Video Event Annotation using Ontologies with Temporal Reasoning

Marco Bertini, Alberto Del Bimbo, Giuseppe Serra
University of Florence - Italy
Via S. Marta, 3 - 50139 Firenze
{bertini,delbimbo,serra}@dsi.unifi.it

Abstract

Annotation and retrieval tools for multimedia digital libraries have to cope with the complexity of multimedia content. In particular, when dealing with video content, annotation and retrieval tools have to use appropriate knowledge structures that can effectively relate high level concepts to low and mid level visual features and, at the same time, integrate temporal information which is crucial when defining an abstract model for video. In this paper we present a multimedia ontologies that include both linguistic and visual ontology. Moreover provided that appropriate low level descriptors are used to detect simple events, subjects or objects, we propose usage of Semantic Web Rule Language in order to provide a formal definition of complex events based on temporal relations between simple entities. Results for complex event inferencing are shown for the news broadcast domain.

1. Introduction

Digital video is the media that has probably the highest importance in the construction and management of future digital libraries. In fact every day broadcasters, media companies, government institution and also individuals are producing large amounts of digital video data for news, personal entertainment, educational and institutional purposes. In order to be able to exploit the large video digital libraries that are currently being created, new tools and methodologies that allow annotation, retrieval and management have to be developed.

Effective examples of retrieval by content of video clips using textual keywords have been presented for news [3, 7, 12] and sports video domains [4, 16].

But for a richer annotation of digital video are required that more complex linguistic structures are to be used to represent knowledge about video at a deeper semantic level. Ontologies are defined as the representation of the semantics of terms and their relationships. They consist of concepts, concept properties, and relationships between concepts, all expressed in linguistic terms [6]. For the digital video domain, ontologies are used to describe either the video content domain or the structure of the media. In the first case they are static descriptions of entities and highlights present in the video and their relationships, as codified by human experience; in the second case they can describe the structure of the media, i.e. the component elements of the video, the operations allowed on its parts, and the low-level video descriptors that characterize their content.

However traditional domain ontologies, whether in the form of textual metadata or linguistic abstractions and relations defined on primitive video elements, are substantially inadequate to support complete annotation and retrieval by content of video documents.

For this reason ontologies has be enriched to include structural video information and visual data descriptors, growing the representation upwards, in a sense. In [14], a visual descriptors concepts ontology and a multimedia video structure ontology, respectively based on MPEG-7 visual descriptors and MPEG-7 multimedia description schema, are used together with a domain ontology in order to support video content annotation. Jaimes et al. [9] suggested that concepts that relate to perceptual facts be categorized into classes using *modal keywords*, i.e. keywords that represent perceptual concepts in several categories. This is a key observation that can be used to great advantage once we have a method to classify phenomena into perceptual categories of the domain. In [2], qualitative attributes that refer to perceptual properties like color homogeneity, low-level perceptual features like model components distribution, and spatial relations were included in the ontology. Semantic concepts of video objects were derived from color clustering and reasoning. In [13] the authors have presented video annotation and retrieval based on high-level concepts derived from machine learned concept detectors that exploit low level visual features. The ontology includes both semantic descriptions and structure of concepts and their lexical relationships, obtained from WordNet.

These solutions for annotation and retrieval have focused on static concepts or entities exploiting the same usage of visual features employed for annotation of images without taking into account the temporal dimension that is the main characteristic of video content. The temporal dimension and the consequent temporal evolution and relationships between concepts and entities are key features that have to be taken into account when dealing with annotation and retrieval of video content [10].

For example, consider the problem of recognizing and characterizing *anchor persons* and *interview* events in news videos. Using only visual features extracted from key frames it may be difficult or impractical to distinguish between the two events. In fact, when looking at two key frames of these events without any knowledge of the context from which the frames have been taken (e.g. who is the person speaking, what happened before the moment that is represented, what is going to happen next, etc.) they can only be classified as *person speaking*. Instead, when context information is added (e.g. a known anchorman is recognized or temporal relationships between previous and following events are known) the proper classification of the two different events can be performed.

Authors have proposed different approaches for handling time information in videos based both on events definition and formal temporal relations conceptualizations. In [5] a formal language for describing an ontology of events (VERL) and a companion language to annotate instances of the events described in VERL are proposed. Allen's interval algebra is used to describe the relations among the temporal intervals in which events occur and example of complex events definition are provided for the video surveillance domain. Haghi et al. [11] introduce a framework for video annotation based on a temporal ontology. The ontology is expressed in temporal RDF and simple queries, taking into account temporal relationships of events, can be performed.

In this paper for video event annotation, we proposed a multimedia ontology defined by both linguistic and visual ontologies extended with a formal definition of complex events through standard languages (OWL) and standard rule based inference (SWRL) for annotation of video content. Differently from the other approaches that employed ad hoc languages or rules for video events definition, our work provides a formal definition of complex events using standard languages. Semantic Web Rule Language (SWRL) [8] based on a video and a domain ontology expressed in OWL is used. This brings two main advances: the events definitions can be easily shared with other different ontologies or applications, and the available inference engines can be used to detect complex events. In particular we present a solution for the formal definition of complex events based on occurrences and Allen temporal relationships between simple events. Provided that appropriate low level descriptors are used to detect simple events, subjects or objects. we propose usage of Semantic Web Rule Language in order to provide a formal definition of complex events based on temporal relations between simple entities. The main contribution of our work is the demonstration of effective usage of Multimedia Ontologies . Results for a set of complex events for broadcast news domain are reported.

The organization of the paper is as follows. In the next section structure and the definition of a multimedia ontology for the news video domain are presented. In section 3 the descriptors used to identify simple concepts of the ontology are described. In section 4 the definition of complex events is formalized. We conclude in section 5 with preliminary results for complex events detection in news video domain.

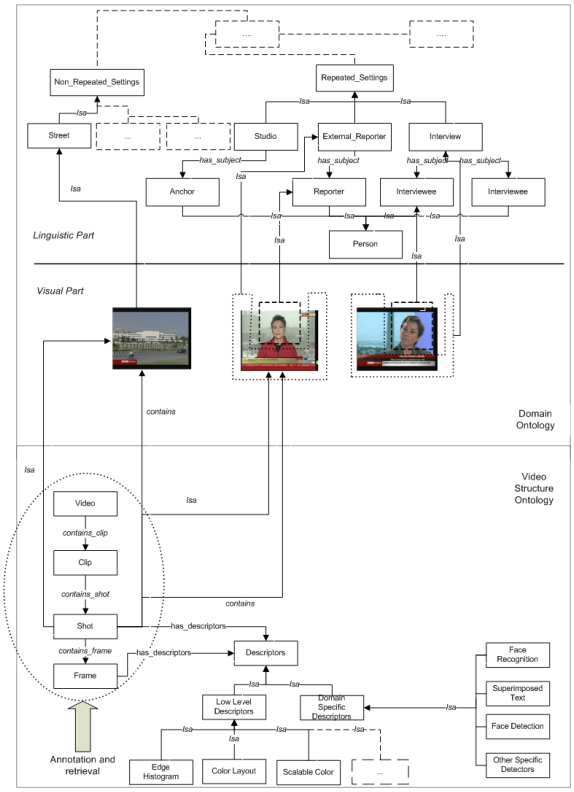


Figure 1. A simplified view of a Multimedia Ontology scheme used to model part of the news broadcast domain

2 Ontology extended with multimedia and temporal features

In figure 1 a simplified view of a multimedia ontology that can be used to model part of the news broadcast domain is shown. The multimedia ontology is composed of a

video structure ontology and a domain ontology. The latter comprises a traditional linguistic part modeling classes, concepts and their relations. The visual part contains shots and clips representing visual instances of linguistic concepts. They link the domain concepts to the structural parts of the video, modeled in the video structure ontology using the appropriate multimedia descriptors.

However, using static multimedia features only (e.g. features computed from a keyframe) it may be difficult to define and recognize dynamic or complex concepts and events (e.g. a person walking, airplane takeoff, etc.). In order to solve this problem we further extend the ontology with the possibility of expressing the occurrence of simple and static concepts, and with temporal relationships based on the Allen logic. An example is that of airplane takeoff that can be described using the Allen operator "meets" applied to the "airplane on the ground" and "airplane flying"; it should be noted that the latter two concepts can be described using static multimedia descriptors.

In order to evaluate our approach we have defined a simple multimedia ontology that comprises videos, objects, subjects and events. The video part of the ontology models the structural aspects such as shots, clips, frames and multimedia descriptors. Objects and subjects are used to define the inanimate objects and actors of the video. Events comprise simple and complex events: a simple event is an event that can be detected and recognized using static or simple multimedia features (e.g. a person talking can be detected using face detection and a simple sound classifier), while complex events can not be directly recognized but require the use of temporal relations and occurrences of simple and complex events (e.g. an external report sequence can be described as a composition of different events such as shots of the anchor, shots of the external reporter possibly interleaved with report scenes, etc.). The multimedia features used in the ontology are low-level descriptors, such as MPEG-7 color and texture descriptors, and higher-level features such as face detectors and recognition.

3 Modeling perceptual concepts

Descriptors of perceptual concepts regard image regions in frames (for entities and subjects) or sequences of frames (for scenes, highlights and events). They include color, texture or pattern descriptors, and their temporal distribution. We model perceptual observations in news videos at three levels: the scene setting, scene subjects (i.e. the actors in the scene), and simple events that can occur in a scene.

3.1 Scene setting modeling

We model four of the most common types of scene settings in news videos (see Fig. 2):

- **studio setting:** the location in which one or two anchorperson are framed, its appearance varies greatly also within the same broadcaster, and also changes in time. Due to the news video structure it is shown several times within the video;
- **external reporter setting:** the location, external from the studio setting, from which a broadcaster journalist reports. It can be a city street, a landscape, a building, etc. Often, since the reporter is framed using a medium view, the setting appears blurred. Typically the setting is shown several times during the report;
- **interview setting:** the location where an interview is taking place, usually it is a room or an external studio, but it can be sometimes also a public location, the setting is usually shown several times, possibly from different angles, to show the questions and answers of the interviewee; and
- **report setting:** the location shown during the external reports. It can be a city street, landscape, building, etc, and compared to the external reporter setting is much more varied, the people usually is not clearly evidenced, and the setting is seldom shown more than once.

These observations suggest that the different settings can be divided in two classes: the settings that appear more than once and those which do not. Other than that it is evident that it is extremely hard to visually characterize them, and the number of their occurrences and the temporal relations of their appearances have to be taken into account to distinguish them.

Following our previous work [1] we exploit generic features such as global color features, layout of homogeneous colour areas, and edge features (defined in the MPEG-7 standard for multimedia content description), to detect the settings that are shown more than once, and group them in different clusters.

3.2 Scene subject modeling

Anchorpersons, reporters, interviewees and persons in general are important subjects for semantic annotation and content-based retrieval of video clips and episodes, and are usually shown using close-up and medium views sequences.

Generally speaking, identification of people is a hard task, and in practice only close-up and medium views are useful, since persons can be automatically identified by exploiting face information.

To detect faces, we use a slightly modified version of the AdaBoost face detector of [15]. Due to the extremely high variability of faces shown in news videos we do not perform direct face recognition, but rather cluster similar

(a) A studio setting (b) An interview setting: interviewee (c) An interview setting: interviewer and interviewee (d) An interview setting: interviewer

(e) External reporter setting (f) A report setting

Figure 2. Keyframes of shots showing different settings.

faces shown within a news video, using the Haar features used to detect the faces.

3.3 Simple events modeling

The most common events shown in news videos are the monologue of an interviewer or interviewee, or one or two anchorpersons talking. These simple events can be described as a person talking in a setting that is shown more than once within the news program. Their detection and recognition can be performed combining the results of the settings detector with the person detector described in the previous sections, e.g. to avoid to classify two similar scenes such as a CGI shots (see Fig.3.2 *d* and *i*) as possible interviews.

These simple events, if combined together taking into account temporal relations, may be used to detect more complex events, such as the anchormen shots and the interview sequences, as described in the following section.

4 Complex event definition

Analyzing the structure of news broadcast it can be noticed, for instance, that we can state that an **anchorman** is framed if there is a **person talking** in a **studio setting**. If two different **persons** are framed alternatively in a sequence we can argue that there is an interview going on.

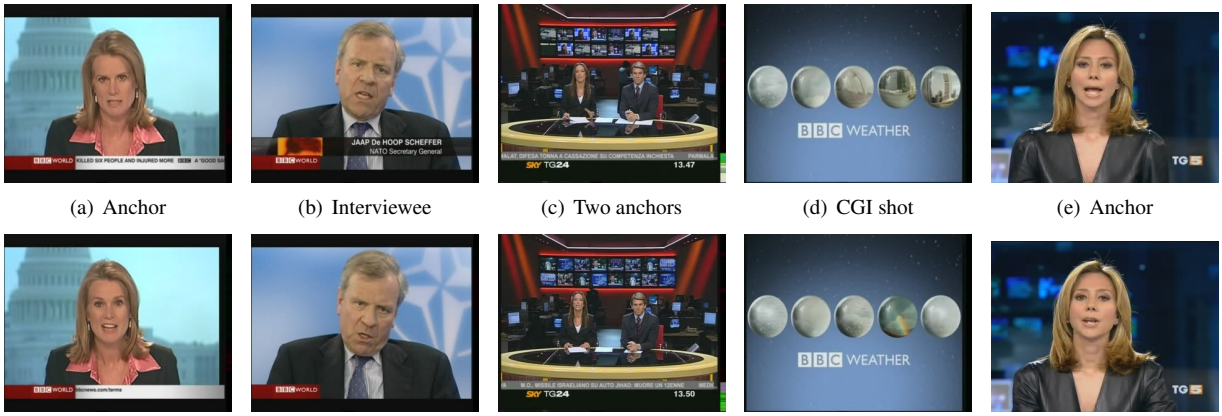
As described in Sect. 3 we have started by taking into account simple events such as "person", and "settings" which we were able to detect using the appropriate descriptors and trying to define other complex concepts or events exploiting temporal relations between the detectable concepts. Each instance of events, subjects or concepts that is detected has its proper time interval represented by the value of `hasTimeInterval` property containing an instance of the SWRL built-in types `TimeInterval`.

Complex events can be defined by means of SWRL rules that evaluate occurrences of simple events, subjects and scenes and their temporal relations. The complex concepts and events that have been defined are the following:

- **anchorman:** If the same instance of a **Person** appears for a certain number of times in a delimited sequence of a news broadcast AND the occurrences of instances have temporal distance greater than a given threshold AND the same instance of **Studio Setting** occurs at the same time THEN that instance of **Person** is also an **Anchorman**;

- **simple interview:** If an instance of **Person** and an instance of **Anchor** appear for a certain number of times in a delimited sequence of a news broadcast AND the two instances alternate each other AND the same instances of **Studio Setting** occurs respectively during **Anchor** and **Anchor** THEN the instance of **Person** is an **Interviewee** AND during the news sequence an **Interview** is occurring;

- **complex interview:** If an instance of **Anchor** and an instance of **Person** alternate between each other (other events can occur in between such as shots of both anchors and the interviewed) with at least 3 occurrences of **Anchor** and 2 of **Person** AND the duration between the first and the last appearance of the anchor is less than 180 seconds THEN **Person** is an **Interviewed**



(a) Anchor (b) Interviewee (c) Two anchors (d) CGI shot (e) Anchor

(f) Anchor (g) Interviewee (h) Two anchors (i) CGI shot (j) Anchor

Figure 3. Keyframes of similar shots detected using MPEG-7 features and Needleman-Wunch distance

AND during the news sequence an **Interview** is occurring.

- **external report:** If two different instances of **Person** appear for a certain number of times in a delimited sequence of a news broadcast AND the two instances alternate between each other AND the same two different instances of **Studio Setting** and **external reporter setting** alternate accordingly THEN during the news sequence an **External Report** is occurring.

- **complex external report:** If at least two occurrences of **Report** and two occurrences of the same instance of **Person** occur in between of two occurrences of the same **Anchor** instance AND, regardless of their order AND the time interval between the two instances of **Anchor** is between 4 and 6 minutes THEN **Person** is **External Reporter** and during the news sequence an **External Report** is occurring.

In the following we present in details the SWRL translations of the above complex events. Please note that for the sake of clarity the number of occurrences required for a classification of person have been limited to 3.

- **anchorman:**

```
Person(?p1)  ^
hasValidPeriod(?p1, ?Vpp1)  ^
hasValidPeriod(?p1, ?Vpp2)  ^
hasValidPeriod(?p1, ?Vpp3)  ^
differentFrom(?Vpp1, ?Vpp2)  ^
differentFrom(?Vpp2, ?Vpp3)  ^
differentFrom(?Vpp1, ?Vpp3)  ^
temporal:hasFinishTime(?Vpp1, ?FTp1)  ^
temporal:hasStartTime(?Vpp2, ?STp2)  ^
temporal:hasFinishTime(?Vpp2, ?FTp2)  ^
temporal:hasStartTime(?Vpp3, ?STp3)  ^
temporal:duration(?dp1dp2,?FTp1, ?STp2, temporal:Seconds)  ^
temporal:duration(?dp2dp3,?FTp2,?STp3, temporal:Seconds)  ^
swrlb:greaterThan(?dp1dp2, 120)  ^
swrlb:greaterThan(?dp2dp3, 120)  ^
StudioSetting(?s1)  ^
hasValidPeriod(?s1, ?Vps1)  ^
hasValidPeriod(?s1, ?Vps2)  ^
hasValidPeriod(?s1, ?Vps3)  ^
differentFrom(?Vps1, ?Vps2)  ^
differentFrom(?Vps2, ?Vps3)  ^
differentFrom(?Vps1, ?Vps3)  ^
temporal:equals(?Vps1, ?Vpp1, temporal:Seconds)  ^
temporal:equals(?Vps2, ?Vpp2, temporal:Seconds)  ^
temporal:equals(?Vps3, ?Vpp3, temporal:Seconds)
->
? Anchor(?p1)
```

The *hasValidPeriod* property can be applied to any entity, object or event in a video and may have multiple values expressed as instances of a *ValidPeriod* class, according to the occurrences of the corresponding object or event. Each *ValidPeriod* has a *temporal:hasFinishTime* and *temporal:hasStartTime* property. Each *ValidPeriod* and temporal instance can be compared using the *temporal:duration*, *temporal:equals*, *temporal:before*, *temporal:meets* primitives. In this case *p1* is an instance of **Person** and *Vpp1*, *Vpp2*, *Vpp3* are three different time intervals corresponding to the occurrences of *p1*. The *temporal:greaterThan* conditions applied to *temporal:duration* of the time intervals require that the occurrences of *p1* have an interleaving greater than 120 seconds. *s1* is an instance of **StudioSettings** and *Vps1*, *Vps2*, *Vps3* are three different time intervals corresponding to the occurrences of *s1*. The rule requires that *p1* and *s1* have to occur at the same time. If all the above conditions are met then *p1* is classified as **Anchor**. The *differentFrom* constraint imposes that two instances of the same class have to be different. See Fig.4.

- **interview:**

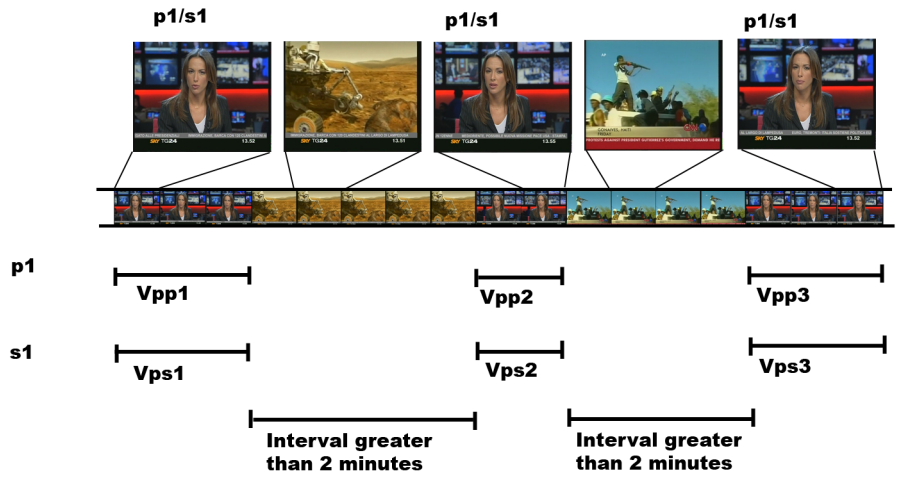


Figure 4. SWRL anchor rule schema

```
Anchor(?a1)  ^
Person(?p1)  ^
hasValidPeriod(?a1, ?Vpa1)  ^
hasValidPeriod(?a1, ?Vpa2)  ^
hasValidPeriod(?a1, ?Vpa3)  ^
differentFrom(?Vpa1, ?Vpa2)  ^
differentFrom(?Vpa2, ?Vpa3)  ^
differentFrom(?Vpa1, ?Vpa3)  ^
temporal:before(?Vpa1, ?Vpa2) ^
temporal:before(?Vpa2, ?Vpa3) ^
differentFrom(?p1, ?a1)  ^
hasValidPeriod(?p1, ?Vpp1)  ^
hasValidPeriod(?p1, ?Vpp2)  ^
differentFrom(?Vpp1, ?Vpp2)  ^
temporal:before(?Vpp1, ?Vpp2) ^
temporal:meets(?Vpa1, ?Vpp1, temporal:Seconds)  ^
temporal:meets(?Vpp1, ?Vpa2, temporal:Seconds)  ^
temporal:meets(?Vpa2, ?Vpp2, temporal:Seconds)  ^
temporal:meets(?Vpp2, ?Vpa3, temporal:Seconds)
StudioSetting(?s1)  ^
hasValidPeriod(?s1, ?Vps1)  ^
hasValidPeriod(?s1, ?Vps2)  ^
hasValidPeriod(?s1, ?Vps3)  ^
differentFrom(?Vps1, ?Vps2)  ^
differentFrom(?Vps2, ?Vps3)  ^
differentFrom(?Vps1, ?Vps3)  ^
temporal:equals(?Vps1, ?Vpa1, temporal:Seconds)  ^
temporal:equals(?Vps2, ?Vpa2, temporal:Seconds)  ^
temporal:equals(?Vps3, ?Vpa3, temporal:Seconds)  ^
StudioSetting(?z1)  ^
hasValidPeriod(?z1, ?Vpz1)  ^
hasValidPeriod(?z1, ?Vpz2)  ^
hasValidPeriod(?z1, ?Vpz3)  ^
differentFrom(?Vpz1, ?Vpz2)  ^
differentFrom(?Vpz2, ?Vpz3)  ^
differentFrom(?Vpz1, ?Vpz3)  ^
temporal:equals(?Vpz1, ?Vpp1, temporal:Seconds)  ^
temporal:equals(?Vpz2, ?Vpp2, temporal:Seconds)  ^
temporal:equals(?Vpz3, ?Vpp3, temporal:Seconds)  ^
-> Interviewee(?p1)
```

Where *a1* and *p1* are two different instances of **Anchor** and **Person**, respectively; *Vpa1*, *Vpa2*, *Vpa3* are three different time intervals corresponding to the occurrences of *a1* and *Vpp1*, *Vpp2*, *Vpp3* are three different time intervals corresponding to the occurrences of *p1*. The *temporal:before* conditions, together with the *temporal:meets* conditions, applied to time intervals of both *a1* and *p1* state that the occurrences of *a1* has to be followed, without any other content in between, by the occurrences of *p1*. It is also required that n instance *s1* of **StudioSetting** occurs during *a1*.

If all the above conditions are met then *p1* is classified as **Interviewee**. Then programmatically an event **Interview** is added to the ontology with *temporal:hasStartTime* set to the start time of *Vpa1* and *temporal:hasFinishTime*set to the finish time of *Vpa3*. See Fig.5.

- **complex interview:**

```
Anchor(?a1)  ^
Person(?p1)  ^
hasValidPeriod(?a1, ?Vpa1)  ^
hasValidPeriod(?a1, ?Vpa2)  ^
hasValidPeriod(?a1, ?Vpa3)  ^
differentFrom(?Vpa1, ?Vpa2)  ^
differentFrom(?Vpa2, ?Vpa3)  ^
differentFrom(?Vpa1, ?Vpa3)  ^
hasValidPeriod(?p1, ?Vpp1)  ^
hasValidPeriod(?p1, ?Vpp2)  ^
differentFrom(?Vpp1, ?Vpp2)  ^
temporal:before(?Vpa1, ?Vpp1)  ^
temporal:before(?Vpp1, ?Vpa2)  ^
temporal:before(?Vpa2, ?Vpp2)  ^
temporal:before(?Vpp2, ?Vpa3)  ^
temporal:hasStartTime(?Vpa3, ?STa3)  ^
temporal:hasFinishTime(?Vpa1, ?FTa1)  ^
temporal:duration(?diff, ?FTa1, ?STa3, temporal:Seconds)  ^
swrlb:lessThan(?diff, 180)
-> Interviewee(?p1)
```

Where *a1* is an instance of **Anchor** and *p1* is an instance of **Person**; *Vpa1*, *Vpa2*, *Vpa3* are three different time intervals corresponding to the occurrences of *a1* and *Vpp1*, *Vpp2*, are two different time intervals corresponding to the occurrences of *p1*. The *temporal:before* constrains state that *a1* and *p1* has to alternate each other (with at least three

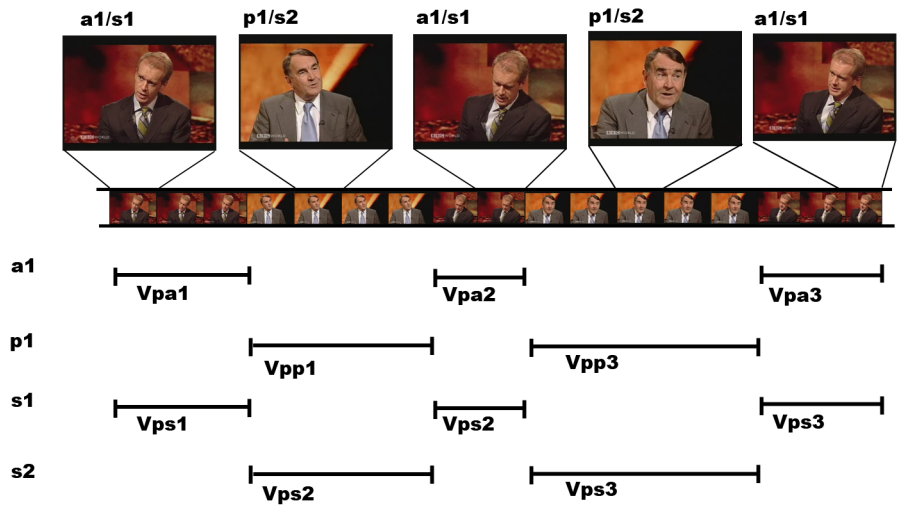


Figure 5. SWRL interview rule schema

occurrences for *a1* and two occurrences *p1*) and it is required that the *temporal:duration* of the interval between the first and last occurrences of *a1* is less than 180 seconds (*swrlb:lessThan*). If all the above conditions are met then *p1* is classified as **interviewee**. See Fig.6.

- **external report:**

```
Person(?p1)  ^
Person(?q1)  ^
hasValidPeriod(?p1, ?Vpp1)  ^
hasValidPeriod(?p1, ?Vpp2)  ^
hasValidPeriod(?p1, ?Vpp3)  ^
differentFrom(?Vpp1, ?Vpp2)  ^
differentFrom(?Vpp2, ?Vpp3)  ^
differentFrom(?Vpp1, ?Vpp3)  ^
temporal:before(?Vpp1, ?Vpp2) ^
temporal:before(?Vpp2, ?Vpp3) ^
differentFrom(?q1, ?p1)  ^
hasValidPeriod(?q1, ?Vpq1)  ^
hasValidPeriod(?q1, ?Vpq2)  ^
differentFrom(?Vpq1, ?Vpq2)  ^
temporal:before(?Vpq1, ?Vpq2) ^
temporal:meets(?Vpp1, ?Vpq1, temporal:Seconds)  ^
temporal:meets(?Vpq1, ?Vpp2, temporal:Seconds)  ^
temporal:meets(?Vpp2, ?Vpq2, temporal:Seconds)  ^
temporal:meets(?Vpq2, ?Vpp3, temporal:Seconds)  ^
OutdoorSetting(?o1)  ^
hasValidPeriod(?o1, ?Vpo1)  ^
hasValidPeriod(?o1, ?Vpo2)  ^
temporal:equals(?Vpo1, ?Vpq1)  ^
temporal:equals(?Vpo2, ?Vpq2) ->
Anchor(?p1)  ?
ExternalReporter(?q1)
```

In this example all the statements are the same than the ones defined for the **Interview** event. The only difference is in that is required an instance of **OutdoorSetting** occurring during *p1*.

- **complex external report:**

```
Anchor(?a1)  ^
Person(?p1)  ^
hasValidPeriod(?p1, ?Vpp1)  ^
hasValidPeriod(?p1, ?Vpp2)  ^
differentFrom(?Vpp1, ?Vpp2)  ^
Report(?r1)  ^
hasValidPeriod(?r1, ?Vpr1)  ^
Report(?r2)  ^
hasValidPeriod(?r2, ?Vpr2)  ^
differentFrom(?r1, ?r2)  ^
hasValidPeriod(?a1, ?Vpa1)  ^
hasValidPeriod(?a1, ?Vpa2)  ^
differentFrom(?Vpa1, ?Vpa2)  ^
temporal:before(?Vpa1, ?Vpa2)  ^
temporal:hasFinishTime(?Vpa1, ?FTa1)  ^
temporal:hasStartTime(?Vpa2, ?STa2)  ^
temporal:before(?Vpr1, ?STa2, temporal:Seconds)  ^
temporal:before(?Vpr2, ?STa2, temporal:Seconds)  ^
temporal:before(?Vpp1, ?STa2, temporal:Seconds)  ^
temporal:before(?Vpp2, ?STa2, temporal:Seconds)  ^
temporal:before(?FTa1, ?Vpr1, temporal:Seconds)  ^
temporal:before(?FTa1, ?Vpr2, temporal:Seconds)  ^
temporal:before(?FTa1, ?Vpp1, temporal:Seconds)  ^
temporal:before(?FTa1, ?Vpp2, temporal:Seconds)  ^
temporal:duration(?dur, ?FTa1, ?STa2, temporal:Seconds)  ^
swrlb:greaterThan(?dur, 240) ^
swrlb:lessThan(?dur, 360)
-> ExternalReporter(?p1)
```

Where *a1* is an instance of **Anchor**, *p1* is an instance of **Person** *r1* and *r2* are instances of **Report**, *Vpa1* and *Vpa2* are two time intervals corresponding to the occurrences of *a1*, *Vpp1* and *Vpp2* are two time intervals corresponding to the occurrences of *p1*. The *temporal:before* constraints state that occurrences of **Report** and **Person** have to occur in an interval between the two occurrences of **Anchor**. The *swrlb:greaterThan* and *swrlb:lessThan* constraints set the duration of that interval between 4 and 6 minutes. If all the above conditions are met then *p1* is an **External Reporter**. See Fig.7.

5 Experimental Results

The proposed approach has been tested on several news videos from different countries and broadcasters (BBC

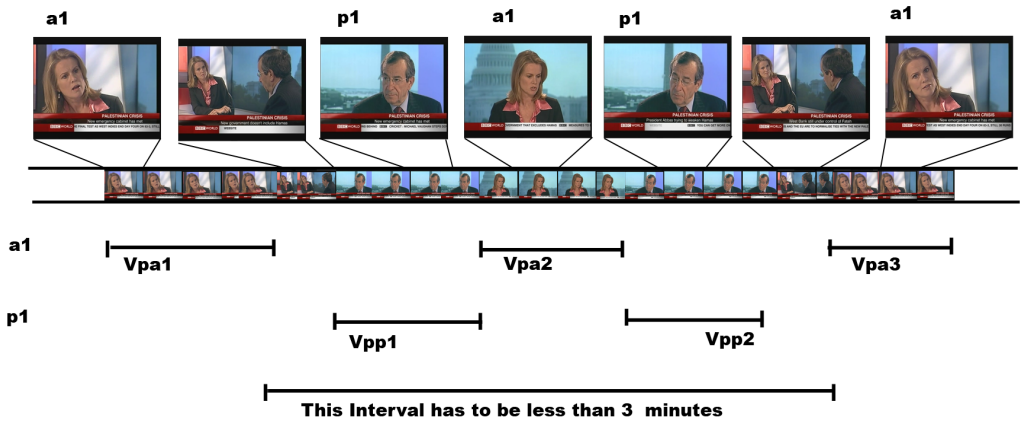


Figure 6. SWRL complex interview rule schema

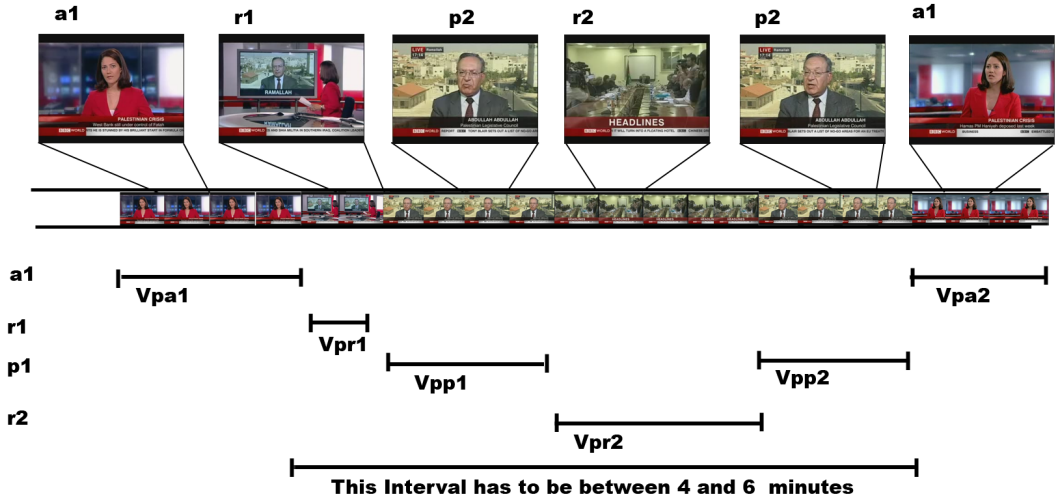


Figure 7. SWRL complex external report rule schema

World, CNN, RAI, Mediaset Canale 5, Sky TV). The videos have been acquired at full PAL frame rate and frame resolution and compressed using MPEG-2, for a total length of more than 2.5 hours of videos. The shots obtained using a simple video segmentation based on color histograms have been processed (resized to 360×288 pixels) to extract the MPEG-7 color and texture features and similar scene settings have been identified using the approach described in Sect. 3. Face detection has been performed on the whole videos at full PAL resolution. In order to avoid errors due to false detections shots have been considered as containing faces only if they were detected in consecutive frames for more than one third of the frames of a shot.

Table 1 reports results of the detection of complex events using multimedia ontology and the SWRL rules described in the previous sections. The good figures of anchormen shots detection are due to the good behaviour of the combination of the visual features used: face detection and shot similarity based on MPEG-7 descriptors can accurately model the pattern of the anchormen shots. The only false detection is due to an unfortunate case in which faces were shown during the titles of a news video, while the title background matched the layout of the studio setting. The low number of missed detections are due to a few extremely

Complex event	Occurrences	Detected	Miss	False
Anchormen shots	105	95	10	1
External report	6	6	0	0
Interview	20	20	0	0

Table 1. Complex events recognition

short anchormen shots that did not allow to asses a robust similarity to the other shots within the same video.

The well defined patterns of events such as external reporters and interviews combined with the effective modeling that can be achieved with SWRL are responsible of the good figures shown in the table. We expect that for these types of events "misses" are more likely to happen instead of "false detections" since the intrinsic nature of their pattern would require a large number of false detections of simple events in a short time interval.

However, we have to report that in some news videos six shots that actually were external reporters were not detected since they did not match the model: these are isolated shots that usually appear in the middle of a news report, in which a journalist, that does not re-appear anymore during the news video, performs a short monologue. In order to recognize this kind of external reporters we should add higher-level semantic knowledge that may allow to state that the framed person is a journalist.

6 Conclusions

Temporal specification of high-level, composite events is of critical importance, since many of the interesting events that occur in most video domains cannot be described in terms of individual, atomic events that can be characterized using only simple visual descriptors or static concepts. In this paper we have described our initial work on multimedia ontologies extended with temporal rules inference, that uses the Allen interval algebra, to model complex dynamic concepts.

In our future work we will deal with extension of the test set, using a standard corpus such as the TrecVid set to test the performance and the behaviors of the framework.

Acknowledgments

This work is partially supported by the Information Society Technologies (IST) Program of the European Commission as part of the VIDI-Video project (Contract FP6-045547).

References

[1] M. Bertini, A. Del Bimbo, and W. Nunziati. Video clip matching using MPEG-7 descriptors and edit distance. In *In Proc. of Conference on Image and Video Retrieval (CIVR)*, August 2006.

[2] 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.

[3] A. Eickeler and S. Muller. Content-based video indexing of tv broadcast news using hidden markov models. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2997–3000, March 1999.

[4] A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, 12(7):796–807, July 2003.

[5] 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.

[6] T. Gruber. Principles for the design of ontologies used for knowledge sharing. *Int. Journal of Human-Computer Studies*, 43(5-6):907–928, 1995.

[7] A. Hauptmann and M. Witbrock. Informedia: News–on–demand multimedia information acquisition and retrieval. *Intelligent Multimedia Information Retrieval*, pages 213–239, 1997.

[8] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean. Swrl: A semantic web rule language combining owl and ruleml. Technical report, W3C Member submission 21 may 2004, 2004.

[9] A. Jaimes, B. Tseng, and J. Smith. Modal keywords, ontologies, and reasoning for video understanding. In *Int'l Conference on Image and Video Retrieval (CIVR)*, July 2003.

[10] L. Kennedy. Revision of lscom event/activity annotations, dto challenge workshop on large scale concept ontology for multimedia. Technical Report 221-2006-7, Columbia University ADVENT Technical Report, December 2006.

[11] B. Qasemizadeh, H. Haghi, and M. Kangavari. A framework for temporal content modeling of video data using an ontological infrastructure. In *In Proc. of Semantics, Knowledge and Grid, 2006.*, November 2006.

[12] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith. Video ocr for digital news archive. *IEEE International Workshop on Content–Based Access of Image and Video Databases CAIVD' 98*, pages 52–60, 1998.

[13] C. 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), August 2007.

[14] J. Strintzis, S. Bloehdorn, S. Handschuh, S. Staab, N. Simou, V. Tzouvaras, K. Petridis, I. Kompatsiaris, and Y. Avrithis. Knowledge representation for semantic multimedia content analysis and reasoning. In *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, Nov. 2004.

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

[16] X.Yu, C. Xu, H. Leung, Q. Tian, Q. Tang, and K. W. Wan. Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video. In *ACM Multimedia 2003*, volume 3, pages 11–20, Berkeley, CA (USA), 4-6 Nov. 2003.