the important role of ontology in standardizing activity definitions, allowing easy portability to specific deployments and enabling systems interoperability. The use of ontologies for Human Behavior Recognition is also reviewed in [61], where Rodriguez *et al.* also emphasize ontological models as the most promising tools for their flexibility and representation capabilities as long as for reasoning and information sharing objectives. They present a set of

[4] http://www.w3.org/Submission/2011/SUBM-spin-overview-20110222/

upper ontologies designed to represent human activity, as well as domain ontologies that can serve the same aim in context-aware intelligent environments. In [59] Poppe *et al.* provide a comprehensive review of the advantages of using Semantic Web Technologies in video surveillance domain exploiting the need for describing video analytics with a common metadata format and discussing semantic reasoning by means of rules. A survey on semantic web technologies in pervasive computing have been proposed by Ye *et al.* [83]. They focus on information modeling and reasoning along with streaming data and uncertainty handling for exploiting the use of Semantic Web technologies to represent contextual information and enhance data exchange. Another interesting survey has been recently proposed by Onofri *et al.* [56], where particular attention is devoted to the problem of activity recognition in video streams as addressed by systems incorporating a priori knowledge and context information. To the extend of our knowledge, this is the first work focusing on semantic web technologies applied to video analysis problems.

4. Semantic technologies for video analysis

As described in section 2 the semantic web technology stack is built around two central pillars: the ontology (and related representation schemas) and reasoning. For that reason, the following sections present, in different video analysis areas, the use of ontology for formally representing information and annotating contents; moreover, we will explore the use of reasoning technology for inferring new knowledge at different levels.

4.1. Low-level analysis: detection, segmentation and tracking

A first interesting work is presented by Dasiopoulou *et al.* [19]: here a multimedia ontology encodes the semantic concepts of the domain with qualitative attributes, low-level features, object spatial relations and processing methods, while F-logic rules are defined to govern the application of analysis methods. The proposed approach has been experimented to Formula One, soccer, and beach vacations videos producing interesting results. The paper is particularly interesting since it demonstrates different uses of SW technologies. In particular, ontologies (even using a less expressing language i.e. RDFS in spite of OWL) are used for both representing relevant semantic concepts for the domain in terms of object classes (e.g. for the formula one domain we have car, road, grass, etc.), feature classes (like size, color, motion, etc.) together with classes describing specific properties like color models, etc.. The ontology is also used for representing the properties of the different algorithms to be applied for content analysis (like k-means, motion clustering, etc.). Instances of the ontology are connected through properties (i.e. a car hasFeature a specific size, or hasColor a color, etc.). The reasoning part, implemented through F-Logic rules, uses the information available in the ontology for deciding which algorithm to use according to the characteristics of the content to be analyzed (low level features). The technological choices for knowledge representation (RDFS) and rule-based reasoning (F-Logic) are imposed by the development environment used for the framework (ontoEdit and OntoBroker respectively). Through the experimentation, the authors demonstrate a significant improvement with respect to traditional approaches for region identifications (i.e. for car identification in formula one domain there is an improvement of 31% on correct identification and only 7% of misses instead of 33% of traditional ones). The downside highlighted by the authors is the collection and analysis of information to be included in the ontology in particular with respect to the features attributes like color, homogeneity, motion, etc.

On the same line is the paper proposed by García *et al.* [27] that aims at defining a knowledge-based framework for video object segmentation where relationships among analysis stages are exploited. The key idea is to provide a rich description of the scene at low, mid and high semantic levels through an ontology; in particular, a first stage comprise the low-level analysis modules (background subtraction, short-term change detection, point or region tracking, motion field estimation...) which are provided with a structure to collaborate and achieve consistent results with the considered application context; results of these algorithms are modeled as classes in the analysis ontology, representing the lowest abstraction level for occurrences in the scene. The following stages, starting from results of low levels, build Point Hypothesis Maps (PHM) and Region Hypothesis Maps where the most probable states (occurrences) of each point and region respectively are coded according to the ScenePoint and SceneRegion hierarchies of the analysis ontology. A feedback path allows to evaluate the quality of the results at each stage repeating iteratively the association

process until consistency is reached. As a measure of the quality of the whole system, the authors provide quantitative background segmentation results obtained by comparing the system with other state of art approach (Mixture of Gaussian and Bayesian approach) on VSSN06 dataset showing a minimum improvement of 50%. Gomez *et al.* [29] propose a computer vision framework that relies on an ontology based representation of the scene and combines contextual information and sensor data. The key aspect is the application of logical reasoning starting from data coming from a classical tracker with the aim of constructing an ontological model of the objects and the activities happening in a observed area. Reasoning procedures are used to detect and predict tracking errors, sending feedback to the tracker in order to attune the low-level image-processing algorithms. The proposed system relies on two main modules: a general tracking layer and a context layer. The general tracking layer is a software program that performs movement detection, blob-track association, track creation and deletion as well as trajectory generation. The contextual layer receives data from the general tracking layer producing as a result a high level interpretation of the scene together with a feedback for the general tracking layer (a set of recommendations to be performed). In particular, the ontology models Camera Data, Tracking Data, Scene Objects, Activities, Impact and threats, Feedback. The authors use RACER reasoner (with new RACER Query Language - nRQL) for scene interpretation since it allows abductive reasoning. In fact, abductive rules are defined in the framework to interpret what is happening in the scene from the basic tracking data (creating a new Person instance when a track bigger than a predefined size not already assigned is detected in the image). The authors show the use of the framework in a surveillance application using the PETS2002 dataset. An ontology-driven approach for the semantic analysis of video is also proposed in [58] where a multimedia ontology for segmentation is designed. The knowledge architecture is actually made of four modules: the Core Ontology, the Mid-Level Ontology, the Domain Ontology and the Multimedia Ontology. While the Core and Mid-Level ontologies (based on DOLCE [26]) contain specification of domain independent concepts, Multimedia Ontology is defined to model the content of multimedia data and includes algorithms for processing the content. The Information Object (IO) design pattern, previously proposed for extending the DOLCE core ontology, was extended so that the Multimedia Information Object (MMIO) could combine DOLCE IO pattern with MPEG-7 standard for the representation of media content and multimedia features. As a sub-class of MMIO, the Visual Information Object is exploited since it carries the visual information. The whole ontology infrastructure is linked with the signal domain by a combined use of a temporal and a spatial segmentation algorithm, a layered structure of Support Vector Machines (SVMs)-based classifiers to associate the keyframe with the appropriate concept. Since the association mechanism requires a fusion step to make decisions starting from local and global features, a Genetic Algorithm is implemented to address the optimization problem. These processing methods support the decomposition of visual information and the detection of the defined domain-specific concepts. In particular, the performances of the procedure have been assessed in the domain of disaster news videos, where the ability to achieve an accurate keyframe-concept association was evaluated with interesting results (about 98% accuracy in fire detection for keyframe-concept association using global-level information). The use of semantic technologies has also been exploited in conjunction with traditional bottom-up video analysis methods with the aim of detecting and correcting common errors associated with the low-level tracking step. In particular, Greco *et al.* demonstrate in [31] how an hybrid solution, that involves the semantic annotation of the output of a standard tracking algorithm, can improve the overall system performances on standard datasets (PETS09) thanks to the use of a custom defined tracking ontology and a set of SPIN functions/rules that help identifying common mistakes such as false positives, misses and ID switches.

4.2. Mid and high level analysis: Visual event and activity recognition

One of the first proposals for representing video events through ontology is the Video Event Representation Language (VERL) [25]. The key contribution is the idea of modeling events as composable so that complex events can be constructed by composing simpler activities in a hierarchical framework. The lowest-level events are *primitive* events while composite events are defined by compositions of lower-level events by using operations such as sequencing. For example, an event involving a person getting out of a car and going into a building is described using the following sequence: opening car door, getting out of car, closing car door (optional), walking to building, opening building door, and entering building. More com-

plex composition operations such as iteration and alternation as well as composite events containing multiple simultaneous events (multithreaded) can be defined. Temporal relationships between subevents can be handled by using Allen's interval algebra, which defines qualitative relations between intervals. The authors provide a detailed example of VERL application to the description of a "tailgating" event (an action that involves gaining access to a secure facility by entering behind an authorized individual) in surveillance video, which also demonstrates the power of such a representation. The VERL representation and the design of ontologies for visual activity recognition starting from general design principles have been also discussed by Akdemir *et al.* in [3]. They mention qualitative evaluation principles (such as clarity, coherence, extendibility...) [34] and provide several examples from existing ontologies (Cruise Parking Lot, Tail-gate definition in Perimeter and Internal Security, Shoplifting in Store Security). As a main contribution, some improvements are proposed with respect to the above mentioned principles and a discussion on granularity issues in ontology design is provided focusing also on the role of context in determining the ontological complexity. The authors demonstrate the effectiveness of the approach with a comparison between the cases of Bank Surveillance (a controlled environment where a fine granularity is not needed) and Airport Surveillance, where the system achieves from 80% to 100% accuracy .

The idea of representing complex events as a sequence of simple events can be found also in other works such as in [69] where Snidaro *et al.* propose an ontology to describe contextual knowledge in video surveillance domain. Ontology is organized into three branches of classes that model background, entities and events. The event class is specialized into subclasses that describe simple events, spatial events and transitive events, allowing to show how complex events can be described through simple events sequencing. They formalize a set of rules in SWRL language to perform event detection and provide some preliminary qualitative results by evaluating such rules (for example to detect a stop of a vehicle in an unauthorized place for an amount of time over a specified threshold value) through the Jess rule engine[5]. In the same way, SanMiguel *et al.* [62] propose an ontology for representing the prior knowledge related to video event analysis.

Such a knowledge is described in terms of scene-related entities (Object, Event, and Context), system-related entities (Capabilities, Reactions...). The key contribution of the work is the integration of different types of knowledge in an ontology with the purpose of detecting the objects and events in a video scene. In addition, it is shown how the complex events can be described by simple events providing a domain extension for the Underground Video-surveillance domain and mapping the knowledge described in the ontology to a visual analysis framework dependent on the defined events to be detected. Experiments were carried out on several sequences from the the i-LIDS dataset for AVSS2007 and the PETS2006 dataset achieving from 74% to 79% precision.

A different approach to the application of ontological languages to video event detection has been proposed by Town [73]. Here the innovative aspect lays in the possibility to learn effective high-level state and event recognition mechanisms from a set of annotated training sequences by incorporating syntactic and semantic constraints represented by an ontology. The author uses video sequences and ground truth from the CAVIAR project [6] to define an ontology of visual content descriptors arranged in a hierarchy of scenarios, situations, roles, states, and visual properties and then proposes to train Bayesian networks to perform inference over ontology terms.

The use of ontology and rules can improve video semantic analysis also by providing an integrated representation of low-level features and video content analysis algorithms as shown by Bai *et al.* in [5]. They use a domain ontology (described in OWL language) to define high level semantic concepts and their relations within the considered domain while providing rules in Description Logic which describe how algorithms for video analysis should be applied according to different perception content and low-level features. Furthermore, temporal Description Logic is used by the authors to describe the semantic events and a reasoning algorithm is proposed for events detection, demonstrating performances in a soccer video domain and achieving from 91,7% to 100% precision. In the same way, Zaidenberg *et al.* [85] propose an ontology based framework for event recognition called *ScReK (Scenario Recognition based on Knowledge)* with a

[5]http://www.jessrules.com/docs/71/ last access on September 2016)

[6]EC Funded CAVIAR project/IST 2001 37540, http://homepages.inf.ed.ac.uk/rbf/CAVIAR/ last access on September 2016.

novel group tracking approach. The ontology models the events and vision primitives defined by domain and vision experts, while the application domain knowledge to be encoded in the ontology is described by means of a custom proposed declarative language. In particular, an ontology composed of 45 generic event models (re-usable in any applications with groups) and 4 specific event models (defined with the help of metro surveillance staff) has been used for the evaluation of the group tracking approach using both real and standard datasets (CAVIAR) obtaining up to 96% precision. The work in [64] also shows a distributed framework for video analysis that automatically estimates the optimal workflow needed to analyze different application domains. The key aspect is to describe the considered domain, the capabilities of analysis and the user preferences by means of a hierarchical ontological semantic representation. The relations between the semantic descriptions allow to select the most appropriate tools available (algorithms and detection procedures). In particular, a rule-based approach is applied to extract the entities relevant to solve the analysis problem and compute their execution order. The selection process for specific algorithm implementations (when multiple choices are available) is modeled as a constraints satisfaction problem (CSP), so that an *automatic workflow* composition is actually obtained. The experimental stage is carried out in four domains, representing real scenarios for detecting abandoned objects in video surveillance. For each domain, sequences from standard datasets have been selected (AVSS2007, PETS2006, CANTATA and HERMES). Results show that the method can obtain up to 80% precision in event detection. The authors emphasize as main advantage of the framework the integration of ontology-based descriptions and video analysis tools so that any domain properly described by the ontology can be analyzed; the approach is also suitable for distributed settings due to its scalable nature. Tani *et al.* [44] use an ontology based-approach to detect single/multiple objects events through a set of SWRL rules. The scene is described according to the concepts modeled in the ontology. Once a video analysis module extracts the blobs from the video stream using low level features, the bounding box that enclose them are provided as input to the semantic module that instantiates the object and the properties within the ontology. By using a set of SWRL rules, the reasoner first classifies the bounding boxes to identify their meaning (Person or Group) and then select the appropriate event class by means of another set of rules. To represent the scene, a manual segmentation is performed; then blobs are extracted by grouping foreground pixels obtained from a background subtraction algorithm. The authors use the PETS2012 as a case study to depict qualitatively the efficacy of the approach in detecting walking events, group running, group formation and splitting.

Some works make large use of ontology to address the problem of human activity recognition in daily living. For example, Chen *et al.* [17] introduce an approach to activity recognition based on the use of ontological modeling, representation and reasoning, by analyzing in particular the nature and characteristics of Activities of Daily Living (ADLs) and modeling the related concepts through ontologies. The authors describe the algorithms of activity recognition making full use of the reasoning power of semantic modeling and representation. The work shows that ontological ADL models are very flexible and can be easily customised, deployed and scaled up; a simplified application scenario, the recognition of “MakeDrink" ADL, is used to demonstrate the proposed approach, its implementation and operation. In the same way, Riboni *et al* [60] define the formal semantics of human activities by means of a novel OWL2 activity ontology and use reasoning modules to recognize that a user is performing a certain activity by aggregating sensor data and information about people and objects. They also exploit OWL2 operators in order to represent most rule-based activity definitions by ontological axioms, preserving decidability and formal semantics. A real-world dataset of sensor readings, annotated with the activities performed by a person living in a smart-home for 28 days is used to test performance: eight ADLs are monitored, including having dinner, toileting, showering and sleeping obtaining up to 80% accuracy . Recently, Meditskos *et al* [52] have also proposed an ontology-based hybrid framework for activity recognition in Ambient Assisted Living (AAL) environments. The proposed solution combines OWL2 activity patterns and a SPARQL-based approach to overcome OWL2 limitation in supporting temporal reasoning and dynamic assertion of structured individuals. Here the reasoning framework is enhanced with a conceptual layer that allows the formal representation of activity meta-knowledge by means of DOLCE+DnS Ultralite (DUL) ontology patterns. Anyway, the authors do not report comparisons with existing methods or quantitative results at this stage. Other works have investigated the possibility of recognizing events in video taking advantages of the accuracy of probabilistic approaches as well as the descriptive capabili-

ties of semantic-based approaches as presented in [63]. The work shows how the formalization of knowledge relevant to video analysis within a specific domain can be used to define strategies for the event recognition. A two-layer strategy is proposed to recognize events handling the uncertainty of the low-level analysis: the short-term layer to recognize timeless events (changes in object features); the long-term layer to detect events with a temporal relation among their counterparts. The authors evaluate the approach on controlled environments (knowledge about object types that can appear) and uncontrolled environments (public places) selecting sequences from different datasets (VISOR, HERMES, AVSS...) obtaining up to 88,5% accuracy.

In recent times, there is an increasing interest towards the recognition of abnormal events with the use of ontologies. An approach to detect high level events using SWRL has been presented by Pantoja *et al.* [57]: it combines middle-level events and information (about actors and actions) extracted from a Visual Analysis module with a semantic rules inference system to detect meaningful high level crime scenarios. Rules are created manually using an empirical criteria, where experts look at instances of the events in videos and create a natural language description of the rules (the authors provide a detailed example of pick-pocketing where the thief is not alone, and carries out a real “relay" with an accomplice, passing to him the crime body, in order to not being caught with the stolen bag or wallet). Such description is then encoded in SWRL language and embedded in the knowledge base’s TBox. The experimental stage is carried out using two components: a first component performs visual analysis on CCTV scenes; then results are stored in a semantic framework where reasoning is performed with description logic rules. First experiments are made by using Madrid Police videos and show that the statistical performance of the reasoner in detecting crimes in real world situations is encouraging, 60% accuracy in the average, up to 83% when recognizing vandalism against the walls. As the authors observe, the SWRL standard does not allow the creation, by inference, of new individuals, but only the addition of data and object properties to the existing ones. In this way it is not possible, directly, to create a new event from deductions made on two already defined events, but only relationships between them can be expressed. The work proposed by Greco *et al.* [33] overcomes this problem by using SPIN technology. Here contextual knowledge is modeled through a general tracking ontology used to annotate a tracker output and enable reasoning. In particular, the ontology models information coming from the tracking component (frames, bounding box...), knowledge about the scene (static and dynamic objects, occluding objects...), Situations and Events (people leaving scene, falling ground, fighting...). SPIN rules and functions are used to determine when a particular event occurs while SPARQL queries are employed for analytics tasks. The system has proven to successfully recognize Mid level events (ex. people falling to ground) and High level events (ex. person being attacked) on PETS2016 dataset.

Interesting applications have been also recently proposed in the field of aerial surveillance by unmanned aerial vehicles (UAV). In [16] an approach exploiting the synergy between the tracking methods and semantic technologies for automatic object labeling has been proposed. The aim of the work is to enhance and understand situation awareness, as well as critical alerting situations: the UAV with an embedded camera is used to recognize moving and permanently fixed objects within the scene as well as relations and interactions between them. Here contextual informations are used to detect alerting and dangerous situations. The authors propose a first test of the system by using videos captured from a drone flying on the university campus of Salerno. In particular, a case is reported where the system is able to recognize a dangerous situation by means of SWRL rules associated to mid level activities "man kicking a ball on the road" and "car passing through the same road".

4.3. Video retrieval applications

In the following, some interesting works exploiting the capabilities of semantic technologies for video search and retrieval problems are described.

One of the first ontology based method to address the information retrieval problem in generic multimedia content collections, where no key images or textual annotations are available, has been proposed by Kompatsiaris *et al.* [45]. Still-image and video segmentation tools are used to enable time-efficient and unsupervised analysis of visual information within spatial or spatio-temporal objects to allow “content-based" access and manipulation via the extraction of MPEG-7 compliant low-level indexing features for each object. The key point is the use of ontologies to associate low-level indexing features and descriptors (texture, dominant color, contour shape) with higher-level

concepts (or keywords) that humans are more familiar with. In particular, the ontologies are employed to allow the user to query an image and video collection using semantically meaningful concepts (semantic objects), without the need for performing manual annotation of visual information. The ontology paradigm is coupled with a relevance feedback mechanism, based on support vector machines, to achieve better precision in retrieving the desired content. The resulting retrieval scheme provides flexibility in defining the desired semantic object/keyword and bridges the gap between automatic semantic indexing for specific domains and query-by-example approaches. The authors test the proposed methodology by using standard image dataset (5000 images from Corel library) and 812 video shots created by digitizing parts of movies and collecting video clips available on the Internet. The experimentation results demonstrate the use of ontologies and their importance in associating low-level features to high-level concepts in a flexible manner. In particular two object ontology have been defined: one for describing a blue car and another for describing a brown horse. The evaluation demonstrates better performance (both in terms of precision and recall) with respect to traditional approach based on global histogram. In the same way, Town [73] presents a query and retrieval method called OQUEL (ontological query language) to facilitate formulation and evaluation of queries consisting of sentences expressed in a language designed for general purpose retrieval of photographic images. Sentences in such a language are linked to visual evidence iteratively, using the ontology as a structured probabilistic prior to tie together different recognition and processing methodologies. The most interesting aspect lies in the retrieval process that, entirely acting within the ontological domain defined by the syntax and semantics of the user query, utilizes automatically extracted image segmentation and classification information, as well as Bayesian networks to infer higher level and composite terms. The proposed approach becomes an effective mechanism for addressing two key problems of content based image retrieval: (i) the ambiguity of image content and user intention and (ii) the semantic gap which exists between user and system notions of relevance. By basing such a language on an extensible ontology, one can explicitly state ontological commitments about categories, objects, attributes and relations without having to pre-define any particular method of query evaluation or image interpretation. OQUEL queries (sentences) are prescriptive rather than descriptive, i.e. the focus is on making it easy to formulate desired image characteristics as concisely as possible. In order to allow users to enter both simple keyword phrases and arbitrarily complex compound queries, the language grammar features constructs such as predicates, relations, conjunctions and a specification syntax for image content. The latter includes adjectives for image region properties (i.e. shape, colour, and texture) and both relative and absolute object location. Desired image content can be denoted by nouns such as labels for automatically recognized visual categories of stuff (grass, cloth, sky, etc.) and through the use of derived higher level terms for composite objects and scene description (e.g. animals, vegetation, winter scene). The OQUEL language has been implemented has part of the ICON content-based image retrieval system. The author compares the method with other query composition and retrieval approaches available in ICON system (query-by-example and a combination of sketch and feature-based retrieval) demonstrating its superior efficiency and flexibility. In particular, 670 images (extracted from the Corel images library) and 412 amateur digital pictures of highly variable quality and content were chosen to build the experimentation dataset. Twelve OQUEL query has been defined and assessments (in terms of relevant vs non-relevant items retrieved) were carried out manually for all 1082 images. Example of queries are: "indoors & people in foreground"; "some water in the bottom half which is surrounded by trees and grass, size at least 10%"; "city or countryside". For only three query out of 12 the system presents a normalized Rank value equal or slight greater than the other methods.

A comparison between traditional keyword based image retrieval and an ontology based image retrieval by constructing ontologies not only from text annotations, but also employing a combination of text annotation and image feature is proposed by Wang *et al.* [76]. The authors choose a challenging experimental domain, *canine* (a sub-domain of *animal*), and derive the formal definition and domain knowledge for the animal ontology from the BBC Science & Nature Animal category. They also define a textual description ontology (purely based on text to encapsulate high-level description) and a visual description ontology that incorporates classes and relationships extracted from low-level features. The experimental stage compares the ontology based image retrieval system with the Google image search. The experiment dataset is set up by a total of 4000 images using the top 200 Google images of each of the 20 canine subspecies. For the

comparison, the authors use the Google image Search with text ontology-based retrieval and multi-modality ontology-based retrieval. For semantic matchmaking of ontologies, they choose RACER as reasoner since it is able to provide consistency checking of the knowledge base, computing entailed knowledge via resolution and processing queries through complex reasoning. The experimentation results show that the multi-modality ontology-based retrieval outperforms others by returning more relevant images with higher ranking. In the best case “arctic fox", the multi-modality ontology-based retrieval almost overlaps the optimum by returning the N correct images in the first N ranking positions. The experimentation shows how the performance of the system is strictly related to the accuracy of image feature classification used for extracting low level features from the images (indeed for White-Fur, the ACCR value is 0.826, the highest among all species). The authors have also evaluated the precision of their system demonstrating that their approach outperforms the others since retrieved image lay in the top 20, 40, 60 and 80 for all 20 canine subspecies. This demonstrates that by combining the high-level textual information with low-level image features it is possible to improve retrieval precision by about 5 to 30 percent.

Snoek *et al.* [70] propose a multimedia thesaurus consisting of a set of machine learned concept detectors enriched with semantic descriptions and semantic structure obtained from WordNet. The key contribution is an approach to identify, given a multimodal user query, one of three possible strategies to select a relevant detector: text matching (where the text specification of a query is matched with the textual description of a detector), ontology querying (where the concept detector that maximizes Resnik’s measure of information content is selected), and semantic visual querying (where the detector with the maximum posterior probability with respect to all available visual models is selected). In the paper, a general-purpose ontology (with over 100,000 concepts) has been linked to a specific detector set (with several hundreds of concepts). In particular, the authors establish a link between WordNet and a set of 363 detectors learned from both MediaMill and LSCOM annotations. The method has been evaluated against the automatic search task of the TRECVID2005 video retrieval benchmark, using a 85 hours long news video archive. The evaluation assesses three different aspects: the influence of thesaurus size on video search performance (experiment 1); evaluate and compare the multimodal selection strategies for concept detectors (experiment 2); and discuss their combined potential using oracle fusion (experiment 3). For the experiments 2 and 3 the search results are compared against the best possible concept detector score for each topic in relative percentages of average precision (AP%). The results show that semantically-enriched detectors enhance results in semantic retrieval tasks in the majority of the cases. For the experiment 2, in particular, the Semantic Visual Querying approach outperforms the text matching and the simple ontology querying achieving the best score in 12 out of the 24 search queries.

An interesting effort to join together the ontology based annotation and retrieval concepts and the requirements of the computer vision and video surveillance communities has been made by Vezzani *et al.* [75] that propose an open platform for collecting, annotating, retrieving, sharing surveillance videos called the VISOR (VIdeo Surveillance Online Repository) project. The ViSOR open repository is based on a reference ontology integrating many concepts from LSCOM and MediaMill ontologies. The system is conceived as a web application with two main sections: one working at server side and the other one implementing the client user interface. In particular, the web interface allows to browse videos, query by annotated concepts or by keywords, preview compressed video, download and upload media. The repository contains metadata annotations, which can be either manually created as ground truth or automatically generated by video surveillance systems. In the ViSOR framework the performance evaluation tool named ViPER-PE has been integrated, allowing to compare two different annotation files and to report performance results.

Yao *et al.* [82] present a framework (I2T - Image to Text) for parsing image and video content, extracting video event and providing semantic and text annotations. The key contribution is the AoG (And-or-Graph) visual knowledge representation that allows to learn categorical image representations and symbolic representations simultaneously from a large scale image. The AoG embodies vocabularies of visual elements including primitives, parts, objects, scenes as well as a stochastic image grammar that specifies syntactic relations (i.e. compositional) and semantic relations (e.g. categorical, spatial, temporal, and functional) between these visual elements. It connects low-level image features with high-level semantically meaningful concepts so that the parsed image can be transformed to a semantic metadata format and finally to a textual description. The I2T framework provides richer and semantically oriented annotation of visual contents

since image and video contents are expressed in both OWL and text format and allows the integration with a full text search engine, as well as SPARQL queries, to provide accurate content-based retrieval. Users can retrieve images and video clips via keyword searching and semantic-based querying. The approach has been validated in maritime and urban video surveillance contexts and through a real-time automatic driving scene understanding system. For both domains, the authors evaluate event detection and metadata/text generation with sequences of different scenes. The data set is composed by ten sequences of urban and maritime scenes, with a total duration of about 120 min, that contain more than 400 moving objects. Visual events have been extracted and text descriptions have been generated. The authors define 12 different kind of events like: entering and exiting the scene, moving at abnormal speed, approaching traffic intersection, watercraft approaching a maritime, an object following another object, etc.. For automatic driving scene understanding the authors do not use a specific dataset (it was an ongoing project) bus simply demonstrate how it is possible to identify relevant objects (cars, pedestrian, etc.) in the foreground. In both experimentation domains, authors do not provide performance data (precision and recall) of the proposed framework, but show with several images how the system is able to provide a textual description of the objects and the event depicted in the images.

Xue *et al.* [81] propose an ontology-based content archive and retrieval framework for surveillance videos. They define a surveillance domain ontology that acts as a content description schema to allow hierarchical analysis of video data and then build OWL description files that represent semantic information of video clips as a resource ontology. Such an ontology models the basic feature description in the low level, the video object description in the mid level and activity/event description in the high level. In particular, the low-level part includes metadata specialization and MPEG-7 descriptors, the mid level defines Object classes (fix, mobile and contextual object), the high level models common activities decomposing complex events representation into temporal, spatial or logic combination of objects. For reasoning capabilities, the Apache Jena framework is used together with OWL query API to search and retrieve results of the input query via web browser and locate feedback onto the video data. The authors test the system for object (walking people) and event (car parking) retrieval using the PETS 2001 dataset, reporting an average accuracy classification rate of 93.15% without providing information about the number of objects and event present in the scene.

Xu *et al.* in [80] and [79] propose a method to annotate video traffic events by defining concepts (people, vehicle, traffic signs...) and their spatial and temporal relations. The major contribution regards the introduction of Video Structural Description (VSD), a hierarchical semantic data model including three different layers: pattern recognition layer (that extracts and represents the contents of the video through the ontology), video resources layer (that links video resources with their semantic relations), user demands layer (the retrieval engine). The define concepts in the ontology such as persons, vehicles, and traffic signs, that can be used for annotating and representing video traffic events. In addition, the spatial and temporal relation between objects in an event is defined. As a case study, an application to annotate and search traffic events is considered. The application uses a video annotation ontology that reuses RDF vocabularies from different knowledge repository (the traffic law of China, the basic features of car and the basic features of person). The video annotation module allows to load video resources and annotate them according to the given ontology while the video search module allows to search concepts according to the ontology content. The system has been tested for identifying the illegal cars that use others' licenses (in China, each car must have a sole license and a sole license number). The dataset is composed by 1.19 billion data from the traffic speed camera. The dataset (stored in ten servers) has been processed for extracting information about car license, GIS position, time, car color and model. The rule for detecting cars with illegal license number has been defined as follow: "the distance between cars with the same license number should be lower than 15 km in the time interval of 10 min". The system running on a Map-Reduce cluster in 50 min has processed the whole dataset identifying 395 illegal cars.

Sobhani *et al.* [71] propose an advanced intelligent forensic retrieval system by taking advantage of an ontological knowledge representation. A relevant use case drawn from a real-world event (the UK riots 2011) is considered, with event sources obtained from CCTV footage and video footage captured using hand held devices. OWL is chosen to represent the forensic domain ontology which is divided into seven classes (entities, event, event category, place, resource, source and types of damage); in particular, entities include *person*, *group*, *organization* and *object* while the event

category class is split into two sub classes which are *crowd disaster* and *terrorist attack*. The key contribution of the work is then the analysis and the development of the ontology whose major aim is to share a common understanding of the domain structure among forensic investigators for addressing difficulties in handling, investigating and filtering a large amount of data. The ontology is also used to demonstrate how high level reasoning can be incorporated into an automated forensic system and has been designed according to a well defined process (Top-Down development process) using 7 competency questions. A dataset of 3.07TB data of CCTV surveillance footage provided by the Scotland Yard as a part of European LASIE project has been used. A portion of the dataset of about 3.46GB has been used for the domain ontology population. The reduced dataset has been manually annotated for the test case scenario related to a sports riot event for identify 9 different activities like: Loitering around the town center; Smashing vehicles (buses) in the middle of the road; Attack the police and getting involve in a fight; etc.. The populated ontology contains 12 individuals in the *event* class, 8 individuals in the *person*.

4.4. Multimedia visual content annotation

First uses of ontology in multimedia are dated back to 2001, right after the Semantic Web introduction, although they mainly involve the use of ontology as a thesauri-like approach to manual content annotation [65].

Later on, ontology has begun to play an important role also in driving the extraction of semantic description. Some works have investigated the possibility to map MPEG-7 visual metadata to ontologies. For example, Simou *et al.* [66] present the construction of an ontology that represents the structure of the MPEG-7 visual part to enable machines to generate and understand visual descriptions which can be used for multimedia reasoning. In the same direction, Bloehdorn *et al.* [15] propose a software environment (M-Ontomat Annotizer) that links low level MPEG-7 visual descriptions to conventional Semantic Web ontologies and annotations for Multimedia Analysis; in particular, ontologies are used to represent high-level multimedia concept descriptions in conjunction with low-level visual descriptors and links between them. Some qualitative interesting results are shown in the tennis domain. Hollink *et al.* [38] propose a visual ontology (built out of two existing knowledge corpora - WordNet and MPEG-7 - by creating links between visual and general concepts) to aid video annotation in a broad domain and identify requirements for a generic visual ontology. The idea is interesting, although the results, which test precision and reliability of annotations, are at a preliminary level. A first important step towards the creation of a framework for research on semantic analysis of multimedia content has been achieved with the development of the first Large-Scale Concept Ontology for Multimedia (LSCOM) based on a taxonomy of 1000 concepts along with a set of use cases and queries and a large annotated data set of broadcast news video. This has been described by Naphade *et al.* in [55] as a collaborative effort led by IBM, Carnegie Mellon University, and Columbia University with CyC corporation and various other research academic and industrial groups. Fan *et al.* [21] propose to incorporate a concept ontology to boost hierarchical video classifier training and multimodal feature selection. They show a case study in a specific surgery education domain where ontology models common needs and interests of medical students to training certain clinic skills, which are well-defined by specific medical education program. The integration of a domain-specific and user-centric concept ontology for video concept organization and indexing is aimed at enhance medical students' ability on video access. The paper reports interesting results about the hierarchical boosting algorithm for classification tasks, but only qualitative considerations are made for the ontology application. Hudelot *et al.* in [42] propose a methodology where the ontology uses structural information on the spatial arrangement of the structures and is enriched by fuzzy representations of concepts which define their semantics, and allow establishing the link between these concepts and the information that can be extracted from images. This methodological approach is illustrated on a medical example, dealing with knowledge-based recognition of brain structures in 3D magnetic resonance images using the proposed fuzzy spatial relation ontology. No comparison with other methods is reported. The advantages of using multimedia ontologies to perform video annotation have been also investigated by Bertini *et al.* [9] [11] [10]. They show how the knowledge modeled by means of ontology allows to perform more refined annotation with respect to common visual data analysis methods and exploit a case study on soccer videos. They also introduce the Dynamic Pictorially Enriched Ontology model that includes linguistic concepts together with visual prototypes obtained by clustering the instances

of observed visual data. The Ontology Web Language (OWL) is used to model both domain concepts and visual prototypes, and the Semantic Web Rule Language (SWRL) to enhance, through reasoning, the results of the classification and derive new semantic annotations. A method to learn a set of rules for semantic video annotation and event recognition by exploiting the domain knowledge embedded into an ontology is also proposed where they adapt the First Order Inductive Learner (FOIL) technique to the Semantic Web Rule Language (SWRL). The annotation framework was tested on the Formula 1 and soccer domains to show its general applicability and achievable performance improvements and the use of visual prototypes with SWRL reasoning show an average improvement in precision of 10 percentage points with respect to SVM. *et al.* [6] present an approach for automatic annotation and retrieval of video content based on ontologies and semantic-concept classifiers. Taxonomical relations between concepts are determined, using WordNet, to define the ontology schema; the concept detectors are then linked to the corresponding concepts in the ontology and a rule-based method (with rules expressed using SWRL) allows for automatic semantic annotation of composite concepts and events in videos. The ontology used for the experiments is built from the MediaMill detectors thesaurus while the dataset used is the training set of Trecvid 2005. Results show an overall improvement of 17,64 percent with respect to the baseline detectors.

5. Considerations and challenges

The analysis of the selected papers has been developed according to different perspectives. As mentioned in section 1, we have chosen 75 works: 31 conference papers, 40 journal papers, and 3 book chapters. Fig.1 shows the distribution of the works over the years, according to each category. We can observe that a relevant number of papers in the field of interest has been published starting from 2005; in particular a peak in the number of journal and conference papers per year (about 8) is obtained in 2005, 2006 and 2010 respectively. Interestingly, we can individuate a correlation between such peaks and the diffusion and the standardization of important semantic technologies (reported in the timeline in Fig.2) such as SPARQL (introduced in 2004 and then standardized in 2008) as well as the proposal of OWL2 in 2009. Another intriguing distribution is reported in Fig. 3 where we show the number of papers with respect to the defined taxonomy. In particular, we report: papers about multimedia visual content annotation, mostly representing the early works in the field; papers on low level analysis, where the ontology is used to model low level features and drive detection, segmentation or tracking processes; papers about video retrieval applications; papers about mid and high level analysis, which involve visual event and activity recognition, representing the major contribution of the recent days. Other works discuss the use of semantic web technologies in conjunction with general knowledge based systems or provide a survey of their use in computer vision problems. In Table 1, for each paper, the main contributions have been summarized by taking into account the general topic, the domains described by means of ontology, rules and/or query mechanism involved, dataset (standard or custom made) used for the experimentation.

From our analysis, it becomes evident that while the early works in this field adopted ontology as a mean for annotating data for retrieval purposes, now semantic technologies are crucial for event detection and activity recognition since the most recent papers make large use of ontologies and last reasoning methodologies to infer complex activities from atomic events. The role of ontologies has also become prominent in pervasive computing since semantic annotation is particularly affordable for fusing data acquired from different sensors (not only video) in the big data and the Internet of things domains. In particular, according to the defined taxonomy, works on low level analysis have been proposed since 2005 where ontology, in most cases, models domain related concepts, represents low level features and allows to select appropriate processing methods for analyzing video contents through reasoning. Furthermore, there are recent efforts that leverage semantic technologies for low level error detection and correction. Papers on content annotation and retrieval are also spread in the whole time frame and make use of ontology to describe concept detectors, descriptors (as in the case of MPEG-7 that was first publicly released in 2002 and received the latest amendment in 2011), linguistic concepts, traffic events and for image annotation. Research works in the field of mid level analysis have become popular in the last ten years and typically use ontology to model activities as composed of simple events or to represent scenarios, situations, roles, states and general visual description. More recent works have instead addressed the problem of high level analysis, where the role of ontologies and rea-

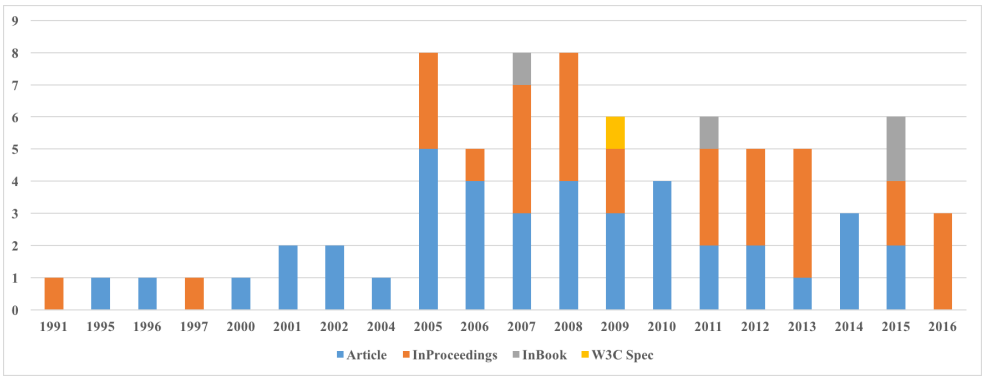

Fig. 1. Surveyed papers' distribution over the years

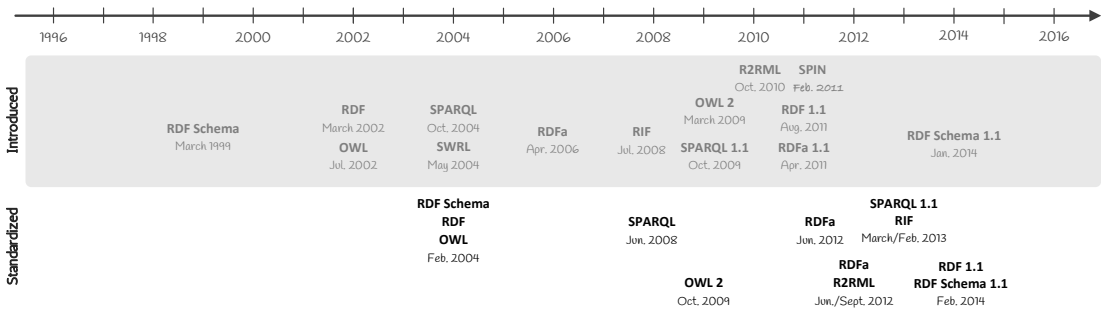

Fig. 2. W3C Semantic Web Standards Timeline [13]

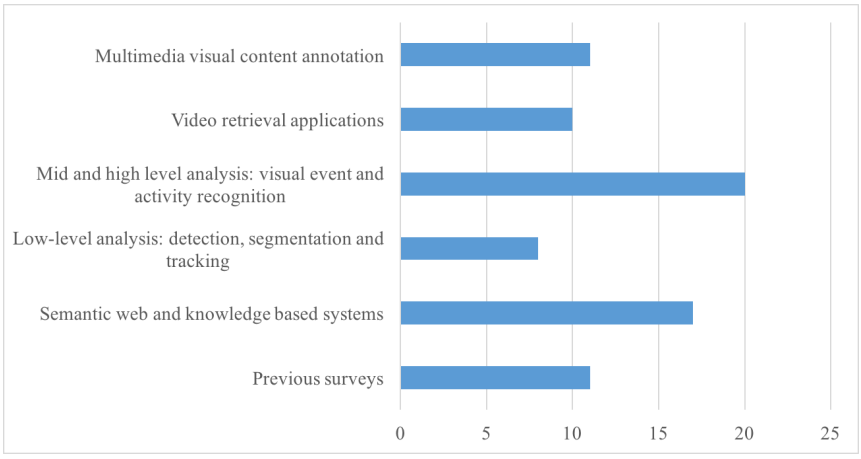

Fig. 3. Surveyed papers' topic distribution

soning is mostly centered on collecting data from heterogeneous sensors (not only video streams) and detect complex activities; in this context, successful applications have been proposed for monitoring Activities of Daily Living (ADL) or in the context of Ambient Assisted Living (AAL). Such approaches represent the trend of semantic technologies making use of OWL ontological knowledge not only to represent domain relationships between low-level observations and high-level activities, but also to support context aggregation and activity interpretation.

What is evident from the overall analysis is that only few of the published works make custom designed ontologies available for public use, except the case of VERL in 2015. Examples supporting knowledge reuse are actually very rare and restricted to some specific schemas like DOLCE, MMIO (Multimedia Information Object), LSCOM (Large Scale Concept Ontology for Multimedia), COMM (Core Ontology for Multimedia Annotation), Ontology for Media Resources 1.0 and M3O (Multimedia Metadata Ontology). The presence of so many multimedia ontologies also confirms a different maturity of such a domain; in fact, multimedia has been an early adopter field, as we can observe from timeline analysis.

Another interesting evidence is about the usage of OWL. What emerges from several papers is a very basic adoption of OWL also in cases where simpler schema like RDFS would be just enough. Only few applications have fully exploited the OWL characteristics by also considering the features of the different profiles. Moreover, a very low adoption of OWL2 and the related profiles is encountered too.

5.1. Challenges

The intelligent video analysis poses new challenges to the SW research community. In the area of Big Data processing, last trends of SW technologies emphasize the role of stream processing and stream reasoning. In particular, there is a need to add temporal dimension to semantic data representation and processing. Many approaches using extension of SPARQL, like C-SPARQL, and technologies like Apache Storm to distribute the processing workload for addressing decision support needs are emerging for instance in smart cities [49], or in problems of topic detection and tracking on microblogging [50]. We think that a future trend in this direction could also involve video analysis as demonstrated in [32].

Our analysis also emphasized the need of developing and sharing large knowledge graphs to support a new set of applications that adopt solutions oriented to the fusion of traditional bottom-up approaches (learning from data) with top-down (starting from the model and background knowledge). In this context, SW technologies have to facilitate the management and information extraction from such graphs through incremental and distributed inference engines by fully exploiting parallel architectures that would allow to execute inferences in a reasonable amount of time. Another emerging issue coming from video analysis is related to uncertainty in knowledge representation. It could be very difficult for a detection algorithm to discriminate whether an adult is walking in a scene with a trolley, a dog or a baby. The possibility to properly annotate actions in the knowledge base contributes to the creation of more reliable applications especially in the domain of intelligent business analytics solutions, that is nowadays one of the hottest research areas in video analysis as demonstrated in [18]. Here, fuzzy and probabilistic ontologies with related reasoners could help in better representing the context and understand the possible situations.

Finally, we conclude our discussion with some considerations about the "Open World" assumption issue, which “affects" SW language chain and is often seen as an obstacle in the development of affordable business applications. Steps ahead have been with the introduction SPIN technology (actually supported by several triple store providers even if with some limitation) and the efforts of SW community in the development of the Shapes Constraint Language (SHACL) under the W3C RDF Data Shapes Working Group seem to lead to promising results.

Table 1
Summary of the surveyed papers

Taxonomy	Papers	Ontology modeling	Rules/Reasoning	Datasets* (standard & custom)
Low level	[19]	Domain concepts with qualitative attributes, Low level features, Object spatial relations and Processing methods	F-logic	Formula One, Soccer
	[27]	Low level features	_	VSSN06
	[29]	Camera Data, Tracking Data, Scene Objects, Activities, Impact and threats, Feedback	RACER	PETS2002
	[58]	Domain concepts (DOLCE Extension)	_	Domain Disaster news
	[31]	Domain concepts, Low Level Features	SPIN	PETS2009
Mid level	[25]	Composable events with temporal relationships	_	Bank Surveillance
	[3]	Composable events , Contextual information		Bank Surveillance, Airport Tarmac Surveillance
	[69]	Composable events , Contextual information	SWRLJess	_
	[62]	Object, Event, Context, Capabilities, Reactions	_	i-LIDS , PETS2006
	[73]	Visual content descriptors (scenarios, situations, roles, states, and visual properties)	_	CAVIAR project
	[5]	High level semantic concepts and their relations within the considered domain	DL	Soccer
	[85]	Events and Vision Primitives	_	CAVIAR project
	[64]	Domain Concepts, Capabilities of the analysis and User preferences	_	AVSS2007, PETS2006, CANTATA and HERMES
High level	[44]	Low level features, Events	SWRL	PETS2012
	[17]	Activity of Daily Living concepts, Algorithms for activity recognition	_	_
	[60]	Human activities, Sensor data , Information about people and objects	_	
	[52]	Human activities, Sensor data , Information about people and objects	SPIN	_
	[63]	Low level features, events and temporal relations	_	
	[57]	Middle level events, actors and actions	SWRL	
	[33]	Domain concepts, mid/high level events, low-level features	SPIN	PETS2016
	[16]	Domain concepts, low-level features	_	_
Content retrieval	[45]	Low-level indexing features and descriptor with higher-level concepts	_	Corel library
	[6]	Domain concepts	_	
	[73]	Structured probabilistic prior to tie together different recognition and processing methodologies	_	_
	[76]	High-level descriptions, classes and relationships extracted from low-level features	_	_
	[70]	Concept detectors	_	TRECVID2005
	[75]	Domain concepts	_	_
	[82]	Low level features, high level concepts	_	_
	[81]	Low level features, mid level and activity level description.	_	PETS2001
	[80]	Domain concepts (traffic events)	_	_
	[71]	Forensic domain concepts	_	UK Riots 2011 CCTV
Content annotation	[65]	Photo annotation	_	_
	[66]	MPEG-7 descriptors and structure	_	_
	[15]	MPEG-7 descriptors and structure	_	_
	[38]	Multimedia Domain Concepts	_	Broadcast News Video
	[21]	Domain concepts	_	Surgery education
	[42]	Structural information on spatial arrangement, fuzzy representations of concepts	_	3D magnetic resonance
	[10]	Visual descriptors, Linguistic concepts	_	_

6. Conclusions

In this work, we have proposed a survey of many relevant papers concerning the application of Semantic Web technologies to video analysis. Our work was aimed at characterizing the potentiality offered by semantic web technologies for improving the performance of existing algorithms and solutions and enabling advanced video analytic functionalities. A taxonomy of the SW technology adoption for video analysis has been proposed and the surveyed papers have been analyzed according to this taxonomy. Moreover, an analysis of the considered papers with respect to the time frame and use of the SW languages has been carried out too. As a result, our study revealed an increasing trend in the use of Semantic Web technologies for event detection and activity recognition thanks to the advances in reasoning techniques that allows to infer effectively complex activities from atomic events. The use of ontologies is also spreading very quickly in pervasive computing since semantic annotation is particularly affordable for fusing data acquired from different sensors in the big data and the Internet of things domains. That said, we expect more and more video analysis applications to come in the near future thanks to the maturity of stream processing and reasoning and the advances in parallel processing technologies in the Big Data analysis domain, that allow to distribute the workload in an efficient way.

References

[1] A motion-based image processing system for detecting potentially dangerous situations in underground railway stations. *Transportation Research Part C: Emerging Technologies*, 14(2):96 – 113, 2006.

[2] Trevor Ainsworth. Buyer beware. *Security Oz*, 19:18–26, 2002. cited By 10.

[3] Umut Akdemir, Pavan Turaga, and Rama Chellappa. An ontology based approach for activity recognition from video. In *Proceedings of the 16th ACM International Conference on Multimedia*, MM '08, pages 709–712, New York, NY, USA, 2008. ACM. cited By 53.

[4] A. Aydm Alatan, E. Tuncel, and L. Onural. A rule-based method for object segmentation in video sequences. In *Image Processing, 1997. Proceedings., International Conference on*, volume 2, pages 522–525 vol.2, Oct 1997. cited By 17.

[5] L. Bai, S. Lao, G. J. F. Jones, and A. F. Smeaton. Video semantic content analysis based on ontology. In *Machine Vision and Image Processing Conference, 2007. IMVIP 2007. International*, pages 117–124, Sept 2007. cited By 47.

[6] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, and Giuseppe Serra. Video annotation and retrieval using ontologies and rule learning. *IEEE MultiMedia*, 17(4):80–88, 2010. cited By 54.

[7] H. Bannour and C. Hudelot. Towards ontologies for image interpretation and annotation. In *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pages 211–216, June 2011. cited By 31.

[8] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001. cited By 20023.

[9] M. Bertini, A. D. Bimbo, C. Torniai, C. Grana, R. Vezzani, and R. Cucchiara. Sports video annotation using enhanced hsv histograms in multimedia ontologies. In *Image Analysis and Processing Workshops, 2007. ICIAPW 2007. 14th International Conference on*, pages 160–170, Sept 2007. cited By 12.

[10] Marco Bertini, Alberto Del Bimbo, Giuseppe Serra, Carlo Torniai, Rita Cucchiara, Costantino Grana, and Roberto Vezzani. Dynamic pictorially enriched ontologies for digital video libraries. *IEEE MultiMedia*, 16(2):42–51, April 2009. cited By 13.

[11] Marco Bertini, Alberto Del Bimbo, and Giuseppe Serra. Learning ontology rules for semantic video annotation. In *Proceedings of the 2nd ACM workshop on Multimedia semantics*, pages 1–8. ACM, 2008. cited By 22.

[12] Claudio Bettini, Oliver Brdiczka, Karen Henricksen, Jadwiga Indulska, Daniela Nicklas, Anand Ranganathan, and Daniele Riboni. A survey of context modelling and reasoning techniques. *Pervasive Mob. Comput.*, 6(2):161–180, April 2010. cited By 852.

[13] Nikos Bikakis, Chrisa Tsinaraki, Nektarios Gioldasis, Ioannis Stavrakantonakis, and Stavros Christodoulakis. *The XML and Semantic Web Worlds: Technologies, Interoperability and Integration: A Survey of the State of the Art*, pages 319–360. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[14] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):167–177, June 2005. cited By 138.

[15] Stephan Bloehdorn, Kosmas Petridis, Carsten Saathoff, Nikos Simou, Yannis Avrithis, Siegfried H, Yiannis Kompatsiaris, and Michael G. Strintzis. Semantic annotation of images and videos for multimedia analysis. In *In Proceedings of the 2nd European Semantic Web Conference, ESWC 2005*, volume 3532, pages 592–607, 2005. cited By 229.

[16] D. Cavaliere, S. Senatore, M. Vento, and V. Loia. Towards semantic context-aware drones for aerial scenes understanding. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 115–121, Aug 2016.

[17] Liming Chen and Chris Nugent. Ontology-based activity recognition in intelligent pervasive environments. *International Journal of Web Information Systems*, 5(4):410–430, 2009. cited By 137.

[18] S. Chen, K. Clawson, M. Jing, J. Liu, H. Wang, and B. Scotney. Uncertainty reasoning based formal framework for big video data understanding. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 2, pages 487–494, Aug 2014.

[19] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, and M. G. Strintzis. Knowledge-assisted semantic video object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1210–1224, Oct 2005. cited By 113.

[20] Stamatia Dasiopoulou, Eirini Giannakidou, Georgios Litos, Polyxeni Malasioti, and Yiannis Kompatsiaris. *A Survey of Semantic Image and Video Annotation Tools*, pages 196–239. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. cited By 56.

[21] 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):939–957, Aug 2007. cited By 31.

[22] James Ferryman and Anna-Louise Ellis. Performance evaluation of crowd image analysis using the {PETS2009} dataset. *Pattern Recognition Letters*, 44(0):3 – 15, 2014. Pattern Recognition and Crowd Analysis.

[23] Sandro Rama Fiorini and Mara Abel. A review on knowledge-based computer vision. 2010.

[24] P. Foggia, G. Percannella, A. Saggese, and M. Vento. Real-time tracking of single people and groups simultaneously by contextual graph-based reasoning dealing complex occlusions. In *Performance Evaluation of Tracking and Surveillance (PETS), 2013 IEEE International Workshop on*, pages 29–36, 2013. cited By 7.

[25] A.R.J. Francois, R. Nevatia, J. Hobbs, and R.C. Bolles. Verl: An ontology framework for representing and annotating video events. *IEEE Multimedia*, 12(4):76–86, 2005. cited By 196.

[26] Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. Sweetening ontologies with dolce. *Knowledge engineering and knowledge management: Ontologies and the semantic Web*, pages 223–233, 2002.

[27] Alvaro García and Jesús Bescós. Video object segmentation based on feedback schemes guided by a low-level scene ontology. In *Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems*, ACIVS '08, pages 322–333, Berlin, Heidelberg, 2008. Springer-Verlag. cited By 8.

[28] Benoit Gaüzère, Pierluigi Ritrovato, Alessia Saggese, and Mario Vento. *Human Tracking Using a Top-Down and Knowledge Based Approach*, pages 257–267. Springer International Publishing, Cham, 2015.

[29] Juan Gomez-Romero, Miguel A. Patricio, Jesús García, and José M. Molina. Ontology-based context representation and reasoning for object tracking and scene interpretation in video. *Expert Systems with Applications*, 38(6):7494 – 7510, 2011. cited By 47.

[30] Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter Patel-Schneider, and Ulrike Sattler. {OWL} 2: The next step for {OWL}. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):309 – 322, 2008. Semantic Web Challenge 2006/2007.

[31] L. Greco, P. Ritrovato, A. Saggese, and M. Vento. Improving reliability of people tracking by adding semantic reasoning. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 194–199, Aug 2016.

[32] L. Greco, P. Ritrovato, and M. Vento. Advanced video analytics: An ontology-based approach. volume Part F129475, 2017. cited By 0.

[33] Luca Greco, Pierluigi Ritrovato, Alessia Saggese, and Mario Vento. Abnormal event recognition: A hybrid approach using semantic web. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016. cited By 0.

[34] Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5):907–928, 1995. cited By 9190.

[35] Amarnath Gupta, Terry E. Weymouth, and Ramesh Jain. Semantic queries with pictures: The vimsys model. In *Proceedings of the 17th International Conference on Very Large Data Bases*, VLDB '91, pages 69–79, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc. cited By 210.

[36] Allan Hanbury. A survey of methods for image annotation. *Journal of Visual Languages & Computing*, 19(5):617–627, 2008. cited By 120.

[37] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F Patel-Schneider, and Sebastian Rudolph. Owl 2 web ontology language: Primer. w3c recommendation (2009).

[38] L. Hollink, M. Worring, and A. Th. Schreiber. Building a visual ontology for video retrieval. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 479–482, New York, NY, USA, 2005. ACM. cited By 81.

[39] M. Horridge and S. Bechhofer. The owl api: A java api for owl ontologies. *Semantic Web*, 2(1):11–21, 2011. cited By 373.

[40] Ian Horrocks, Peter F Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, Mike Dean, et al. Swrl: A semantic web rule language combining owl and ruleml. *W3C Member submission*, 21:79, 2004. cited By 2513.

[41] Chih-Cheng Hsu, Wesley W. Chu, and Ricky K. Taira. A knowledge-based approach for retrieving images by content. *IEEE Trans. on Knowl. and Data Eng.*, 8(4):522–532, August 1996. cited By 121.

[42] CÃl'line Hudelot, Jamal Atif, and Isabelle Bloch. Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems*, 159(15):1929 – 1951, 2008. From Knowledge Representation to Information Processing and ManagementSelected papers from the French Fuzzy Days (LFA 2006).

[43] P. Kannan, P. Shanthi Bala, and G. Aghila. A comparative study of multimedia retrieval using ontology for semantic web. In *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on*, pages 400–405, March 2012. cited By 6.

[44] MohammedYassine Kazi Tani, Adel Lablack, Abdelghani Ghomari, and IoanMarius Bilasco. Events detection using a video-surveillance ontology and a rule-based approach. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops*, volume 8926 of *Lecture Notes in Computer Science*, pages 299–308. Springer International Publishing, 2015. cited By 1.

[45] Ioannis Kompatsiaris, Vasileios Mezaris, and Michael G Strintzis. Multimedia content indexing and retrieval using an object ontology. *Multimedia Content and Semantic Web-Methods, Standards and Tools*, pages 339–371, 2005. cited By 34.

[46] Benjamin Laxton, Jongwoo Lim, and David Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. cited By 109.

[47] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. pages 2036–2043, 2009. cited By 402.

[48] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level seman-

tics. *Pattern Recogn.*, 40(1):262–282, January 2007.

[49] Carmen De Maio, Giuseppe Fenza, Vincenzo Loia, and Francesco Orciuoli. Distributed online temporal fuzzy concept analysis for stream processing in smart cities. *Journal of Parallel and Distributed Computing*, 110(Supplement C):31 – 41, 2017. High Performance and Parallelism for Large Data Sets.

[50] Carmen De Maio, Giuseppe Fenza, Vincenzo Loia, and Francesco Orciuoli. Unfolding social content evolution along time and semantics. *Future Generation Computer Systems*, 66(Supplement C):146 – 159, 2017.

[51] B. McBride. Jena: a semantic web toolkit. *IEEE Internet Computing*, 6(6):55–59, Nov 2002. cited By 596.

[52] G. Meditskos, S. Dasiopoulou, V. Efstathiou, and I. Kompatsiaris. Sp-act: A hybrid framework for complex activity recognition combining owl and sparql rules. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pages 25–30, March 2013. cited By 21.

[53] B. Motik, R. Shearer, and I. Horrocks. Hypertableau reasoning for description logics. *Journal of Artificial Intelligence Research*, 36:165–228, 2009. cited By 357.

[54] Boris Motik, Ulrike Sattler, and Rudi Studer. Query answering for owl-dl with rules. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(1):41 – 60, 2005. Rules Systems.

[55] M. Naphade, J. R. Smith, J. Tesic, Shih-Fu Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, July 2006. cited By 603.

[56] Leonardo Onofri, Paolo Soda, Mykola Pechenizkiy, and Giulio Iannello. A survey on using domain and contextual knowledge for human activity recognition in video streams. *Expert Systems with Applications*, 63(Supplement C):97 – 111, 2016.

[57] C. Pantoja, A. Ciapetti, C. Massari, and M. Tarantelli. Action recognition in surveillance videos using semantic web rules. In *Imaging for Crime Prevention and Detection (ICDP-15), 6th International Conference on*, pages 1–6, July 2015. cited By 0.

[58] Georgios Th. Papadopoulos, Vasileios Mezaris, Ioannis Kompatsiaris, and Michael G. Strintzis. *Semantic Multimedia: Second International Conference on Semantic and Digital Media Technologies, SAMT 2007, Genoa, Italy, December 5-7, 2007. Proceedings*, chapter Ontology-Driven Semantic Video Analysis Using Visual Information Objects, pages 56–69. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. cited By 81.

[59] Chris Poppe, Gaëtan Martens, Pieterjan De Potter, and Rik Van De Walle. Semantic web technologies for video surveillance metadata. *Multimedia Tools Appl.*, 56(3):439–467, February 2012. cited By 7.

[60] D. Riboni, L. Pareschi, L. Radaelli, and C. Bettini. Is ontology-based activity recognition really effective? In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*, pages 427–431, March 2011. cited By 54.

[61] Natalia Díaz Rodríguez, M. P. Cuéllar, Johan Lilius, and Miguel Delgado Calvo-Flores. A survey on ontologies for human behavior recognition. *ACM Comput. Surv.*, 46(4), March 2014. cited By 33.

[62] J.C. SanMiguel, J.M. Martinez, and A. Garcia. An ontology for event detection and its application in surveillance video. In *Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, pages 220–225, Sept 2009. cited By 35.

[63] Juan C. Sanmiguel and José M. Martínez. A semantic-based probabilistic approach for real-time video event recognition. *Computer Vision and Image Understanding*, 116(9):937 – 952, 2012.

[64] JuanC. SanMiguel and José M. Martínez. A semantic-guided and self-configurable framework for video analysis. *Machine Vision and Applications*, 24(3):493–512, 2013. cited By 1.

[65] A Th Guus Schreiber, Barbara Dubbeldam, Jan Wielemaker, and Bob Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, (3):66–74, 2001. cited By 376.

[66] N. Simou, V. Tzouvaras, Y. Avrithis, G. Stamou, and S. Kollias. A visual descriptor ontology for multimedia reasoning. In *Proc. of WIAMIS '05*, 2005. cited By 46.

[67] Tomo Sjekavica, Ines Obradović, and Gordan Gledec. Ontologies for multimedia annotation: An overview. In *4th European Conference of Computer Science (ECCS'13)*, 2013. cited By 2.

[68] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, December 2000. cited By 6461.

[69] L. Snidaro, M. Belluz, and G. L. Foresti. Representing and recognizing complex events in surveillance applications. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 493–498, Sept 2007. cited By 19.

[70] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 9(5):975–986, Aug 2007. cited By 208.

[71] F. Sobhani, N. F. Kahar, and Q. Zhang. An ontology framework for automated visual surveillance system. In *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–7, June 2015.

[72] Huang Tiejun. Surveillance video: The biggest big data. *Computing Now*, 7(2):online, 2014. cited By 24.

[73] Christopher Town. Ontological inference for image and video analysis. *Machine Vision and Applications*, 17(2):94–115, 2006. cited By 73.

[74] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, Nov 2008. cited By 1041.

[75] Roberto Vezzani and Rita Cucchiara. Annotation collection and online performance evaluation for video surveillance: the visor project. In *Advanced Video and Signal Based Surveillance, 2008. AVSS'08. IEEE Fifth International Conference on*, pages 227–234. IEEE, 2008. cited By 13.

[76] H. Wang, S. Liu, and L.-T. Chia. Does ontology help in image retrieval?: a comparison between keyword, text ontology and multi-modality ontology approaches. pages 109–112, 2006. cited By 81.

[77] Jiang Wang, Zhuoyuan Chen, and Ying Wu. Action recognition with multiscale spatio-temporal contexts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3185–3192. IEEE, 2011. cited By 79.

[78] Leonard P. Wesley. Evidential knowledge-based computer vision. *Optical Engineering*, 25(3):363–379, 1986. cited By 24.

[79] Zheng Xu, Yunhuai Liu, Lin Mei, Chuanping Hu, and Lan Chen. Semantic based representing and organizing surveillance big data using video structural description technology. *Journal of Systems and Software*, 102(Supplement C):217 – 225, 2015.

[80] Zheng Xu, Lin Mei, Yunhuai Liu, and Chuanping Hu. Video structural description: a semantic based model for representing and organizing video surveillance big data. In *Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on*, pages 802–809. IEEE, 2013. cited By 4.

[81] Ming Xue, Shibao Zheng, and Chongyang Zhang. Ontology-based surveillance video archive and retrieval system. In *Advanced Computational Intelligence (ICACI), 2012 IEEE Fifth International Conference on*, pages 84–89, Oct 2012. cited By 4.

[82] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S. C. Zhu. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, Aug 2010. cited By 162.

[83] Juan Ye, Stamatia Dasiopoulou, Graeme Stevenson, Georgios Meditskos, Efstratios Kontopoulos, Ioannis Kompatsiaris, and Simon Dobson. Semantic web technologies in pervasive computing: A survey and research roadmap. *Pervasive and Mobile Computing*, 23:1 – 25, 2015. cited By 13.

[84] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4), 2006. cited By 3990.

[85] Sofia Zaidenberg, Bernard Boulay, and François Brémond. A generic framework for video understanding applied to group behavior recognition. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 136–142. IEEE, 2012. cited By 19.