Automatic Live Tagging of Videos Using Chronicles*

Marta Rey-López mrey@det.uvigo.es
Ana Fernández-Vilas avilas@det.uvigo.es
Rebeca P. Díaz-Redondo rebeca@det.uvigo.es
José J. Pazos-Arias jose@det.uvigo.es

Department of Telematic Engineering
University of Vigo, 36310, Spain

ABSTRACT

Creating and editing videos is a costly process in terms of time and money. For this reason, reusability is a major issue in this field; hence, videos should be appropriately labelled, describing its structure and content, but this process is costly too. However, for some types of videos, this information is already expressed as structured text. In this paper, we present an automatic tagging system to describe the structure and semantics of the videos for sports events, extracting information from the chronicles for these events written in the web pages of some journals or sports organisations.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

General Terms

Automatic Tagging, Information Extraction, Metadata, Natural Language Processing

1. INTRODUCTION

The widespread of new information technologies permits broadcasters to record and store a huge amount of video content. In the same manner, and supported by the possibilities of Interactive Digital TV (IDTV), TV viewers have access to plenty of contents. For these reasons, it is important to provide some mechanisms to look for and retrieve the created videos, with the aim of reusing them and offering the interesting ones to the viewers. In order to be able to create systems that perform these tasks, a good description of the videos should be provided.

It is commonly supposed that content creators are the ones in charge of writing these descriptions, as in YouTube (`http://www.youtube.com`) where subscribers use keywords to tag the videos they have uploaded. Nevertheless, this hypothesis has several disadvantages. On the one hand, the subscribers might not label the videos they upload, or they might label them according to their point of view (labelling is a subjective matter). On the other hand, labelling amateur short videos by means of keywords does not require an important amount of time from their creators. However, when dealing with long professional videos which should be described providing semantical information, the process become highly time-consuming.

In order to compensate the lack of descriptions provided by content creators, a new technology has aroused on the web: collaborative tagging. It is a process where many users add metadata to shared content in the form of keywords, so that not only they can categorise information for themselves, but they can also browse the information categorised by others [7]. This technology has been successfully used in pages like Delicious (`http://del.icio.us`) —where collaborative tagging is used to label web pages— or Google Video (`http://video.google.es`) —where videos are indexed using the tags provided by different users. In the field of TV, this technology could be applied so that different users had the possibility of adding tags to the TV programs they are watching. The most popular tags could be added to the metadata that describes this video to be sent to other viewers. However, although this technology is very useful for informal applications, trusting in the users' criterion to describe the videos is not a valid option in commercial ones, where a reliable source of information is needed.

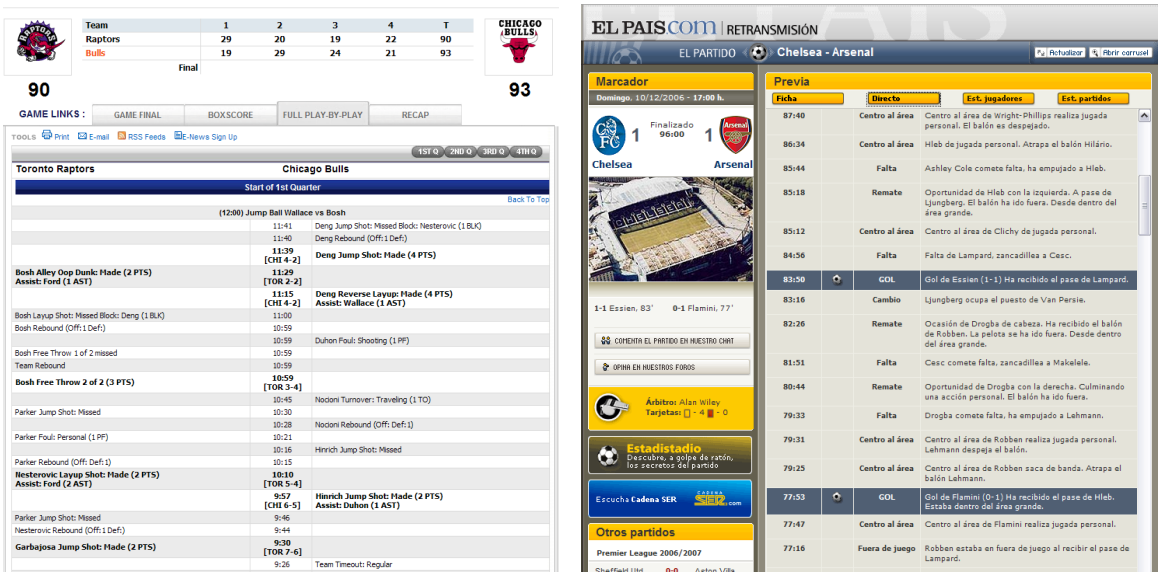
In this paper, we present a dependable application to automatically tag videos of sports events in an objective way —as opposed to the subjective character of users' tagging. This application uses a textual source, parallel to the video, to extract semantical information about the events that take place in relevant moments of the video. In our case, the input textual sources are the chronicles of sports events written by on-line journals and other web pages (Fig. 1) to automatically tag the segments of sports videos (understanding segments as continuous fragments of the video).

This application has been developed into a broader project in the field of Interactive Digital TV (IDTV) where the descriptions of the videos are used to establish relations with some pieces of other videos or additional TV contents, such as advertisements, learning material, etc. For example, if the

*Partly supported by the R+D project TSI 2007-65599 (Spanish Ministry of Education and Science).

AMDIT 2008, February 14, Quebec, Canada

DOI 10.4108/ICST.AMBISYS2008.2818



(a) NBA web page

(b) On-line journal ElPaís

Figure 1: Chronicles of sports events

viewer is watching a soccer match and Ronaldinho scores a goal, he/she could be offered some fragments of other videos where this player scores a goal too. In order for these relationships to be established, semantical descriptions should be provided for the pieces of videos. With the aim of providing these descriptions, the contribution of this paper is a complementary dependable method to the semantic descriptions provided by the content creators. In this application area, our approach has the advantage that, in live transmissions, the additional contents can be offered very soon after the event takes place, since chronicles are written almost while they happen.

The paper is structured as follows, first, we discuss some related approaches. Next, we explain how the developed application works, evaluate and discuss it. Last but not least, we draw some conclusions and motivate our future work.

2. RELATED WORK

In the field of automatically obtaining metadata that describes video content, two main approaches can be identified: on the one hand, there are some research efforts that directly extract the information from the video source: using motion vectors [8], detecting slow motion shots [5], using finite state machines [1] or analysing the ball trajectory [15].

On the other hand, closer to our approach, some other proposals use a related text source to obtain the information.

The closest one is the proposal exposed in [12] where ticker reports of soccer matches are used as one of the sources to extract the information about the match by means of templates. However, the process of parsing the information is more complicated and time-consuming than ours. These two factors are very important in the field of Interactive Digital TV (IDTV) —where our approach is going to be applied— due to the restrictions imposed by the Set-top boxes used in this medium.

Semantic techniques can be also used to extract the information from related text sources. The work presented in [6] uses contextual sources to obtain keywords for TV programs. Its method consists of three steps: lemmatization to identify which lemma a wordform is associated with, semantic tagging of context documents and ranking of the keywords, based on relations between them and the number of occurrences. The approach in [3] consists in using pictorially enriched ontologies, i.e. ontologies that include visual concepts together with linguistic keywords.

A hybrid approach for extracting salient or domain-specific keywords from instructional videos by exploiting joint audio, visual, and text cues is presented in [10]. It uses three steps: keywords extraction from video transcripts, analysis of audiovisual content to identify the domain and adjust of the keyword salience by fusing the audio, visual and text cues together.

3. AUTOMATIC TAGGING USING CHRONICLES

We have already introduced that the mechanism used by the automatic tagging application is based on using the chronicles written in on-line journals or sports web pages. It is now important to explain which one is going to be used. Almost all on-line journals and web pages of TV channels, as well as some additional web pages, like the NBA one, write chronicles about the most relevant sports events. However, two of them have been identified as particularly interesting (Fig. 1): the NBA web page (`http://www.nba.com/games/20061208/TORCHI/playbyplay.html`) and the on-line version of the Spanish diary El País (`http://www.elpais.com/deportes/futbol/partido.html?p=0213_00_17_0287_0235`). As mentioned before, for professional videos, a reliable and objective

EVENT	CHANGE	
IDENTIFIER	*Cambio*	
PROPERTIES	PATTERNS	
	BEFORE	AFTER
comesIn	—	*sustituye a*
	—	*entra por*
	—	*ocupa el puesto de*
goesOut	*sustituye a*	·
	entra por	·
	ocupa el puesto de	·

Table 1: Language patterns for the event "change"

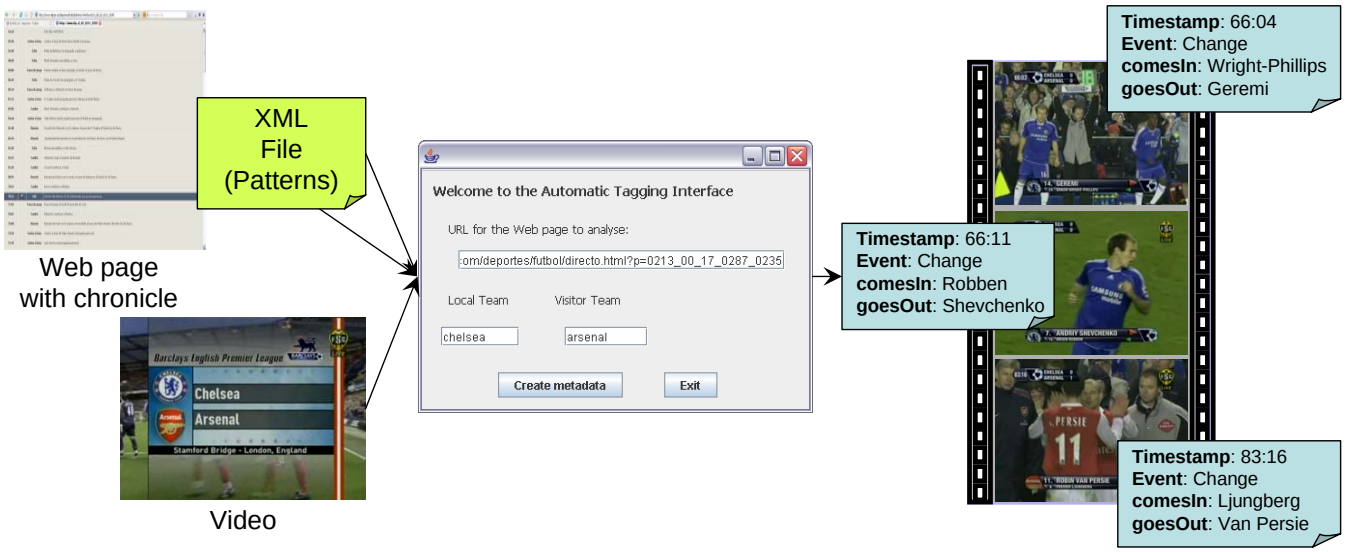


Figure 2: The Automatic Tagging Tool

source is needed. Both of the aforementioned pages fulfil this requirement. In addition, they use a very structured and non-ambiguous way of expressing the information. For the application developed, we have chosen the second one for one reason, although the application has been tested to obtain information from soccer chronicles, this web page offers broader future improvements, since it provides information about different sports and different leagues in the world.

3.1 Language Patterns

The main idea of this application is extracting information from the chronicles, written in natural language, to represent them in a structured, standardised way to be used later for automatic searches. In this manner, it can be considered a simple application of Natural Language Processing techniques —of the subfield of Information Extraction[1]—, that works correctly for the scenario and web pages exposed.

Text recognition for this application is based on an XML file —different for each sport— that defines the different types of events that can occur in a match of this sport, and the structures of the sentences that explain the properties of these events, called patterns. For soccer matches, eight different events have been identified (the properties for each event are shown in parenthesis): goal (*scoredBy* and *result*), foul/penalty (*by* and *to*), shot (*by* and *type*), change (*comesIn* and *goesOut*), centre (*by*), offside (*by* and *passedBy*), into the area (*by*) and card (*type* and *to*). Table 1 shows an example of the different patterns used to obtain the properties for the event "change". The patterns consist of the text string that appears before and after the value of the property[2].

For example, in the soccer match Chelsea vs. Arsenal, the different changes are expressed with the sentences: "*Ljungberg ocupa el puesto de Van Persie.*" (Ljungberg replaces Van Persie), "*Robben entra por Shevchenko.*" (Robben comes in to substitute Shevchenko) and "*Wright-Phillips sustituye a Geremi.*" (Wright-Phillips substitutes Geremi). After identifying the type of event —using the information available in the third column of the table of the web page (Fig. 1(b))— the application can extract the information about the properties of the event from the information available in the fourth column, using the information exposed in Tab. 1.

3.2 The Automatic Tagging Tool

The tool uses the web page with the chronicle of the match from the web page of the on-line journal El País, as well as the XML file for the correspondent sport that contains all the events for this sport and the patterns for the properties of each event. Each row of the table for the chronicle

[1]Information Extraction is defined as the automatic extraction of "meaningful units" from semi-structured documents [4].

[2]As the source used for the chronicles is a Spanish journal, the chronicles are in Spanish, therefore, the patterns too.

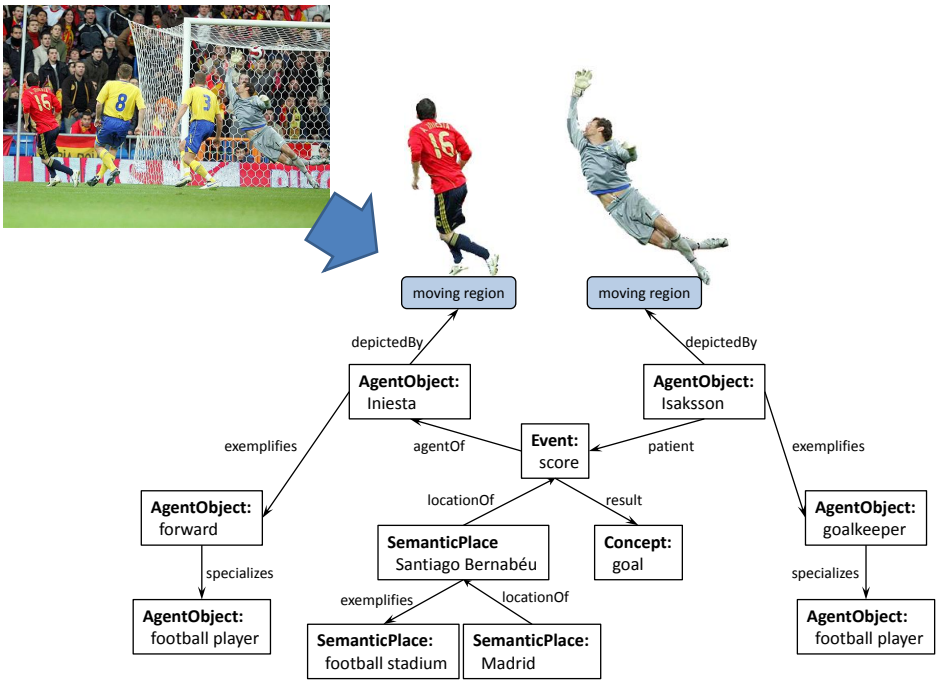


Figure 3: Annotation of sports videos using MPEG-7

contains the moment when the event took place, an image (for some of the events, such as red card, goal...), the type of event and its description (Fig. 1(b)). After identifying the event, the application uses the XML file to obtain the patterns for each property and analyses the description in order to find the actual values for these properties (Fig. 2).

The output of the application is currently a textual file with basic structured information about the events, it indicates the event, the moment when it happens and the values of its properties. However, the intention of this project is providing this information with a standardised structure. We have studied two different standards: TV-Anytime [13] and MPEG-7 [9]. We have chosen the later since it allows structuring the video in segments and writing semantical information about them, whereas TV-Anytime only allows applying keywords to these segments. Fig. 3 shows the annotation of a goal in MPEG-7 using our system.

4. EVALUATION

We have evaluated the results of the information extracted by the application by comparing it with human-made annotations. We have used a group of 30 undergraduate students to describe the events of five different soccer matches. They were provided with the segments of the videos where these events took place (according to the temporal information offered in the on-line journal ElPaís). For each segment, they were asked to identify the type of event and they had the possibility of adding values for the different properties of this event. The results were very encouraging. 95% of the events were identified in the same way as our application. Concerning the properties, 90% of the values extracted by our application have been also been provided by the majority of the students, and some of them have provided some values for additional properties that were not written in the chronicle.

5. PRACTICAL APPLICATION

As we have already introduced in Sec. 1, this application belongs to a broader project where the descriptions of TV contents are used to stablish relations between them. We are particularly interested in applying this approach in the field of t-learning (TV-based interactive learning [2]) where we have proposed two different learning experiences [11] that combine TV programmes with learning elements: *entercation* experiences, where the core element is a TV programme which is enhanced with related contents in order to use it as a bait to engage users in education; and *edutainment* experiences, where the learning elements are improved by adding related fragments of TV programmes.

Our intention is using the application developed to automatically tag sports videos in order to use them as a scenario to test these t-learning experiences. For example, an *entercation* experience can be created from a soccer video if we add a learning element about injuries prevention when a player injures himself, or a documentary about the town where the match takes place after the game. On the other hand, we can create an *edutainment* experience from a course for referees that is complemented with real scenes of soccer matches.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have exposed an application to tag videos automatically. This application uses simple natural language processing techniques to extract information about the events occurred in a sports video from the chronicles about the matches written in the on-line journal El País. Its main advantage with respect to the ones exposed in the related work is its simplicity and dependability, since it uses

a reliable source to obtain the information. Probably, the authors of the chronicles in the on-line version of the journal El País use an automatic chronicles writer where they select the event and the values for its properties, this application would randomly select one of the possible sentences it has for describing this event and fill the gaps with the values of the properties. Thus, the application uses inverse engineering mechanisms, which make the results very accurate.

This approach is very useful, not only for tagging the videos themselves, but also to help metadata creators to find the relevant events in a video to focus on the relevant fragments to write its description. In [14], an application is presented that helps content providers in the manual creation of MPEG-7 metadata. An example is exposed where soccer videos are described. In order to find the relevant events of the video, it uses a voice recogniser and looks for a series of words, for example "penalty". This system could be greatly enhanced substituting the voice recogniser by the application that we have presented in this paper.

As a future line of this project, the application will be completed to look for chronicles from different sources in the Internet and use them to contrast the information and create the metadata as a combination of these sources, applying different weights to them according to their reliability, improving hence the quality of labelling. In addition, we are working on the implementation of an interface to make easier the creation of language pattern files —for the moment these files should be created manually, writing directly the XML files— helping the user with the syntax but also looking in the chronicles for possible patterns.

7. REFERENCES

[1] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati. Semantic annotation of soccer videos: automatic highlights identification. *Computer vision and image understanding*, 92(2-3):285–305, 2003.

[2] P. J. Bates. A Study into TV-based Interactive Learning to the Home. `http://www.pjb.co.uk/t-learning`, 2003.

[3] M. Bertini, R. Cucchiara, A. D. Bimbo, and C. Torniai. Video annotation with pictorially enriched ontologies. In *IEEE International Conference on Multimedia and Expo (ICME 2005)*, The Netherlands, 2005.

[4] P. Buitelaar and T. Declerck. Linguistic Annotation for the Semantic Web. In *Annotation for the Semantic Web*, pages 93–110. IOS Press, 2003.

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

[6] L. Gazendam, V. Malaisé, G. Schreiber, and H. Brugman. Deriving semantic annotations of an audiovisual program from contextual texts. In *First International Workshop on Semantic Web Annotations for Multimedia (SWAMM)*, Edinburgh, 2006.

[7] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[8] R. Leonardi and P. Migliorati. Semantic indexing of multimedia documents. *IEEE Multimedia*, 9(2):44–51, 2002.

[9] Moving Pictures Experts Group (MPEG). Information Technology - Multimedia Content Description Interface. Part 5: Multimedia Description Schemes. In *International Standard ISO/IEC 15938-5*. 2003.

[10] Y. Park and Y. Li. Extracting salient keywords from instructional videos using joint text, audio and visual cues. In *Human Language Technology Conference of the North American Chapter of the ACL*, pages 109–112, New York, USA, 2006. Association for Computational Linguistics.

[11] M. Rey-López, A. Fernández-Vilas, and R. P. Díaz-Redondo. A Model for Personalized Learning Through IDTV. In Springer-Verlag, editor, *Adaptive Hypermedia, Adaptive Web-Based Systems 2006 (AH2006)*, volume 4018, pages 457–461, Dublin, Ireland, 2006.

[12] H. Saggion, H. Cunningham, K. Bontcheva, D. Maynard, O. Hamza, and Y. Wilks. Multimedia indexing through multi-source and multi-language information extraction: the mumis project. *Data Knowledge Engineering*, 48(2):247–264, 2004.

[13] The TV-Anytime Forum. Broadcast and On-line Services: Search, select and rightful use of content on personal storage systems. European Standard ETSI TS 102 822, 2004.

[14] C. Tsinaraki, P. Polydoros, F. Kazasis, and S. Christodoulakis. Ontology-based Semantic Indexing for MPEG-7 and TV-Anytime Audiovisual Content. *Special issue of Multimedia Tools and Application Journal on Video Segmentation for Semantic Annotation and Transcoding*, 26:299–325, 2005.

[15] X. Yu, H. W. Leong, C. Xu, and Q. Tian. Trajectory-based ball detection and tracking in broadcast soccer video. In *ACM Multimedia*, volume 3, pages 11–20, Berkeley, CA (USA), 2003.