TEXT MINING AS SUPPORT FOR SEMANTIC VIDEO INDEXING AND ANALYSIS

Jan Nemrava, Vojtěch Svátek

University of Economics, Prague
W. Churchil Sq. 4, 130 68 Prague-CZ

Paul Buitelaar, Thierry Declerck

DFKI Saarbrücken
Stuhlsatzenhausweg 3, 66123 Saarbrücken-D

ABSTRACT

This paper presents our work in the field of semantic multimedia annotation and indexing with the use of complementary textual resources analysis. We describe the advantages of complementary sources of information as a support for annotation and test whether these data can be used for automatic annotation and event detection.

1. INTRODUCTION

In this paper we present our work using the complementary textual resources in video analysis. This, for the selected domain (soccer in our case) concerns various textual sources such as structured data (match tables with teams, player names, score goals, substitutions, etc.) and semi-structured, textual web data (minute-by-minute match reports – unstructured text accompanied with temporal information). Events and entities detected in these sources are marked up with semantic classes derived from an ontology on soccer by use of information extraction tools. Since the target audience comes from various research areas, this text will be focused on the potential use of this approach rather than on the description of the details. Temporal alignment of primary video data (soccer match videos) with semantically organized events and entities from the textual and structured complementary resources can be used as indicator for video segment extraction and semantic classification; e.g. the occurrence of a 'Header' event in the complementary resources will be used to train and later classify the corresponding video segment accordingly. This information can then be used for semantic indexing and retrieval of events in soccer videos, but also for the targeted extraction of audio/visual (A/V) features (motion, audio-pitch, field-line, close-up). We denote such extraction of A/V features based on textual evidence "cross-media feature extraction".

There is quite a lot research effort carried out in the field of semantic annotation and indexing in the sports domain. Some of them, such as the work by [9], also use the complementary resources, but (in this case) not to the extent as we do. For further related work see [4] [2] [1].

2. RESOURCES COMPLEMENTARY TO A/V STREAMS

The exploitation of related (complementary) textual resources, especially when these are endowed with temporal references can largely increase the quality of the video analysis, indexing and retrieval. Of course the number of domains containing freely available detailed temporal descriptions is limited, but those where this information is available in a large scale can be very effectively used. Multiple parallel descriptions of one event will further increase the coverage and eliminate false events. Good examples can be found in the sports domain. Current research in sports video analysis focuses on event recognition and classification based on the extraction of low-level features and is—when based solely on the low-level features—limited to a very small number of different event types, e.g. 'scoring-event' [8]. On the other hand, complementary resources can serve as a valuable source for more fine-grained event recognition and classification.

We distinguish between two different kinds of information sources according to their direct vs. indirect connection to the video material. Primary complementary resources include such information that is directly attached to the media, namely, overlay texts, audio track and spoken commentaries. Secondary complementary resources include information that is independent from the media itself but related to its content, but it must be identified and processed first.

3. COMPLEMENTARY TEXTUAL RESOURCES AND VIDEO INDEXING

Major sports events, such as the FIFA Soccer World Cup Tournament that was held in Germany in 2006, provide a wide range of available textual resources, ranging from semi-structured data in the form of tables on web sites to textual summaries and other match reports. The video material was analyzed independently of the research described here, see [8]. The results of analysis are taken as input for our research and consist of video segmentation, where each second is defined by a set of feature detectors,

i.e. Crowd detection, Speech-Band Audio Activity, On-Screen Graphics, Motion activity measure and Field Line orientation.

A dataset for ontology-based [7] information extraction and ontology learning from text (SmartWeb corpus) consists of a soccer ontology, a corpus of semi-structured and textual match reports and a knowledge base of automatically extracted events and entities.

Minute-by-minute reports are usually published at soccer web sites and enable people to 'watch' the match in textual form on the web. Processing several such reports in parallel increases the coverage of events and eliminates false positive events. We therefore rely on the following 6 different sources in this case: ARD, bild.de, LigaLive (all in German), and Guardian, DW-World, DFB.de (all in English); we apply the SProUT tool [3] on them. This effort resulted in an interactive non-linear event browsing demo presented in [6]. The next section describes experiments with event detection based on the general A/V detectors.

4. CROSS-MEDIA FEATURE EXTRACTION

The aim of the semantic annotation is to allow (semi)automatic detection of events in the video based on previously learned examples. The aim of this experiment was to test whether the general detectors are able to serve as sufficient source of information. For this experiment we used two manually-annotated soccer match videos, one as a training set and the other for the tests. We created additional derived features describing the previous and the next values of the detectors in the same time range as the event instance itself, providing us with better chance to capture the behavior of the detector in time. We used decision trees as machine learning algorithm and built up a binary classifier for each of the observed events. The task of the classifier was to decide whether the particular segment is or is not an event. By our observation, the detectors we used are too generic for fine-grained event detection but they can help detect a certain event in a given (usually one minute long) time range where the event was identified in the text. The table below shows that different detectors are important for different event types. This potentially allows detecting instances of event types based on observing only those detectors that are discriminative for them (this assumption is also used by the decision tree algorithm). The letters P, C, N represent Previous, Current or Next Value of the detector for particular event type. More details can be found in [5].

	crowd			audio			motion			closeup			field line		
	P	C	N	P	C	N	P	C	N	P	C	N	M	E	O
Foul	x				x				x			x	x		
Free kick		x				x							x	x	
Header			x					x		x			x	x	
Shot on goal					x				x			x	x	x	x
Foul + shot on goal					x				x			x	x	x	

Results of the cross-media feature selection

5. CONCLUSIONS AND FUTURE WORK

We presented an approach to the use of resources that are complementary to A/V streams, such as videos of football matches, for the semantic indexing of such streams. We further presented an experiment with event detection based on general A/V detectors supported by textual annotation. In [5] we showed that such event detection based on general detectors can quite satisfactorily act as binary classifier, but when trained to provide classification for more classes it performs significantly worse. Using classifiers similar to those we have tested together with complementary textual minute-by-minute information (providing minute-based rough estimates where a particular event occurred) can help in refining the video indexing and retrieval. The potential of this work is not only in the annotation for indexing and retrieval of multimedia but also as feedback to the video learning algorithm, so we see its role in the area of OCR and other video analysis areas where we have to deal with text.

6. ACKNOWLEDGEMENTS

This research was supported by the European Commission under contract FP6-027026 for the K-Space project. We thank D Sadlier and Noel O'Connor (DCU, Ireland) for providing the A/V data and analysis results.

7. REFERENCES

[1] Bertini M., et al.: Automatic annotation and semantic retrieval of video sequences using multimedia ontologies. MULTIMEDIA '06. ACM, New York, NY

[2] Castano S., et al.: Ontology Dynamics with Multimedia Information: The BOEMIE Evolution Methodology. In Proc. of International Workshop on Ontology Dynamics (IWOD) ESWC 2007 Workshop, Innsbruck, Austria

[3] Drozdzynski W., et al.: Shallow Processing with Unification and Typed Feature Structures - Foundations and Applications. In KI 1/2004.

[4] Lanagan J and Smeaton A.F.: SportsAnno: What do you think?, RIAO 2007 - Large-Scale Semantic Access to Content, Pittsburgh, PA, USA, 30 May - 1 June 2007.

[5] Nemrava J., et al.: Text Mining Support for Semantic Indexing and Analysis of A/V Streams, OntoImage Workshop at LREC 2008, Marrakech, Morocco, May 2008

[6] Nemrava J., et al.:An Architecture for Mining Resources Complementary to Audio-Visual Streams. In: Proc. of the KAMC workshop at SAMT07, Italy, Dec. 2007.

[7] Oberle D., et al.: DOLCE ergo SUMO: On Foundational and Domain Models in SWIntO (SmartWeb Integrated Ontology) Journal of Web Semantics: Science, Services and Agents on the World Wide Web 5 (2007) 156-174.

[8] Sadlier D., O'Connor N.: Event Detection in Field Sports Video using Audio-Visual Features and a Support Vector Machine. IEEE Transactions on Circuits and Systems for Video Technology, Oct 2005

[9] Xu H., Chua T.: The fusion of audio-visual features and external knowledge for event detection in team sports video. In Proceedings of the 6th ACM SIGMM Workshop on Multimedia information Retrieval, 2004