Text Mining Support for Semantic Indexing and Analysis of A/V Streams

Jan Nemrava[1,2], Paul Buitelaar[2] , Vojtěch Svátek[1] , Thierry Declerck[2]

Department of Information and Knowledge Engineering
University of Economics
Prague, Czech Republic

Language Technology Lab & Competence Center
Semantic Web, DFKI (GmbH)
Saarbrücken, Germany

E-mail: {nemrava,svatek}@vse.cz, {paulb,declerck}@dfki.de

Abstract

The work described here concerns the use of complementary resources in sports video analysis; soccer in our case. Structured web data such as match tables with teams, player names, score goals, substitutions, etc. and multiple, unstructured, textual web data sources (minute-by-minute match reports) are processed with an ontology-based information extraction tool to extract and annotate events and entities according to the SmartWeb soccer ontology. Through the temporal alignment of the primary A/V data (soccer videos) with the textual and structured complementary resources, these extracted and semantically organized events can be used as indicators for video segment extraction and semantic classification, i.e. occurrences of particular events in the complementary resources can be used to classify the corresponding video segment, enabling semantic indexing and retrieval of soccer videos.

1. Introduction

We present an experiment in the use of complementary resources for the semantic indexing and analysis of audio/visual (A/V) streams, i.e. in the domain chosen (soccer matches) this concerns structured web data (match tables with teams, player names, score goals, substitutions, etc.) and unstructured, textual web data (minute-by-minute match reports). Events extracted from these resources are marked up with semantic classes derived from an ontology on soccer by use of an information extraction system. Through the temporal alignment of the primary video data (soccer match videos) with the textual and structured complementary resources, these extracted and semantically organized events can be used as indicators for video segment extraction and semantic classification, i.e. the occurrence of a 'Header' event in the complementary resources will be used to classify the corresponding video segment accordingly.

This information can then be used for semantic analysis, indexing and retrieval of soccer videos, but also for the selection of A/V features (motion, audio-pitch, field-line, close-up, …) for specific soccer event types, e.g. a CornerKick event will have a specific value for the field-line feature (EndLine), a ScoreGoal event will have a high value for the audio-pitch feature, etc. As such identification of characteristic features is based on textual evidence we call this ’cross-media feature selection and extraction’.

The remainder of this paper is organized as follows. In section 2 we will discuss the nature and potential use of complementary resources in video analysis. In section 3 we present the experiment we did on using complementary resources in the analysis and semantic annotation of soccer match videos. In section 4 we discuss our approach to the extraction of ’cross-media features‘ and finally in section 5 we draw some conclusions of our work and look forward to future work.

2. Resources Complementary to A/V streams

Despite the advances in content-based video analysis techniques, the quality of video analysis, indexing and retrieval would strongly benefit from the exploitation of related (complementary) textual resources, especially if these are endowed with temporal references. 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 limited to a very small number of different event types, e.g. ’scoring-event’. On the other hand, complementary resources can serve as a valuable source for a more fine-grained event recognition and classification.

When describing complementary resources 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 – it must be identified and processed first. The next two sections describe each of these in more detail.

2.1 Primary Complementary Resources

Although primary complementary resources are not the main focus of our current research and remain more in the field of low-level analysis, we consider them as a valuable source of relevant information. Apart from the audio track containing spoken commentaries we can make use of overlay text that is present in the video picture. The audio track of sports events is however unfortunately known for a very high Word Error Rate on automatic speech recognition (Sturm et al., 2003), even when dealing with a limited vocabulary such as player names and likely events.

We decided therefore not to use the audio track information in our research. The overlay text (a typical example is the time counter in sport events reporting as shown in Figure 1 below) instead provides us with very important information about the time offset between the video file time and the real match time. This information is crucial for the alignment of events extracted from complementary text resources with the low-level video analysis results.

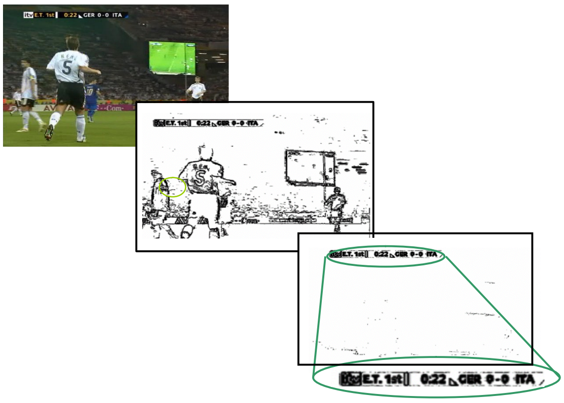


Figure 1: Primary Complementary Resources Example

2.2 Secondary Complementary Resources

The focus of our work is on the use of secondary complementary resources that come in the form of semi-structured tables, containing the summary of statistical, numerical and categorical data connected with events covered by video broadcasts (such as soccer matches) and in the form of unstructured textual reports containing detailed descriptions about particular events covered by video broadcasts including time point information.

Semi-structured as well as unstructured match reports can be readily obtained from web sources and information can be extracted by use of wrappers based on regular expressions in the case of semi-structured tables or of more sophisticated techniques that involve NLP-based information extraction in the case of unstructured text reports (Nemrava et al., 2007).

For our purposes we used semi-structured match tables as well as so-called minute-by-minute match reports, which combine unstructured text information on typical events that are not covered by the tabular match reports with a level of temporal structure through time points, i.e. indication of minute in the match.

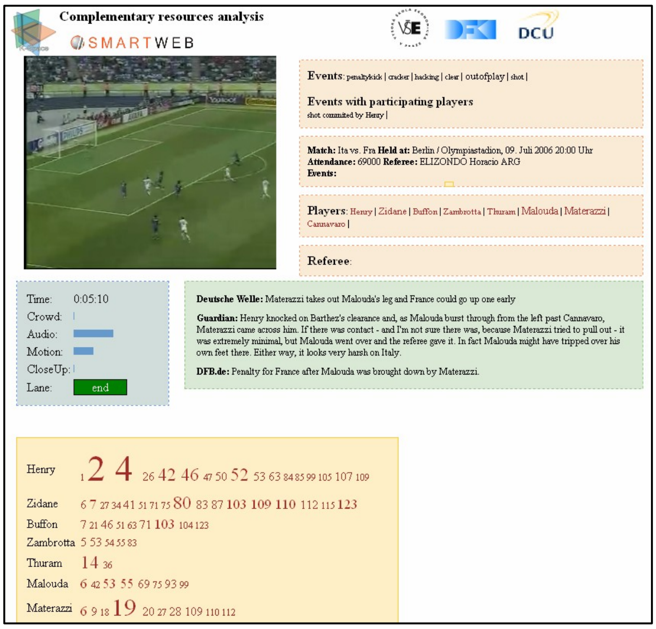
3. Semantic Indexing of A/V Streams with Complementary Resources

Major sports events, such as the FIFA Soccer World Cup Tournament that was held in Germany in 2006, provide a range of readily-available resources, ranging from A/V material broadcasted by television or Internet, semi-structured data in the form of tables on web sites, to textual summaries and other match reports. For the research reported here, we used a data set of original videos of television broadcasted matches as primary data, enriched with complementary information that we extracted from web tables (based on the 'SmartWeb Data Set' described below) and textual minute-by-minute reports.

The video material was analyzed independently from the research described here as described in (Sadlier et al. 2005). The analysis results are simply taken as input for our research and consist of video segmentation, with each segment defined by a set of feature detectors, i.e. Crowd detection, Speech-Band Audio Activity, On-Screen Graphics, Scoreboard Presence/Absence Tracking, Motion activity measure, Field Line (for a more extensive discussion see below).

The SmartWeb Data Set1 is an experimental data set for ontology-based information extraction and ontology learning from text. The data set 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 game in textual form on the web. These reports provide valuable information including the exact time point when each event happened. Combining several of these reports will increase the coverage of events. We therefore identified and collected minute-by-minute reports from the following web sites: ARD, bild.de, LigaLive (in German) and Guardian, DW-World, DFB.de (in English).



Events: penaltykick | cracker | hacking | clear | outofplay | shot |

Events with participating players
shot commited by Henry |

Match: Ita vs. Fra **Held at:** Berlin / Olympiastadion, 09. Juli 2006 20:00 Uhr
Attendance: 69000 **Referee:** ELIZONDO Horacio ARG
Events:

Players: Henry | Zidane | Buffon | Zambrotta | Thuram | Malouda | Materazzi | Cannavaro |

Referee:

Time: 0:05:10
Crowd:
Audio:
Motion:
CloseUp:
Lane:

Deutsche Welle: Materazzi takes out Malouda's leg and France could go up one early

Guardian: Henry knocked on Barthez's clearance and, as Malouda burst through from the left past Cannavaro, Materazzi came across him. If there was contact - and I'm not sure there was, because Materazzi tried to pull out - it was extremely minimal, but Malouda went over and the referee gave it. In fact Malouda might have tripped over his own feet there. Either way, it looks very harsh on Italy.

DFB.de: Penalty for France after Malouda was brought down by Materazzi.

Henry	1 2 4 26 42 46 47 50 52 53 63 84 85 99 105 107 109
Zidane	6 7 27 34 41 51 71 75 80 83 87 103 109 110 112 115 123
Buffon	7 21 46 51 63 71 103 104 123
Zambrotta	5 53 54 55 83
Thuram	14 36
Malouda	6 42 53 55 69 75 93 99
Materazzi	6 9 18 19 20 27 28 109 110 112

Figure 2: Semanic Indexing Demo

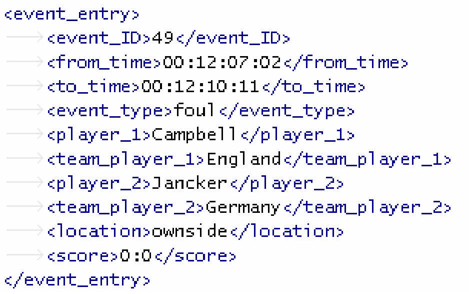
By use of the information extraction system SProUT (Drozdzynski et al, 2004) in combination with the SmartWeb soccer ontology (D. Oberle et al, 2007) we were able to derive a domain knowledge base from these resources, containing information about players (a list of

[1] http://www.dfki.de/sw-lt/olp2_dataset/

players names, their numbers, substitutions etc.), the match metadata (basic information about the game can contain information such as date, place, referee name, attendance, time synchronization information) and events (scoregoals, penalties, headers, etc.).
Obviously, such extracted information can be used to build up a semantic index of players and events in the match. Figure 2 depicts an example application of such semantic indexing using SMIL[2]. Various extracted information is aggregated and displayed along with the match video (A/V stream of a television broadcast), providing the user with direct access to events and entities occurring in the selected minute, while also enabling non-linear browsing through the match video.

4. Cross-Media Feature Extraction

Apart from the indexing and retrieval, information extracted from the complementary resources can be used also for the selection of A/V features specific for particular soccer event types. As such identification of characteristic features is based on textual evidence we call this ’cross-media feature selection and extraction’. Using machine learning techniques we try to determine discriminative features of selected football event types and build classifiers assigning appropriate event type to segments of A/V streams. These classifiers will allow creating a permanent connection between the textual information and the A/V analysis. We test whether the A/V detectors themselves are able to classify events of a certain kind. The following events were selected: foul, free kick, header, shot on goal, corner kick and goal. These events are all of different importance, as reflected also in the A/V streams by the time allocated to replays, crowd reaction, interruption etc.



```
<event_entry>
    <event_ID>49</event_ID>
    <from_time>00:12:07:02</from_time>
    <to_time>00:12:10:11</to_time>
    <event_type>foul</event_type>
    <player_1>Campbell</player_1>
    <team_player_1>England</team_player_1>
    <player_2>Jancker</player_2>
    <team_player_2>Germany</team_player_2>
    <location>ownside</location>
    <score>0:0</score>
</event_entry>
```

Figure 3: Textual Annotation Example

Data: We used two soccer matches from the Euro Cup 2000, one as training and the other as testing data. We used this data because these matches contained very detailed manual annotation (see Figure 3) created in the context of the MUMIS[3] project. Table 3 has the statistics of selected events. Unfortunately for the goal and corner kick event types the number of instances was insufficient for the experiment and we left them out.

	Match 1 **- Training Data -**	**Match 2** **- Test Data -**
Foul	31	28
Free kick	18	14
Header	27	22
Shot On Goal	8	17
Corner kick	3	8
Goal	7	1

Table 1: Training vs. Test data

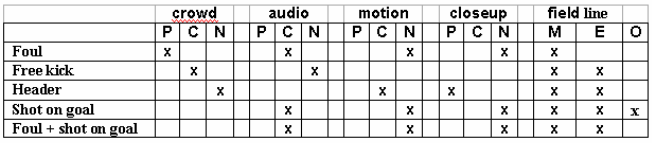
We first aimed at creating a binary classifier for every event type predicting whether the given video segment falls into a particular event type or not, rather than trying to built up one classifier over all event types. In other words, we wanted to know if a particular video segment is for example a foul or not. We later extended the classifier to a ternary classifier aiming at two event types predictions (fouls and shot on goals).

Creating derived values: Two problems occurred when we tried to build up a classifier for soccer events based on the A/V analysis. The first limitation is the generality of video detectors and their low number and the second is the fact that each second (or other time window) of the video analysis will be treated individually without regard to the previous and the next values (and thus behavior in time) of the detectors. We tried to overcome this by adding derived detectors describing the previous and the next values of the detectors in the same time range as the event instance itself (usually 3-5 seconds). We believe that this can help the machine learning algorithms to make a clearer distinction between the different event types. After this preprocessing we had 15 detectors in total. Basic ones are crowd, audio pitch, motion level and close-up detectors, derived ones are the previous and the following average values for each detector and the remaining three denote the proportion between the end-zone, middle zone and other zone of the soccer field based on the field line orientation within the video segment.

Train and test: For the given event type, every event element in the textual annotation file was associated with the appropriate video segment and its A/V analysis for the two matches. These data were labeled as training/testing data. The negative instances (i.e. non-event instances) were created by selecting segments of the A/V streams where none of the selected events occurred.

Building up a model: Decision trees provided the best performance over the given dataset. Table 2 shows the results from the experiment. The first 4 rows are the binary classifier and the results while the last two rows present results from the ternary classifier predicting three classes (2 event types and other)

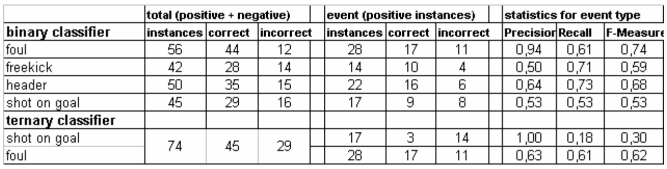
[2] http://www.w3.org/TR/REC-smil/
[3] http://lands.let.kun.nl/TSpublic/MUMIS/



	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	

Table 2: Results table

Apart from the classifier we wanted to test[4] whether we can identify A/V detectors that are more discriminative than others for particular event types:



	total (positive + negative)			event (positive instances)			statistics for event type		
binary classifier	instances	correct	incorrect	instances	correct	incorrect	Precisior	Recall	F-Measure
foul	56	44	12	28	17	11	0,94	0,61	0,74
freekick	42	28	14	14	10	4	0,50	0,71	0,59
header	50	35	15	22	16	6	0,64	0,73	0,68
shot on goal	45	29	16	17	9	8	0,53	0,53	0,53
ternary classifier									
shot on goal	74	45	29	17	3	14	1,00	0,18	0,30
foul				28	17	11	0,63	0,61	0,62

Table 3: Feature Selection

(P, C, N – Previous, Current, Next; M, E, O – middle, end, other)

The results in Table 3 show 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). However, a combination of several event types would lead to a conjunction of these discriminative features and would become too general.

5. Related work

There are several ongoing activities dealing with multimodal analysis and mapping across different resources. Very interesting work has been done by Xu (2004), also in the soccer domain. They also proposed a scalable framework that utilizes both internal AV features and external knowledge sources to detect events and identify their boundaries in full-length match videos. Besides detecting events, they focused on discovering detailed semantics and performing question answering. The difference was in the amount of textual sources they used and the number of features in the video analysis.

Another related work is SportsAnno (Lanagan, Smeaton, 2007), a video browsing system allowing users to read match reports taken whilst viewing the match video associated with the reports. The main difference is that they used videos summarizes and present free text information without any text processing or inf. extraction. The added value is a possibility for users to add comments as the basis for discussion and searching between all the users of the system. Bertini et al. (2006) used a multimedia ontology and MOM (Multimedia Ontology Manager) to automatically annotate manually pre-selected video clips. They also generated automatic clip subtitles.

The EU IST Project BOEMIE (Castano, 2007) focuses on the use of multimedia analysis results for population and enrichment of ontologies, in the athletics domain. Most published results of the project deal with still images.

6. 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. We showed that such event-detection based on general detectors can work as a binary classifier quite satisfactorily, but when trained to provide classification for more classes 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.

7. Acknowledgements

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

8. References

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

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

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

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.

Nemrava J., et al. : Architecture for mapping between results of video analysis and complementary resource analysis., K-Space Public Deliverable 5.10

Nemrava J., et al. :An Architecture for Mining Resources Complementary to Audio-Visual Streams. In: Proc. of the KAMC (Knowledge Acquisition from Multimedia Content) workshop at SAMT07, Italy, Dec. 2007.

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.

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

Sturm J., J. Kessens, M. Wester, F. de Wet, E. Sanders, H. Strik. (2003) Automatic Transcription of Football Commentaries in the MUMIS Project. In EUROSPEECH-2003, p 1853-1856.

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

[4] In this step we used the following attribute selectors from Weka Machine Learning Tool: CfsSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval