Ontology-based Video Annotation in Multimedia Entertainment

Antonella Carbonaro, Rodolfo Ferrini
Department of Computer Science
University of Bologna
Mura Anteo Zamboni 7, I-40127
Bologna, Italy
e_mail: {carbonar, ferrini}@csr.unibo.it

Abstract. In this paper we propose ontology-based video content annotation and recommendation tools. Our system is able to perform automatic shot detection and supports users during the annotation phase in a collaborative framework by providing suggestions on the basis of actual user needs as well as modifiable user behaviour and interests. Annotations are based on domain ontologies expressing hierarchical links between entities and guarantying interoperability of resources. Examples to verify the effectiveness of both the shot detection and the frame matching modules are analyzed.

1. INTRODUCTION

Vast amounts of multimedia information including video are becoming ubiquitous as a result of advances in multimedia computing technologies and high-speed networks. Video is rapidly becoming the most popular media, due to its high information and entertainment power.

The main challenge is to index information retained in video in order to make them searchable and thus (re-) usable. This requires the multimedia content to be annotated, which can either be done manually or automatically. In the first case, the process is extremely work- and thus cost-intensive; in the second case, it is necessary to apply content-analysis algorithms that automatically extract descriptions from the multimedia data. The aim is to create a concise description of the multimedia content features, that is, its metadata. Metadata descriptions may vary considerably in terms of comprehensiveness, granularity, abstraction level, etc. depending on the application domain, the tools used and the effort made for creating the descriptions.

Multimedia annotation systems need standard output, which must be compliant with other tools for browsing or indexing. MPEG-7 [9] standard was defined for this purpose. It represents an elaborate standard in which a number of fields, ranging from low level encoding scheme descriptors to high level content descriptors, are merged to be useful for describing a video or part of it.

In text-based applications, it is often sufficient to annotate only the generic properties of the document (such as title and creator) and to perform keyword search, based on full-text information retrieval approaches. For non-textual resources, however, full text search is only an option if there are sufficient associated textual information (for example, Google's image search based on surrounding text in HTML or video retrieval based on subtitles, closed captions or audio transcripts). In many other cases content descriptions are inevitable. Since content descriptions are not often about the entire document (e.g. a specific shot in a film or a specific region in a picture), it is necessary to implement shot detection and keyframe extraction procedures to deal with video files.

An important step towards efficient manipulation and retrieval of visual media is semantic information representation [3], [1]. In the digital library community a flat list of attribute/value pairs is often assumed to be available. In the Semantic Web community, annotations are often assumed to be an instance of an ontology. Through the ontologies the system will express key entities and relationships describing video in a formal machine-processable representation. An ontology-based knowledge representation could be used for content analysis and object recognition, for reasoning processes and for enabling user-friendly and intelligent multimedia content search and retrieval.

Applications that could benefit from semantic video representation are manifold, from education and training to medical, from entertainment to system analysis and evaluation, etc. For example, home entertainment systems (management of personal multimedia collections, including manipulation of content, home video editing, searching, etc.) need a mechanism to interpret human's queries, and retrieve the closest match. However, this search outcome may result very unsatisfactory due to the blurred link between the low-level measured features and the human semantic queries. This discrepancy between the way video data is coded digitally and the way it is experienced by a human user is called the semantic gap, [10]. Differently, in education, semantic annotations of video recording of lectures distributed over the Internet can be used to augment the material by providing explanations, references or examples, that can be used for efficiently access, find and review material in a student personal manner [4], [5]. Moreover, in television, semantic annotation of programmes, for example news, could produce electronic programme guides, which would allow the user to view details of forthcoming programmes in terms of entities referred to in particular broadcasts [6].

1-4244-0667-6/07/$25.00 © 2007 IEEE

In this paper, we propose how to introduce a semantic-based representation of the information embedded in the video media both by user interaction and by ontology exploiting. Aim of this paper is to introduce *Scout-V* (Semantic-based COntent management Tools - Video), our semantic video content annotation system in multimedia entertainment. *Scout-V* provides following web-based tools to help consumers to manage their video collection:

- a browsing application to enable users efficiently access their video collections,
- a shot detection application to automatically identify shot in the video measuring the similarity among frames,
- an annotation application to enable users easily annotate video sequences,
- an ontology editor application to enable users to modify the ontology tree creating all the necessary classes and instances,
- a video annotation recommendation application enable user to speed up annotation task proposing similar frames which have been annotated also by other users,
- an MPEG-7 file producer application to maintain the obtained video content description.

The paper is organized as follows. Section 2 proposes a description of developed system and of web tools. Section 3 details shot detection techniques implemented within our architecture. Section 4 illustrates experimental sessions. Some final considerations and comments about future developments conclude the paper.

2. SYSTEM DESCRIPTION

The *Scout-V* module assists authors in annotating video sequences. Each shot belonging to the video sequence can be annotated on the base of underlying ontologies. These descriptions are labelled for each shot and are stored as MPEG-7 descriptions in the output XML file. *Scout-V* can also save, open, and retrieve MPEG-7 files in order to display the annotations for corresponding video sequences.

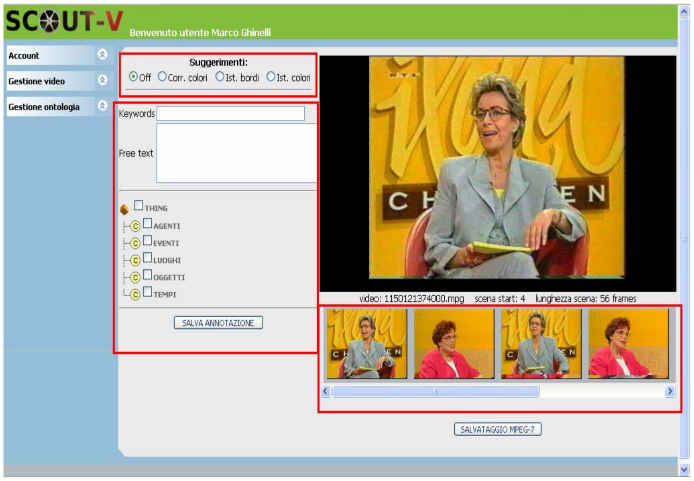
The *Scout-V* main page shows all the videos that should be elaborated performing shot detection, editing or removing. Given the segmentation of video content into video shots, the second step is to define the semantic lexicon to label the shots. A video shot can fundamentally be described by using five basic classes: agents, objects, places, times and events. These five types of lexicon define the initial vocabulary for our video content; they correspond to the *SemanticBase* MPEG-7 tags.

```
1)<SemanticBase xsi:type="AgentObjectType">
2)<SemanticBase xsi:type="ObjectType">
3)<SemanticBase xsi:type="SemanticPlaceType>
4)<SemanticBase xsi:type="SemanticTimeType">
5)<SemanticBase xsi:type="EventType">
```

We have also defined attributes to describe class characteristics. Each attribute corresponds to a specified MPEG-7 tag used in storing phase.

By using the defined vocabulary for static agents, key objects, places, times and events, the lexicon is imported into *Scout-V* for describing and labelling each video shot. The shots are labelled for their content with respect to the selected lexicon. Note that the lexicon definitions are database and application specific, and can be easily modified and imported into the annotation tool.

Scout-V annotation tool is divided into three graphical sections, as illustrated in Figure 1. On the upper left-hand corner is depicted the *Scene Matching* frame in which are shown the algorithms that can be used to obtain video annotation recommendations (Block Truncation Coding, edge histogram, colour histogram). On the bottom left-hand portion of Figure 2 is placed the *Ontology Visualization* frame, providing interactivity to assist authors of the annotation tool. On the right, there is the *Video Presentation* frame with the key frame image display and the frame characteristics.



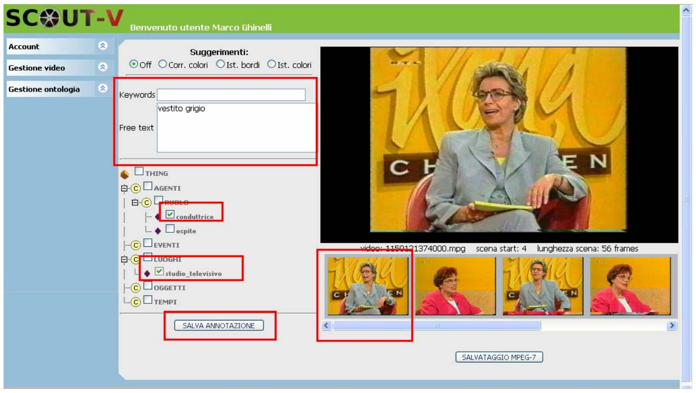
Suggerimenti: ◉ Off ○ Corr. colori ○ Ist. bordi ○ Ist. colori

video: 1150121374000.mpg scena start: 4 lunghezza scena: 56 frames

Fig. 1 Scout-V annotation tool.

The *Ontology Editor* module allows to modify the ontology tree creating and populating all the necessary classes and instances. The aim of the instance creation phase is to effectively represent the domain knowledge, so as to achieve a better precision in the annotation task.

Figure 2 shows annotation procedure applied on the first scene by using checkboxes and comments to better describe the selected video. Annotations are then stored and used by recommendation procedure to help users finding similar frames which have been annotated also by other users (see Figure 3).



● Off ○ Corr. colori ○ Ist. bordi ○ Ist. colori

Keywords

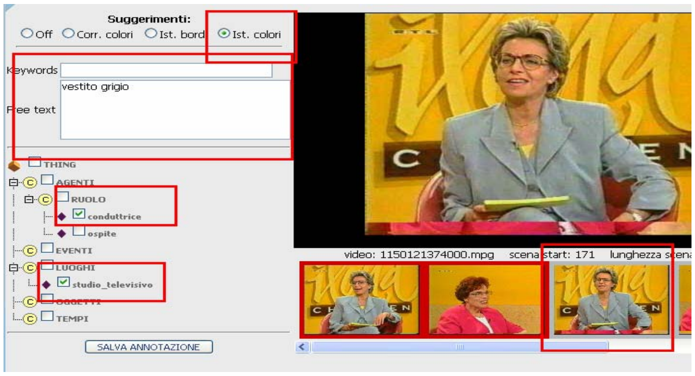
Free text: vestito grigio

□ THING
□ AGENTI
□ RUOLO
☑ conduttrice
□ ospite
□ EVENTI
□ LUOGHI
☑ studio_televisivo
□ OGGETTI
□ TEMPI

video: 1150121374000.mpg scena start: 4 lunghezza scena: 56 frames

Fig. 2 First frame annotation phase

Once the scene as a whole has been annotated, the system produces a MPEG-7 file. Figure 4 shows content description of the first scene as stored in the MPEG-7 standard.



○ Off ○ Corr. colori ○ Ist. bord ● Ist. colori

Keywords

Free text: vestito grigio

□ THING
□ AGENTI
□ RUOLO
☑ conduttrice
□ ospite
□ EVENTI
□ LUOGHI
☑ studio_televisivo
□ OGGETTI
□ TEMPI

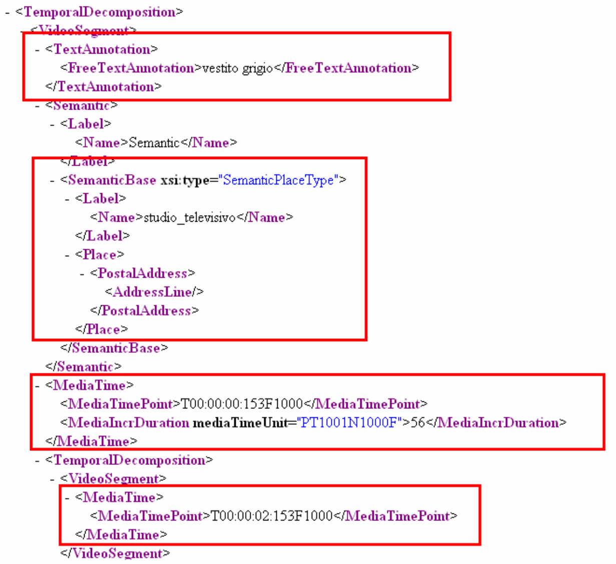
video: 1150121374000.mpg scena start: 171 lunghezza scena

Fig. 3 Recommendation based on edge histogram technique

3. SHOT DETECTION TECHNIQUES

By the shot detection method the video can be automatically segmented into shots. A shot is a contiguous sequence of video frames which have been recorded from a single camera operation [8]. The method is based on the detection of shot transitions (hart cuts, dissolves, and fades). One or more keyframes are extracted from the obtained shots set in dependence on the visual content dynamics.

The basic idea is to measure the similarity among frames. The most classical algorithm for such a task is to compare one by one each pixel in the image; this method is relatively easy to implement but it suffers form problems related to both camera movement and zoom, as well as from the noise produced, in case, by a lossy compression. For this reason in literature several methods are proposed [2] [7], which take into consideration more general aspects of a picture like the frequency of a colour or of its exact location. Below we have listed three approaches we have implemented as a functionality of our architecture:



```
- <TemporalDecomposition>
  - <VideoSegment>
    - <TextAnnotation>
        <FreeTextAnnotation>vestito grigio</FreeTextAnnotation>
      </TextAnnotation>
    - <Semantic>
      - <Label>
          <Name>Semantic</Name>
        </Label>
      - <SemanticBase xsi:type="SemanticPlaceType">
        - <Label>
            <Name>studio_televisivo</Name>
          </Label>
        - <Place>
          - <PostalAddress>
              <AddressLine/>
            </PostalAddress>
          </Place>
        </SemanticBase>
      </Semantic>
    - <MediaTime>
        <MediaTimePoint>T00:00:00:153F1000</MediaTimePoint>
        <MediaIncrDuration mediaTimeUnit="PT1001N1000F">56</MediaIncrDuration>
      </MediaTime>
    - <TemporalDecomposition>
      - <VideoSegment>
        - <MediaTime>
            <MediaTimePoint>T00:00:02:153F1000</MediaTimePoint>
          </MediaTime>
        </VideoSegment>
```

Fig. 4 Scene content description

3.1 Histogram-based method.

The most common approach used in the field of automatic shot detection uses colour histograms. Histogram of an image *f* is a function $p_f(r)$ which specifies the frequency of a colour *r* on the image *f*. For a digital image, since colour is a quantized variable, $p_f(r)$ is a discrete function, only defined for a finite number *M* of colour levels. For this reason, an histogram takes into consideration only the frequency of the pixel instead oft its position.

The algorithm implemented in *Scout-V* uses a particular histogram called RGB in which the picture is divided in the three primary colour i.e. red, green and blue. Moreover, our approach makes the assumption that if two frame histograms belong to the same shot, they are closer than two histograms belonging to different shots.

3.2 Edge-based method.

The colour frequency measurement is not the only feature that one could take into consideration for the automatic shots detections. Another common technique is based on edges. This method suffers from the camera movement but it could intercept the fading between two shots with more accuracy, one of the most difficult tasks in shot detection. In this case, the value we want to obtain is an edges variation factor. Once the value is achieved, it will be normalized and processed in order to obtain the final similarity measurement between frames to detect shots.

1-4244-0667-6/07/$25.00 © 2007 IEEE

3.3 Fourier transform-based method.

The Fourier transform is widely used in signal processing, but could be efficiently used also in the image processing field. In *Scout-V* we use a template-matching method based on the Fourier transform which utilize the correlation function. Mathematical details of the process are not the purpose of this work, so we list only the most important steps of our approach:

- Fourier transform measurement of an image I;
- Fourier transform measurement of the template T;
- Execution of specific formula;
- Result analysis.

In *Scout-V* we implemented a hybrid approach which takes into consideration the proposed algorithms to achieve better results.

4. EXPERIMENTS

We have considered several MPEG videos available online at http://www.open-video.org in order to verify the effectiveness of both the shot detection module and the frame matching module on the annotation process. The first file we have considered shows a boat and a fisherman, alternating the pictures (Figure 5).

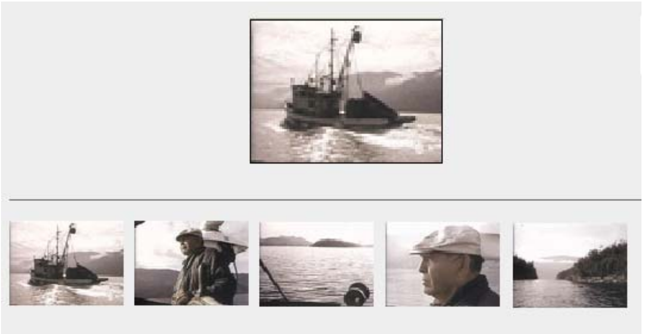


Fig. 5 Scenes from the first video, as showed at http://www.open-video.org

The same videos have been processed using Scout-V, obtaining the same results, as depicted in Figure 6. Once again, such results have been obtained using the majority of input videos available online. Some differences have been noticed on out-of-focus images. We are already working to strengthen and improve shot detection module for such cases. Obtained precision and recall average values are 88.87% and 92.54%.

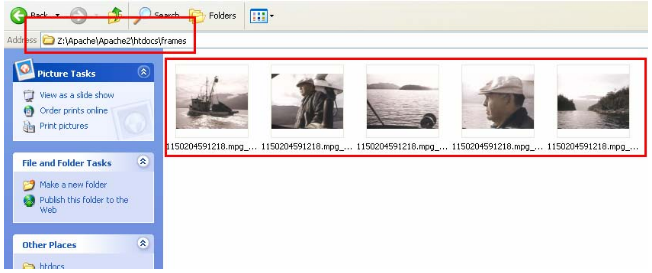


Fig. 6 Scenes obtained using *Scout-V* on the first video file

We have considered several videos in order to verify the effectiveness of the frame matching module to recommend scene on the base of its similarity. For example, Figure 7 shows the different scenes composing a tennis video; let us notice the two players, the referee and the audience.

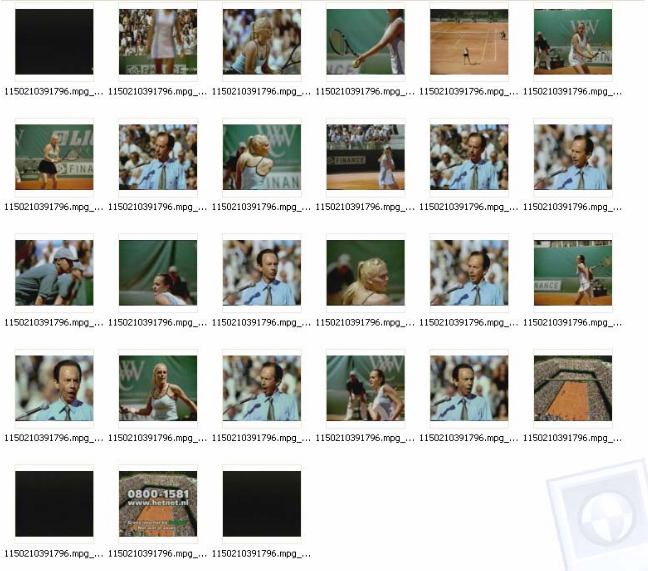
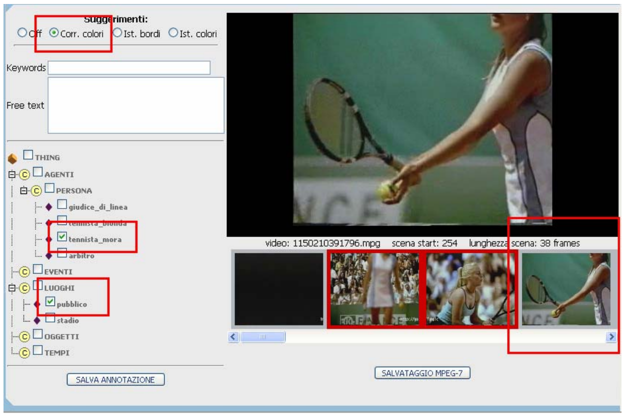

1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_...
1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_...
1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_...
1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_...
0800-1581
www.hetnet.nl
1150210391796.mpg_... 1150210391796.mpg_... 1150210391796.mpg_...

Fig. 7 Tennis video scenes

Figure 8 shows annotation automatically suggested by *Scout-V* for the fourth frame; the system has recognized the similarity with the third scene.

5. CONCLUSIONS

In this paper we have presented a methodology for semantic video content annotation. The system comprises automatic shot detection and scene matching modules to obtain video annotation recommendations in a collaborative framework. Several experiments tested the effectiveness of both the shot detection module and the frame matching module on the annotation process. Future works will include the study and the implementation of an ontology layer able to maintain several existing ontologies the user knows. This approach could allow to compare the knowledge of any user without having a single consensual ontology.



Suggerimenti: ○ Off ● Corr. colori ○ Ist. bordi ○ Ist. colori

Keywords

Free text

- □ THING
  - □ AGENTI
    - □ PERSONA
      - □ giudice_di_linea
      - □ tennista_bionda
      - ☑ tennista_mora
      - □ arbitro
  - □ EVENTI
  - □ LUOGHI
    - ☑ pubblico
    - □ stadio
  - □ OGGETTI
  - □ TEMPI

video: 1150210391796.mpg scena start: 254 lunghezza scena: 38 frames

Fig. 8 Annotation suggested by *Scout-V* for the fourth frame

6. REFERENCES

[1] Bloehdorn, S., K Pet.ridis, N Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab, M. G. Strintzis (2004) Knowledge Representation for Semantic Multimedia Content Analysis and Reasoning. In Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology.

[2] Boreczky, J. S., L. A. Rowe (1996) Comparison of video shot boundary detection techniques J. Electronic Imaging 05(02), 122-128, Edward R. Dougherty; Ed.

[3] Calic, J., N. Campbell, S. Dasiopoulou and Y. Kompatsiaris (2005). A Survey on Multimodal Video Representation for Semantic Retrieval. In: the Third International Conference on Computer as a tool (Eurocon 2005), IEEE.

[4] Carbonaro, A. (2005). Defining personalized learning views of relevant learning objects in a collaborative bookmark management system. Web-Based Intelligent e-Learning Systems: Technologies and Applications" Idea Group Inc.

[5] Carbonaro, A. and R. Ferrini (2005) Considering semantic abilities to improve a Web-Based Distance Learning System, ACM International Workshop on Combining Intelligent and Adaptive Hypermedia Methods/Techniques in Web-based Education Systems.

[6] Dowman, M., V Tablan, H Cunningham, B Popov (2005) Web-assisted annotation, semantic indexing and search of television and radio news Proceedings of the 14th international conference on World Wide Web, pp. 225 – 234.

[7] Gargi, U., R. Kasturi, S.H. Strayer (2000) Performance characterization of video-shot-change detection methods, Circuits and Systems for Video Technology, IEEE Transactions on , vol.10, no.1pp.1-13.

[8] Grana, C., G. Tardini, R. Cucchiara, "MPEG-7 Compliant Shot Detection in Sport Videos" in Proceedings of IEEE International Symposium on Multimedia (ISM2005), Irvine, California, USA, pp. 395-402, December 12-14, 2005.

[9] ISO/IEC. Overview of the MPEG-7 Standard (version 8). ISO/IEC JTC1/SC29/WG11/N4980, Klagenfurt, July 2002.

[10] Smeulders, A. W. M., M. Worring, S. Santini, A. Gupta and R. Jain (2000), Content based image retrievals at the end of the early years. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, issue 12, pp. 1349-1380.