Learning Ontology Rules for Semantic Video Annotation

Marco Bertini, Alberto Del Bimbo, and Giuseppe Serra
Media Integration and Communication Center, University of Florence, Italy
{bertini, delbimbo, serra}@dsi.unifi.it

ABSTRACT

Semantic video annotation using ontologies has received a large attention from the scientific community in the recent years. Ontologies are being regarded as an appropriate tool to bridge the semantic gap. In this paper we present an overview of the state-of-the-art of approaches and algorithms that exploit ontologies to perform semantic video annotation and present an approach to automatically learn rules describing high-level concepts. This approach exploits the domain knowledge embedded into an ontology to learn a set of rules for semantic video annotation. The proposed technique is an adaptation of the First Order Inductive Learner (FOIL) technique to the Semantic Web Rule Language (SWRL) standard: Experiments have been performed in two different video domains: *i)* the TRECVID 2005 broadcast news collection, to detect events related to airplanes, such as taxiing, flying, landing and taking off; *ii)* surveillance videos, to detect if a person enters or exits a specific area. The promising experimental performance demonstrates the effectiveness and flexibility of the proposed framework.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video Analysis*

General Terms

Algorithms, Experimentation

Keywords

Video retrieval, Events detection, Ontology, Learning rules

1. INTRODUCTION

Automatic annotation of video content at the semantic level has received a significant attention from the research community in the recent years, as a fundamental mean to face the explosive growth of video production and the associated growing request for search and retrieval by content of interesting elements. Important fields of application have been news, sports, surveillance, to cite some of those where there is the greatest impact on industry. Recently ontologies have been regarded as an appropriate tool to bridge the semantic gap between the information that can be extracted from the visual data and the interpretation of the same visual data by a user in a given context. An ontology consists of concepts, concept proprieties, and their relationships to provide a formal description of a domain and provides a common vocabulary that overcome semantic heterogeneity of information. Ontology Web Language (OWL) and Semantic Web Rule Language (SWRL) have been proposed by the World Wide Web Consortium (W3C) as language standards for representing ontologies and rules respectively. SPARQL Protocol and RDF Query Language (SPARQL) has been approved as W3C recommendation as query language for the Semantic Web technologies.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MS'08, October 31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-316-7/08/10 ...$5.00.

2. HIGH-LEVEL CONCEPT VIDEO ANNOTATION USING ONTOLOGIES

In the last years many researches have exploited ontologies to perform semantic annotation and retrieval from video digital libraries. Ontologies useful for semantic annotation of videos are those defined by the Dublin Core Metadata Initiative [1], TV Anytime [2] - they have defined standardized metadata vocabularies - and the LSCOM initiative [23] - that has created a specialized vocabulary for news video. In these cases, ontologies include a set of linguistic terms with their associated definitions that formally describe the application domain, through concepts, concept properties and relations, according to some particular view. Other ontologies provide structural and content-based description of multimedia data, similarly to the MPEG-7 standard. Garcia and Celma [15] have produced an OWL-Full ontology obtained through an automatic translation of MPEG-7; this approach has the limitation that computational complexity and decidability of reasoning are not guaranteed. Tsinaraki et al. [29] have manually developed an OWL-DL ontology that captures the full MPEG-7 Multimedia Description Schema (MDS) and the parts of the MPEG-7 video and audio schemas that are required for the complete representation of MDS. In [3] an OWL-DL ontology, designed to provide a high degree of axiomatization, ensuring interoperability through machine accessible semantics, and extensibility has been proposed. This ontology comprises parts of MPEG-7 descriptors such as visual low-level, spatio-temporal decomposition and media information descriptors.

Many researchers have proposed integrated systems where the ontology provides the conceptual view of the domain at the schema level, and appropriate classifiers play the role of observers of the real world sources and classify an observed entity or event in a concept of the ontology. Classifiers have the responsibility of implementing invariance with respect to several conditions that may change the appearance of entities, such as changes in illumination, geometric perspective, occlusion, etc. Once the observations are classified, the ontology is exploited to provide an organized semantic annotation and establishing links between concepts. Ebadollahi, Chang and Smith [12] performed detection of events of the LSCOM ontology. Events were viewed as stochastic temporal processes in the semantic concept space and their pattern was modeled as the collection of the confidences about the elementary concepts associated with the event, computed by the detectors. Snoek et al. [27] proposed a method to perform video annotation with the MediaMill 101 concept lexicon. In this work machine learning technique trains classifiers to detect high-level concepts from low-level features, while WordNet is used to derive high-level concepts relations in order to enhance the annotation performances. Zha et al. [31] have defined an ontology to provide some structure to the LSCOM-lite lexicon, using pairwise correlations between concepts and hierarchical relationships, to refine concept detection of SVM classifiers. Hauptmann et al. [16] proposed a framework to learn relationships between concepts by analysing the co-occurrences between concepts, so as to reinforce the detection made by the classifiers. A methodology for the analysis of low-level features and semantic properties of three flat concepts lexicons has been recently presented in [19] by Koskela, Smeaton et al., showing that modeling inter-concept relations can provide a promising resource for semantic analysis of multimedia data.

Other approaches have directly included in the ontology an explicit representation of the visual knowledge, to perform reasoning not only at the schema level but also at the data level. Bloehdorn et al. [7], defined a Visual Descriptors ontology, a Multimedia Structure ontology and a Domain ontology to perform video content annotation at semantic level. The Visual Descriptors ontology included concept instances represented with MPEG-7 visual descriptors. Dasiopoulou et al. [9] have included in the ontology instances of visual objects. They have used as descriptors qualitative attributes of perceptual properties like color homogeneity, low-level perceptual features like components distribution, and spatial relations. Semantic concepts have been derived from color clustering and reasoning. Maillot and Thonnat [22] have proposed a visual concept ontology that includes texture, color and spatial concepts and relations for object categorization. A set of classifiers for the recognition of visual concepts is trained using features extracted from a set of manually annotated and segmented samples.

The inclusion of data instances in the ontology requires some mechanism for the management of the ontology evolution. A solution was presented by Bertini et al. in [6], using generic and domain specific descriptors, identifying visual prototypes as representative elements of visual concepts and introducing mechanisms for their updating, as new instances of visual concepts are added to the ontology; the prototypes are used to classify events and objects observed in video sequences. Castano et al. [8] have addressed the problem of temporal evolution of ontologies at the schema and visual data level. Each visual instance is checked in order to determine whether it can be associated to the existing abstract concepts or a new concept has to be defined in the ontology. Evolution patterns have been proposed to define the kinds of action to be performed over the ontology: instance population, leveraging detected mid and high-level concepts relations to perform annotation or ontology evolution, defining new concepts and enriching the domain ontology with these new concepts and their relations.

In the attempt of having richer annotations, other authors have explored the usage of reasoning over multimedia ontologies. In this case spatio-temporal relationships between concept occurrences are analyzed so as to distinguish between scenes and events and provide more precise and comprehensive descriptions. Neumann and Möller [24] have proposed a framework for scene interpretation using Description Logic reasoning techniques over "aggregates"; these are composed of multiple parts and constrained by temporal and spatial relations to represent high-level concepts, such as objects configurations, events and episodes. In Espinosa et al. [13] manually annotated regions of images are used as visual representations of concepts, and relations between concept instances are obtained automatically. Inference from observation to explanation (abduction) is then used to check, among detected entities, relations and constraints that lead to consistent interpretation of image content. Jain et al. [20] have employed a two-level ontology of artistic concepts that includes visual concepts such as color and brushwork in the first level, and artist name, painting style and art period for the high-level concepts of the second level. A transductive inference framework has been used to annotate and disambiguate high-level concepts. In Staab et al. [10] automatically segmented image regions are modeled through low-level visual descriptors and associated to semantic concepts using manually labeled regions as training set. Context information is exploited to reduce annotation ambiguities. The labeled images are transformed into a constraint satisfaction problem (CSP), that can be solved using constraint reasoning techniques.

Several authors have exploited the ontology schema using rule-based reasoning over objects and events. Snoek et al. [28] performed annotation of sport highlights using rules that exploited face detection results, superimposed captions, teletext and excited speech recognition, and Allen's logic to model temporal relations between the concepts in the ontology. Francois et al. [14] defined a special formal language to define ontologies of events and used Allen's logic to model the relations between the temporal intervals of elementary concepts, so as to be able to assess complex events in video surveillance. Hollink et al. [17] defined a set of rules in SWRL to perform semi-automatic annotation of images of pancreatic cells. Bai et al. [5] defined a soccer ontology and applied temporal reasoning with temporal description logic to perform event annotation in soccer videos. All these methods have defined rules that are created by human experts; thus, these approaches are not practical for the definition of a large set of rules.

To overcome this problem some researchers have studied techniques to learn automatically a set of rules. Dorado et al. [11] performed video annotation based on learned rules that infer high-level concepts from low-level features using decision tree technique. Shyu et al. [26] proposed a method to annotate rare events and concepts based on set of rules

that use low-level and middle-level features. A decision tree algorithm is applied to the rule learning process. Moreover they addressed the imbalance problem of positive and negative examples in the case of rare event/concept using data mining techniques. Liu et al. [21] proposed a method to enhance accuracy of semantic concepts detection, using association mining techniques to imply the presence of a concept from the co-occurrence of other high-level concepts. None of these three works is based on ontologies and the type of rules that can be learned with these approaches can not be directly applied to an ontology-based framework. Moreover, these methods that learn a set of rules by exploiting decision tree algorithms and low-level features, or simple junctions of high-level concepts, are not enough expressive to describe complex concepts and in particular events.

For example consider the event *a person crosses a street from left to right*. This event can not be described using only the low-level descriptors of the person and of the street, or using the co-occurrence of the high-level concepts *person* and *street* since the person may stay always on a sidewalk in the same part of the street; instead it is required to take into account the temporal evolution of objects and entities, with their properties. This event can be fully described and modeled using first-order logic. A sentence that describes the event is: IF a person is in the left sidewalk of a street in the time interval t_1 AND the same person is in the right sidewalk of the same street in the time interval t_2 AND t_1 is before t_2 THEN that person has crossed the street from left to right; this sentence can be translated in the following fragment of first-order logic language:

$$IF\ person(p)\ \wedge\ personIsOnLeftSidewalk(p,t_1)\ \wedge$$
$$personIsOnRightSidewalk(p,t_1) \wedge before(t_1,t_2)$$
$$THEN\ personCrossTheStreetFromLeftToRight(p)$$

where p is a variable that can be bound to any person and t_1 and t_2 are variables that are used to represent time intervals.

In this paper we propose an adaptation of the First Order Inductive Learner technique (FOIL [25]) to the Semantic Web technologies to learn rules; for convenience this method will be referenced in the following as FOILS. The proposed method exploits the knowledge embedded into the ontology to learn new rules for describing video entities and events. The ontology used in this paper follows the Pictorially Enriched Ontology model [6], and includes: high-level concepts, concept properties and concept relations, used to define the semantic context of the examined domain; concept instances, with their visual descriptors, enrich the video semantic annotation. The learned rules, defined using the SWRL language, can be applied directly to an ontology defined using OWL to allow automatic semantic annotation of video sequences.

Moreover the learning approach used is more expressive than the previous methods because it defines rules through the first-order logic theory. To demonstrate applicability to automatic video annotation our approach has been tested to learn rules that model some events, defined in the LSCOM ontology, related to airplane entities and events related to the video surveillance domain.

3. ONTOLOGY RULES LEARNING

To describe correctly the learning algorithm, let us introduce some basic terminology from formal logic. All the expressions are composed of constants (e.g. *Joe*, *Boeing-747*), variables (e.g. x, y), predicate symbols (e.g. *HasTrajectory*, *GreaterThan*) and function symbols (e.g. *duration*). The difference between predicates and functions is that predicates can assume only boolean values, whereas functions may have any constant as their value. In the following we will use lowercase for functions and capitalized symbols for predicates. A term is any constant, any variable, or any function applied to any term. A literal is any predicate or its negation applied to any term. If a literal contains a negation symbol ($\neg$), it is called *negative literal*, otherwise *positive literal*. A *clause* is any disjunction of literals, where all variables are assumed to be universally quantified. A *Horn clause* is a clause containing at most one positive literal, as shown in the following:

$$H \vee \neg L_1 \vee \neg L_2 \ldots \vee \neg L_n$$

where H is the positive literal, and $\neg L_1 \vee \neg L_2 \ldots \vee \neg L_n$ are negative literals. It is equivalent to:

$$(L_1 \wedge L_2 \ldots \wedge L_n) \rightarrow H$$

which is equivalent to the following:

$$IF\ (L_1 \wedge L_2 \ldots \wedge L_n)\ THEN\ H$$

The Horn clause precondition $L_1 \wedge L_2 \ldots \wedge L_n$ is called *body*; the literal H that forms the post-condition is called the *head*.

FOILS is an adaptation of the FOIL algorithm to the SWRL standard. Similarly to FOIL, the hypotheses learned by FOILS are sets of first-order rules, where each rule is similar to a Horn clause, with the limitation that literals can not contain function symbols, in order to reduce the complexity of the search in the hypothesis space. The algorithm starts with an initial rule, composed by the *head* that we want to find in the rule and an empty or initial *body*. The algorithm iterates searching new literals that have to be added to the *body* of the rule. The search is a general-to-specific search through the space of hypotheses, beginning with the most general preconditions possible (the empty or initial precondition), and adding literals one at a time to specialize the rule until it avoids all negative examples, or when no more negative examples are excluded for a certain number of loops l. A schema of the algorithm is shown in Alg. 1; in our experiments l has been set to 3.
Two issues have to be addressed: the generation of hypothesis candidates and the choice of the most promising candidate.

Algorithm 1 FOILS algorithm schematization

```
Pos ← Positive examples
Neg ← Negative examples
Rule ← Initial rule
repeat
    Candidate_literals ← Generating hypothesis candidates
    Best_literal ← arg max_L Rule_Gain(L,Rule)
    Add Best_literal to Rule preconditions
    Pos ← subset of Positive examples that satisfy Rule
    Neg ← subset of Negative examples that does not satisfy Rule
until Neg is empty or no more Neg examples are excluded for l loops
```

3.1 Generating hypothesis candidates

Suppose that the current rule being considered is:

$$(L_1 \wedge L_2 \ldots \wedge L_n) \rightarrow P(x_1, x_2, \ldots, x_k)$$

where $(L_1 \wedge L_2 \ldots \wedge L_n)$ are literals forming the current rule preconditions and where $P(x_1, x_2, \ldots, x_k)$ is the literal that form the rule *head*. FOILS generates candidate specializations of this rule by considering new literals L_{n+1} that fit one of the following forms:

- $Q(v_1, \ldots, v_r)$ where Q is any predicate name occurring in the ontology and where the v_i are either a new variable or a variable already present in the rule. At least one of the v_i in the created literal must already exist as a variable in the rule.
- $Equal(x_j, x_k)$ where x_j and x_k are variables already present in the rule.

We observe that in FOIL there is another rule for generation of new candidates: it is the negation of either the above form of literals. This rule can not be exploited in our algorithm because it is not permitted by SWRL.

3.2 Most promising literal

To select the most promising literal from the candidates generated at each step, FOILS, similarly to FOIL, considers the performance of the rule over the training data. The evaluation function used to estimate the utility of adding a new literal is based on the number of positive and negative bindings covered before and after adding the new literal. More precisely consider a rule R and a candidate literal L that might be added to the *body* of R. Let R' be the rule created by adding the literal L to rule R. The value of adding L to R is defined as:

$$Rule\_Gain(L, R) \equiv t \left(log_2 \frac{p_1}{p_1 + n_1} - log_2 \frac{p_0}{p_0 + n_0} \right)$$

where p_0 is the number of positive bindings of rule R, n_0 is the number of negative bindings of R, p_1 is the number of positive bindings of rule R' and n_1 is the number of negative bindings of R'. Finally, t is the number of positive binding of rule R that are still covered after adding literal L to R. When a new variable is introduced into R by adding L, then any original binding is considered to be covered as long as some binding extending it is present in the bindings of R'.

4. PICTORIALLY ENRICHED ONTOLOGY

The ontology used in this work follows the Pictorially Enriched Ontology model, presented in [6]. In this model the ontology contains linguistic concepts, their relationships and instances of visual concepts. The linguistic concepts can be related to concrete concepts, that represent entities and events that have some visual manifestation in the reality, or can be abstract concepts, i.e. are related to more immaterial elements. The concept instances are related to the concrete concepts of the schema, and include object identifiers, time information, sets of visual descriptors and link to the raw multimedia data.

These instances are included in the ontology and used to select matching references for the visual descriptors of the entities observed in videos. These descriptors, that may be generic or domain specific, are then used to characterize concepts instances; this characterization allows to select the most representative concepts as visual prototypes of a concept, and allow to perform reasoning based on the visual appearance of a concept. This approach is used when the entities exhibit a variety of complex changes in shape or visual appearance. External classifiers are instead used to assess the presence of concepts that refer to entities with little changes in their appearance, such as human faces or body. In both cases the instances and their visual descriptors are included to allow reasoning on their descriptors values. In Fig. 1 is shown a simplified view of the main concepts used to model the events used in the experiments; for the sake of simplicity the visual descriptors associated to the concepts are not reported. Since both concepts and concept instances in the ontology are defined using OWL, SWRL learned rules can be used to effectively perform reasoning over both concepts and concept instances, so as to disambiguate the results of the classification or derive new semantic annotations.

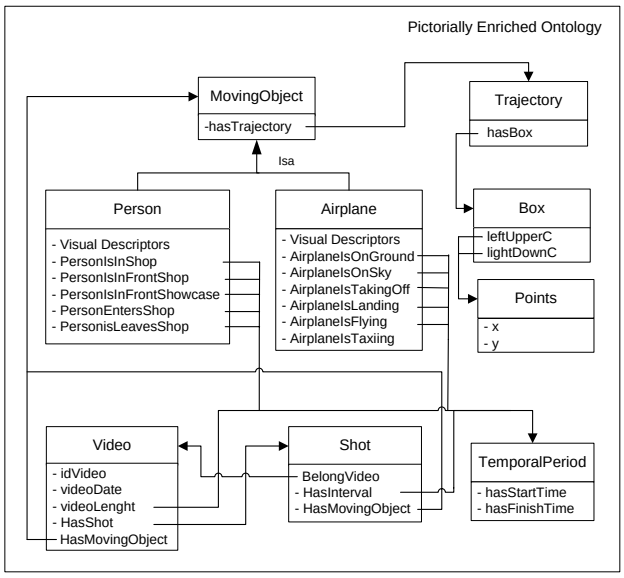

Figure 1: Simplified view of the Pictorially Enriched Ontology used in the experiments.

Video segmentation involves temporal partitioning of the video into units which serve as the basis for descriptor extraction and semantic annotation. In this work, shots are adopted as the basic syntactic unit, while video clips (video sequences possibly composed by more than one shot) are used as annotation units.

5. EXPERIMENTAL RESULTS

We have applied the proposed method to learn rules that describe actions and events related to two different domains: news videos and surveillance videos. For the first domain we have considered four events related to airplanes: airplane flying, airplane takeoff, airplane landing, airplane taxiing. These events are selected from the revised list of LSCOM events/activities [18]. The events related to the video surveillance domain are: person enters in a specific area and person leaves from it. In particular, in our experiments the specific area is a mall shop.

These events can be recognized using the detection results of some particular classifier, the variation of their spatio-temporal relationships over time and context information. In the case of the airplane actions the classifiers used are: airplane, sky and ground detectors. For the other domain we have used a person detector. The airplane detector has been created using the Viola&Jones object detector [30] whereas the person detector used is the one available in the OpenCV library.
The positive and negative examples used to train the airplane detector have been selected from standard image datasets such as Caltech, VOC2005 and VOC2006. The negative examples used are images of man-made objects (e.g. other vehicles like cars, buses and motorcycles), outdoor scenes, animals and persons, various objects. The sky and ground detectors implemented are not used to classify all the parts and segments of each frame, but only locally, next to the airplane position because it is enough to know if the airplane is on ground or in the sky. The sky/ground detector evaluates statistical parameters of the luminance of the blobs around the detected airplane.

Finally using a tracker, based on an improved version of the particle filter [4], we can determine the temporal evolution of the trajectory. In our experiments both the person and airplane concepts are associated with color and luminance histograms, that are used by the tracker to identify the instances of the concepts in a video sequence. For each moving object its bounding box trajectories is inserted in the ontology, after performing a Gaussian smoothing, to reduce measurement noise. In Fig. 2 two sequences showing an airplane take-off and a person entering in a shop are shown.

In the first part of the experiments we evaluate the performance of the airplane detector. We have trained five different detectors, using five configurations, with different numbers of positive and negative examples, image window sizes, and learning steps. Results are reported in Tab. 1. To train the fifth detector the number of positive examples of airplanes has been increased, adding more images of frontal and rear views of airplanes. The first three detectors did not provide an acceptable performance in terms of precision, as shown in the table. The decrease of the precision value between the fourth and fifth detector is mainly due to the fact that the detector may provide multiple detections for the same airplane, whose bounding boxes are overlapping, and these multiple detections have been counted as falses; without considering this overlapping effect the precision is comparable with that of the fourth detector. Considering this fact, the fifth detector has been selected and used in the following experiments.

To test the effectiveness of the learned rules we have used them to recognize events in a large dataset, that comprises 100 videos containing airplane events taken from the web[1], 65 Trecvid 2005 videos and videos of the public CAVIAR[2] surveillance videos dataset, selected from the front views of the 2^{nd} set.

The set of videos selected from the web video sharing sites (called in the following as Web Dataset) is available online,

[1]YouTube (http://www.youtube.com),
Alice Video (http://dailymotion.alice.it),
PlanesTV (http://www.planestv.com/planestv.html),
Yahoo! Video (http://it.video.yahoo.com)
[2]CAVIAR Dataset
(http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/)

Figure 3: Surveillance video dataset: view of the mall shop.

along with the airplane detector[3]. The Trecvid videos were selected from those of the TrecVid development set that are reported to contain the LSCOM concepts *airplane_takeoff*, *airplane_landing* and *airplane_flying*, after a manual inspection that eliminated some errors of the ground truth (e.g. videos that contained rockets or helicopters instead of airplanes). Since the concept *airplane taxiing* is not defined in LSCOM we inspected the videos annotated as containing *airplane* to select some videos that contained this event. The videos of the CAVIAR dataset have been filmed from a fixed position camera that frames a mall shop and the area in front of the shop. In the experiments the scene framed has been divided in three parts as shown in Fig. 3 to determine when a person is in the shop, in front of it or in front of the showcase of the shop.

We have used an implementation of the FOILS algorithm, described in Sect. 3, to learn SWRL rules that model events. To illustrate how the FOILS algorithm works we consider, for example, the target literal *person enters in a shop*. The process starts with an initial rule written in SWRL:

$$Person(?p) \wedge Clip(?c) \rightarrow PersonEntersInShop(?p, ?c)$$

The initial candidates are all the classes and properties defined in the ontology domain and temporal properties used to encode Allen's logic. At each step the most promising literal is added, considering the performance of the rules over the training data until the recognition performance does not improve. For each event we randomly select one third of the videos containing that event as positive examples, and one third of the videos of the other events as negative examples. In Tab. 2 and 3 the learned rules are shown. For each rule we present the initial rule and the final rule obtained using FOILS. The learned rules recognize events within clips; this allows to cope with the case in which an event is shown using more than one shot. In some cases we can observe that FOILS learns some literals that are not necessary for the event definition, however this does not affect negatively the performance of the rule. This fact may happen since FOILS does not take into account the structure of the ontology; an example is the $MovingObject(?p)$ literal in the landing and taking-off rules, that is not necessary due to the fact that in our ontology this concept is an hypernym of airplane.

We have then applied the rules to the videos, evaluating

[3]http://www.micc.unifi.it/dome

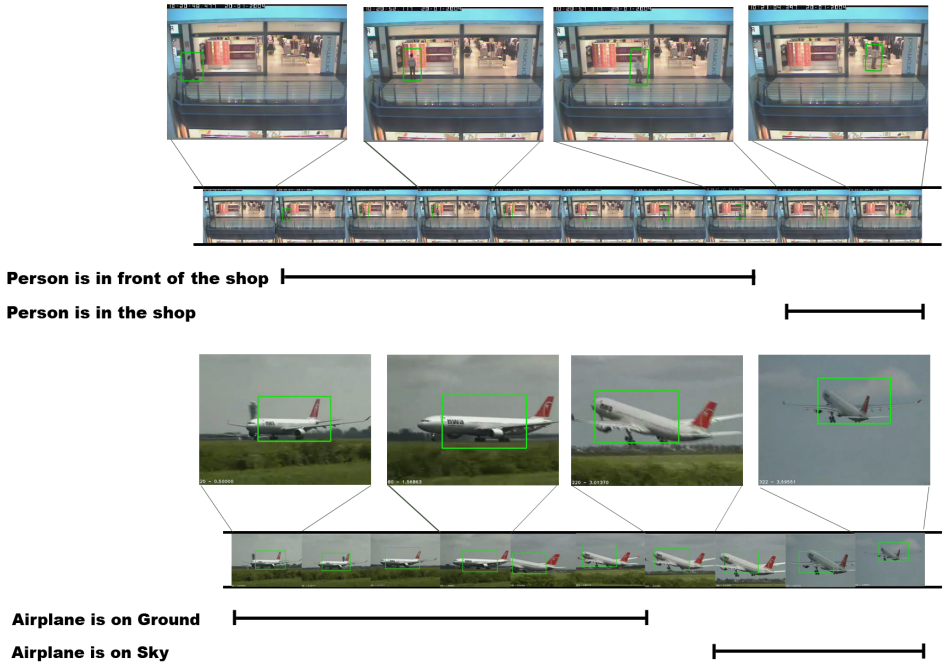

Figure 2: Examples of airplane and person detection and tracking in video sequences. The temporal sequence of basic object properties is shown.

N. detector	N. steps	Neg. examples	Pos. examples	Window	Precision	Recall
1	17	3000	800	50×30	0.20	0.74
2	18	1500	800	50×30	0.19	0.83
3	20	1500	800	50×30	0.32	0.65
4	20	1500	800	25×10	0.75	0.55
5	22	1500	1040	50×30	0.41	0.66

Table 1: Precision and recall of airplane detector.

the results, in term of precision and recall, for all datasets, as show in Tab. 4. As it can be observed the overall results for all the rules are extremely promising. Since the rules that describe flying and landing are more simple, their performance is better than that of the rules that model landing and taking-off. The main difference in the performance results between the Web Dataset and Trecvid videos is related to the quality of the images and to the presence of superimposed graphics, that were present only in the Trecvid news videos. Since the performance of the rules is dependent on the performance of the detectors and tracker we have investigated the cases in which the rules failed. In the news video domain the main cause of failure is due to the performance of the simple sky/ground detector, that uses only the luminance information. In a few cases the fault was the airplane detector, especially when superimposed graphics and text covered the appearance of the airplane. The results of the recognition of video surveillance actions show a good performance both in terms of precision and recall. This is due to the type of the scene and events in the dataset: the fixed camera and lighting conditions reduce the variability of the appearance of the observed events and objects. This lead to have a good performance of the person detectors and of the tracker. The recall performance is mainly dependent on the errors of the tracker, that may happen when multiple persons' trajectories overlap.

6. CONCLUSIONS

In this paper we have presented an overview of approaches and algorithms that exploit ontologies to perform semantic video annotation and an adaptation of the First Order Inductive Learner technique to the Semantic Web Rule Language. This technique exploits the knowledge embedded into an ontology to automatically learn a set of rules that describe events and use them to perform automatic video annotation. The proposed approach has been tested using different datasets and domains to demonstrate the approach can be generalized. Our future work will investigate techniques to incorporate learning of constants and function symbols, to permit to insert numerical temporal specifications in the concept description.

Acknowledgments. This work is partially supported by the Information Society Technologies (IST) Program of the European Commission as part of the VIDI-Video project.

7. REFERENCES

[1] Dublin Core Metadata Initiative - http://dublincore.org/.

Rule: Airplane TakingOff
Initial rule: $Airplane(?a) \wedge Clip(?c) \rightarrow AirplaneIsTakingOff(?a, ?c)$ Result rule: $Airplane(?a) \wedge Clip(?c) \wedge AirplaneIsOnSky(?a, ?g1) \wedge AirplaneIsOnGround(?a, ?g2) \wedge$ $Temporal : after(?g1, ?g2) \wedge HasTemporalPeriod(?c, ?g3) \wedge Temporal : contains(?g3, ?g1) \wedge$ $Temporal : contains(?g3, ?g2) \wedge MovingObject(?a) \rightarrow AirplaneIsTakingOff(?a, ?c)$
Rule: Airplane Landing
Initial rule: $Airplane(?a) \wedge Clip(?c) \rightarrow AirplaneIsLanding(?a, ?c)$ Result rule: $Airplane(?a) \wedge Clip(?c) \wedge AirplaneIsOnSky(?a, ?g1) \wedge AirplaneIsOnGround(?a, ?g2) \wedge$ $Temporal : notafter(?g1, ?g2) \wedge HasTemporalPeriod(?c, ?g3) \wedge Temporal : contains(?g3, ?g1) \wedge$ $Temporal : contains(?g3, ?g2) \wedge MovingObject(?a) \rightarrow AirplaneIsLanding(?a, ?c)$
Rule: Airplane Flying
Initial rule: $Airplane(?a) \wedge Clip(?c) \rightarrow AirplaneFlying(?a, ?c)$ Result rule: $Airplane(?a) \wedge Clip(?c) \wedge AirplaneIsOnSky(?a, ?g1) \wedge$ $HasTemporalPeriod(?c, ?g2) \wedge Temporal : contains(?g2, ?g1) \rightarrow AirplaneIsFlying(?a, ?c)$
Rule: Airplane Taxiing
Initial rule: $Airplane(?a) \wedge Clip(?c) \rightarrow AirplaneIsTaxiing(?a, ?c)$ Result rule: $Airplane(?a) \wedge Clip(?c) \wedge AirplaneIsOnGround(?a, ?g1) \wedge$ $HasTemporalPeriod(?c, ?g2) \wedge Temporal : contains(?g2, ?g1) \rightarrow AirplaneIsTaxiing(?a, ?c)$

Table 2: Rules for airplane events recognition, obtained using FOILS.

Rule: PersonEntersShop
Initial rule: $Person(?p) \wedge Clip(?c) \rightarrow PersonEntersShop(?p, ?c)$ Result rule: $Person(?p) \wedge Clip(?c) \wedge PersonIsInFrontShop(?p, ?g1) \wedge PersonIsInShop(?p, ?g2) \wedge$ $Temporal : notOverlaps(?g2, ?g1) \wedge Temporal : notBefore(?g2, ?g1) \wedge$ $Temporal : notMetBy(?g2, ?g1) \wedge HasTemporalPeriod(?c, ?g3) \wedge$ $Temporal : contains(?g3, ?g1) \wedge Temporal : contains(?g3, ?g2) \rightarrow PersonEntersShop(?p, ?c)$
Rule: PersonLeavesShop
Initial rule: $Person(?p) \wedge Clip(?c) \rightarrow PersonLeavesShop(?p, ?c)$ Result rule: $Person(?p) \wedge Clip(?c) \wedge PersonIsInFrontShop(?p, ?g1) \wedge PersonIsInShop(?p, ?g2) \wedge$ $Temporal : notAfter(?g2, ?g1) \wedge Temporal : notContains(?g1, ?g2) \wedge$ $HasTemporalPeriod(?c, ?g3) \wedge Temporal : contains(?g3, ?g1) \wedge$ $Temporal : contains(?g3, ?g2) \rightarrow PersonLeavesShop(?p, ?c)$

Table 3: Rules for human action recognition, obtained using FOILS.

[2] TV Anytime Forum - http://www.tv-anytime.org/.

[3] R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura. Comm: Designing a well-founded multimedia ontology for the web. In *Proc. of Int'l Semantic Web Conference*, 2007.

[4] A. D. Bagdanov, A. Del Bimbo, F. Dini, and W. Nunziati. Improving the robustness of particle filter-based visual trackers using online parameter adaptation. In *Proc. of IEEE Int'l Conference on Advanced Video and Signal Based Surveillance*, 2007.

[5] L. Bai, S. Lao, G. Jones, and A. F. Smeaton. Video semantic content analysis based on ontology. In *Proc. of Int'l Machine Vision and Image Processing Conference*, 2007.

[6] M. Bertini, A. Del Bimbo, C. Torniai, R. Cucchiara, and C. Grana. Dynamic pictorial ontologies for video digital libraries annotation. In *Proc. ACM Int'l Workshop on the Many Faces of Multimedia Semantics*, 2007.

[7] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, I. Kompatsiaris, S. Staab, and M. Strintzis. Semantic annotation of images and videos for multimedia analysis. In *Proc. of European Semantic Web Conference*, 2005.

[8] S. Castano, S. Espinosa, A. Ferrara, V. Karkaletsis, A. Kaya, S. Melzer, R. Moller, S. Montanelli, and G. Petasis. Ontology dynamics with multimedia

Data Set	Action/Event	Precision	Recall
Web Dataset	Airplane flying	0.96	0.94
Web Dataset	Airplane takeoff	0.76	0.80
Web Dataset	Airplane landing	0.84	0.96
Web Dataset	Airplane taxiing	1	0.76
TRECVID 2005	Airplane flying	0.94	0.5
TRECVID 2005	Airplane takeoff	0.3	0.42
TRECVID 2005	Airplane landing	0.66	0.66
TRECVID 2005	Airplane taxiing	1	0.76
Web Dataset + TRECVID 2005	Airplane flying	0.96	0.71
Web Dataset + TRECVID 2005	Airplane takeoff	0.61	0.70
Web Dataset + TRECVID 2005	Airplane landing	0.90	0.90
Web Dataset + TRECVID 2005	Airplane taxiing	0.94	0.84
CAVIAR	Person enters in the shop	1	0.77
CAVIAR	Person leaves from the shop	1	0.88

Table 4: Precision and recall of Airplane flying, Airplane takeoff, Airplane landing, Airplane taxiing, Person enters in the shop, Person leaves from the shop for different datasets.

information: The boemie evolution methodology. In *Proc. of Int'l Workshop on Ontology Dynamics*, 2007.

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

[10] S. Dasiopoulou, C. Saathoff, P. Mylonas, Y. Avrithis, Y. Kompatsiaris, S. Staab, and M. Strintzis. *Semantic Multimedia and Ontologies Theory and Applications*, chapter Introducing Context and Reasoning in Visual Content Analysis: An Ontology-Based Framework, pages 99–122. Springer, 2008.

[11] A. Dorado, J. Calic, and E. Izquierdo. A rule-based video annotation system. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):622–633, May 2004.

[12] S. Ebadollahi, L. Xie, S.-F. Chang, and J. Smith. Visual event detection using multi-dimensional concept dynamics. In *Proc. of IEEE Int'l Conference on Multimedia & Expo*, 2006.

[13] S. Espinosa, A. Kaya, S. Melzer, R. Moller, and M. Wessel. Towards a media interpretation framework for the semantic web. In *Proc. of Int'l Conference on Web Intelligence*, 2007.

[14] A. Francois, R. Nevatia, J. Hobbs, R. Bolles, and J. Smith. VERL: an ontology framework for representing and annotating video events. *IEEE Multimedia*, 12(4):76–86, Oct-Dec. 2005.

[15] R. Garcia and O. Celma. Semantic integration and retrieval of multimedia metadata. In *Proc. of the Knowledge Markup and Semantic Annotation Workshop*, 2005.

[16] A. Hauptmann, M. Chen, M.-Y.and Christel, W.-H. Lin, and J. Yang. A hybrid approach to improving semantic extraction of news video. In *Proc. of IEEE Int'l Conference on Semantic Computing*, 2007.

[17] L. Hollink, S. Little, and J. Hunter. Evaluating the application of semantic inferencing rules to image annotation. In *Proc. of Int'l Conference on Knowledge Capture*, 2005.

[18] L. Kennedy. Revision of LSCOM event/activity annotations, DTO challenge workshop on large scale concept ontology for multimedia. Advent technical report #221-2006-7, Columbia University, 2006.

[19] M. Koskela, A. F. Smeaton, and J. Laaksonen. Measuring concept similarities in multimedia ontologies: Analysis and evaluation. *IEEE Transactions on Multimedia*, 9(5):912–922, August 2007.

[20] L. Leslie, T.-S. Chua, and J. Ramesh. Annotation of paintings with high-level semantic concepts using transductive inference and ontology-based concept disambiguation. In *Proc. ACM Multimedia*, 2007.

[21] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen. Association and temporal rule mining for post-filtering of semantic concept detection in video. *Multimedia, IEEE Transactions on*, 10(2):240–251, Feb. 2008.

[22] N. Maillot and M. Thonnat. Ontology based complex object recognition. *Image Vision Computing*, 26(1):102–113, 2008.

[23] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, July-Sept. 2006.

[24] B. Neumann and R. Moeller. On scene interpretation with description logics. In *Cognitive Vision Systems: Sampling the Spectrum of Approaches*, LNCS, pages 247–278. Springer, 2006.

[25] J. R. Quinlan. Learning logical definitions from relations. *Machine Learning*, 5(3):239–266, 1990.

[26] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen. Video semantic event/concept detection using a subspace-based multimedia data mining framework. *Multimedia, IEEE Transactions on*, 10(2):252–259, Feb. 2008.

[27] C. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. 9(5):975–986, Aug. 2007.

[28] C. Snoek and M. Worring. Multimedia event-based video indexing multimedia event-based video indexing using time intervals. *IEEE Transactions on Multimedia*, 7(4):638–647, 2005.

[29] C. Tsinaraki, P. Polydoros, F. Kazasis, and S. Christodoulakis. Ontology-based semantic indexing for MPEG-7 and TV-Anytime audiovisual content. *Multimedia Tools and Applications*, (26):299–325, Aug. 2005.

[30] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[31] Z.-J. Zha, T. Mei, Z. Wang, and X.-S. Hua. Building a comprehensive ontology to refine video concept detection. In *Proc. of ACM Int'l Workshop on Multimedia Information Retrieval*, 2007.