EVENT ALIGNMENT FOR CROSS-MEDIA FEATURE EXTRACTION IN THE FOOTBALL DOMAIN

Jan Nemrava[1,2], Paul Buitelaar[2] , Vojtech Svatek[1] , Thierry Declerck[2]

[1] Department of Information and Knowledge Engineering,
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
e-mail: {nemrava,svatek}@vse.cz
[2] DFKI (German Research Center for Artificial Intelligence) GmbH,
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
e-mail: {paulb,declerck}@dfki.de

ABSTRACT

This paper describes an experiment in creating cross-media descriptors from football-related text and videos. We used video analysis results and combined them with several textual resources – both semi-structured (tabular match reports) and unstructured (textual minute-by-minute match reports). Our aim was to discover the relations among six video data detectors and their behavior during a time window that corresponds to an event described in the textual data. Based on this experiment we show how football events extracted from text may be mapped to and help in anlysing corresponding scenes in video.

Index Terms— cross-media descriptors, text-to-video mapping, unstructured textual data mapping.

1. INTRODUCTION

Current research in sports video analysis focuses on event recognition and classification based on the extraction of low-level features. Such approaches are however mostly limited to a very small number of different event types, e.g. 'scoring-event'. On the other hand, there are vast textual and semi-structured data sources that can serve as a valuable source for finer-grained event recognition and classification. In particular, such complementary data can be exploited as background knowledge in filtering the video analysis results and thus for improving the corresponding algorithms.

The idea behind the research described here is to combine analysis results from textual and semi-structured data that are available as part of a football corpus [9] developed at DFKI in the context of the SmartWeb project [8] with video analysis results of corresponding football matches as made available by the Centre for Digital Video Processing at Dublin City University (DCU). The video data were analyzed using an approach [4] described in section 2.2.2, by which 'shot-on-goal events are extracted based on the combined information provided by six video descriptors.

The video analysis approach has a focus on the extraction of just one event type: 'shot-on-goal'. Instead, the work reported here has a focus on the extraction of additional football event types such as 'goals, substitutions, penalties, header, cards, whistles, passes, free kicks, corner kicks' and others that can be extracted from corresponding semi-structured and textual data. However all the information we extract from such resources must be accompanied by temporal information in order to enable alignment with the video analysis results. We therefore compiled a corpus of minute-by-minute reports as available on the web. The purpose of the experiment then concerns the alignment of extracted events from this unstructured, textual data with the events that are provided by the semi-structured data in the SmartWeb corpus and with events that were recognized by the video analysis results. This leads to a filtering of the video analysis results (as the textual/semi-structured data can pinpoint more accurately to the scoring-events) and a further event type classification (as some of the recognized events can be now more accurately identified as a 'penalty, header, card, …'). Additionally, the approach we describe here provides cross-media descriptors that can serve as feedback information in the DCU algorithm as well as other algorithms for (sports) video analysis that can be used in combination with corresponding textual and semi-structured data.

2. DATA PREPROCESSING

2.1 Data Resources

The Football World Cup 2006 held in Germany provides a wide range of information resources, ranging from textual and semi-structured data to video reports. We used this data for our research as follows: a data set of *semi-structured*

(tabular) match reports maintained by DFKI in the context of the SmartWeb project, a *video data set of corresponding matches* with video analysis results carried out at DCU, and *a textual data set of minute-by-minute reports* collected specifically for the research reported here.

2.1.1 SmartWeb Data Set

The SmartWeb Data Set is an experimental data set for ontology-based information extraction and ontology learning from text. The data set consists of:

- An ontology on the football (soccer) domain
- A corpus of semi-structured and textual match reports (German and English documents) that are derived from freely available web sources.
- A knowledge base of events and entities in the world cup domain that have been automatically extracted from the German documents.

For the purposes of the experiment described here we were mostly interested in the events that are described by the semi-structured data. On average there are 12 events for every match in the semi-structured data files.

2.1.2 Minute-by-minute reports

Minute-by-minute reports published at sport oriented web sites 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 can increase the probability of covering many of the events in the video that were detected by the DCU approach. We therefore identified and collected a number of German minute-by-minute reports from the following web sites: ARD [10] , bild.de [1] and LigaLive [11].

2.1.3 DCU Sports Video Analysis

DCU has a number of projects focusing on the extraction of scoring-events from field sports such as soccer, American football, rugby, Australian rules football, Gaelic football, hurling and others. The advantage of this approach is therefore that it is not limited to a specific sports-type. The video analysis provided to us by DCU have MPEG-7 descriptors with time points and confidence values. For the final 5 matches of the 2006 World Cup we got detailed information about the output from each of the descriptors used for event detection (see section 2.2.2). Each second of the match is described with several output values for the selected detectors. A combined confidence value for these selected detectors is computed by use of SVM (Support Vector Machines).

2.1.4 Information aggregation from different resources

We aggregated all information for each match into one file, containing extracted information from the *semi-structured data*, *minute-by-minute reports* and from the *video analysis*. The video analysis segments correspond to 90 seconds window around a detected event - 30 seconds before the event and 60 after the event - with attributes describing outputs from particular video detectors. We searched for a *single highest* value within the time interval from *t-30s* to *t+60s*, where *t* is the time when the event occurred in textual sources. After that we searched for a highest *sum of five seconds* values within the same time interval.

2.2 Preprocessing Tools

2.2.1 Text analysis with SProUT

SProUT (Shallow Processing with Unification and Typed Feature Structures) is a platform for development of multilingual shallow text processing and information extraction systems. It consists of several reusable Unicode-capable online linguistic processing components for basic linguistic operations ranging from tokenization to co-reference matching. SProUT takes the minute-by-minute reports as input, parses them and extracts relevant events. This results in approximately 250 events for each World Cup final match.

2.2.2 DCU video analysis

DCU created a framework for event detection in broadcast video of multiple different field sports [4]. An event model is inferred from evidence from feature detectors, which are chosen such that they are recyclable across multiple sports genres within the field sport domain. The techniques have been applied and tested generically across four distinct genres of field sport video. The preprocessing phase which tries to clear irrelevant time periods consists of Shot Boundary Detection, Probing Domain Restriction and Close-Up Image Detection and Shot Filtering. The following feature detectors are applied and combined by use of SVM (see also Figure 1):

- Crowd detection
- Speech-Band Audio Activity
- On-Screen Graphics
- Scoreboard Presence/Absence Tracking
- Motion activity measure
- Field Line

Second of video	Confidence	Crowd detection	Audio above thrshld	Missing scoreboard	Endzone	Visual Motion	Close-up scene
3	0.000000	0.000000	0.733333	1.000000	0.428571	0.117241	0.025000

Figure 1. DCU video data example

3. MAPPING ACROSS RESOURCES

Given the data resources and tools described above, we identified 45 different football event types from the minute-by-minute reports and SmartWeb semi-structured data set, which resulted in approximately 1200 event occurrences. A further alignment between the results from the different minute-by-minute resources reduced the number of events to 850 by identifying duplicate items (see Section 3.1). These events were then in a final step compared with the feature detectors from DCU in order to capture the relation between these different media resources (see section 3.2).

3.1 Minute-by-Minute Report Alignment

The first step in alignment was to align events that were extracted from the different textual resources (ARD, Bild, LigaLive). We extracted and aligned only those events that SProUT labeled as "player_action", "sportaction_results", "refereeaction", "goalkeeper_action", i.e. only those with important activity of the players. Here, we start with an aggregated list of events that contains duplicate events close to or within the same time period. This results from different styles of reporting, leading to slightly different time allocations for the same events. We assume that the first occurrence of an event within a sequence of events may be the closest to the time when the event actually happened or occurred in the game. We estimate that events within the following two minutes can be considered as the same event described by a "slower" reporter from a different resource

3.2 Match vs. Video Time

Additionally we had to cope with a difference between the time in the video data and actual match time in the semi-structured and textual reports. The fact that the time of the match mostly does not correspond to the time of the video may be a significant problem, because if we want to map events extracted from textual data (i.e. match time) on the data provided by DCU (video time) we need to know the actual difference between them. Even though the approach from DCU includes a preprocessing phase with a focus on detection and elimination of irrelevant time periods, there are cases where the length of the video exceeds the time spans defined in corresponding textual data. This derives from the fact that there is a coarse-grained level of temporal structure (expressed in minutes) in the textual data vs. a fine-grained temporal structure (expressed in seconds) in the video data. This was partially solved by searching within a time window in the video data around every event from the textual data. We tracked these differences manually for our purposes here, using an approach that does not rely on the confidence values in the DCU data but instead concentrates on detecting the time differences based on on-screen information such as scoreboard. In [3] we suggest an automatic approach for this using OCR.

4. CROSS-MEDIA FEATURE EXTRACTION

Using the aligned events from the semi-structured data, text reports and from the video analysis we are now able to extract cross-media features for these events. Cross-media features describe information contained in both textual/semi-structured data and video data in a way that they can be used as additional support in video analysis. In our experiment we aimed at discovering systematic behavior of video detectors in the context of events extracted from corresponding semi-structured and textual data in order to improve video analysis algorithms or for mapping video on to textual data. For some event types some rules can be expected (e.g. when we have "cornerkick" event type the "endzone" video detector should be significantly high), while some would require more detectors than those specifically used by DCU

```
<cornerkick id="6" occured="55" >
        <Confidence Max="0.958193"  Min="0.526706" Avg="0.7614" stdev="0.0951" />
        <Crowd Max="1"  Min="0" Avg="0.3465" stdev="0.2892" />
        <Audio Max="1"  Min="0.533333" Avg="0.9042" stdev="0.1361" />
        <ScoreBoard_Abs Max="1"  Min="0" Avg="0.7087" stdev="0.338" />
        <Endzone Max="1"  Min="0.142857" Avg="0.8016" stdev="0.2106" />
        <Motion Max="1"  Min="0.408" Avg="0.7253" stdev="0.1286" />
        <CloseUp Max="0.8125"  Min="0.102128" Avg="0.3602" stdev="0.1732" />
</cornerkick>
```

Figure 2. Cross-media descriptor for "cornerkick"

```
<penaltykick id="13" occured="12" >
        <Confidence Max="0.991206"  Min="0.629317" Avg="0.811" stdev="0.1018" />
        <Crowd Max="1"  Min="0.2" Avg="0.5313" stdev="0.2904" />
        <Audio Max="1"  Min="0.746667" Avg="0.9522" stdev="0.0937" />
        <ScoreBoard_Abs Max="1"  Min="0.072" Avg="0.869" stdev="0.2466" />
        <Endzone Max="1"  Min="0.285714" Avg="0.7381" stdev="0.2163" />
        <Motion Max="0.924832"  Min="0.434899" Avg="0.7192" stdev="0.1463" />
        <CloseUp Max="0.666667"  Min="0.27027" Avg="0.446" stdev="0.1198" />
</penaltykick>
```

Figure 3 Cross-media descriptor for "penaltykick"

for shot-on-goal detection. For this purpose we extracted event-specific video detectors given the events extracted from text and tabular data. For instance, we extract all occurrences of the event 'cornerkick' from the textual data (three minute-by-minute report in our case) and searched for the highest value of each detector within a window of 90 seconds (-30 +60) around the time when the event was described in textual data. From these detector characteristics we extracted the minimal, maximal and average number to characterize the detectors behavior.

As a first example, the detector behavior for event type "cornerkick" is shown in Figure 2. Here we can see that out of 55 cornerkick events detected in our data the Audio and EndZone detectors contain values which are significantly higher than the others. Relatively small standard deviation shows that even though the range of Audio values is quite large, the values are usually distributed closely around average. On the other hand the low average values for the CloseUp and Crowd detectors show that in case of cornerkicks CloseUp and Crowd shots are not very common.

A second example, as depicted in Figure 3 shows detector behavior for another frequent event type: "penaltykick". In this case the Audio and ScoreBoard Absence features are relatively high. However the Audio detector is above other detectors in most event types we extracted as most football events are accompanied by a crowd noise or commentators excitement. The standard deviation is very low, which indicates that throughout the time window that we searched the audio remains on a very high level.

The purpose of the cross-media descriptors is to capture the features and relations in multimodal data so as to be able to retrieve complementary information when dealing with one of the data sources. Starting from the audiovisual stream processing, cross-media descriptors may provide potential explanations for conspicuous values of lower-level detectors and thus refine the notion of 'highlight' event towards more semantic categories. On the other hand, audiovisual stream, thanks to its higher granularity, allows to label events detected in textual data such as news reports with more precise time stamps (moving from the minute scale to the second one).

5. RELATED WORK

The idea of combining textual information with features extracted from either still photographs or video shots has been used for some time now. For instance in their 1994 paper, Srihari and Burhams focused on capturing visual semantics by identifying information from image captions and extracting features described in an ontology [6]. In recent years a large number of similar approaches have been described for making multimedia content indexable, searchable and manageable. A nice summary of these can be found in [2]. The contribution of our work in contrast to these is mainly in its use of complementary textual and semi-structured resources collected from web resources. These are more loosely connected to corresponding image or video data than image captions and surrounding text, which are more typically used in similar approaches.

The work that is most closely related to our approach is that of the MUMIS project [5], which focused on alignment of extracted events across multiple, multilingual resources. The main aim of MUMIS was to provide textual content for better multimedia indexing and search. It did however not focus on the role of textual or semi-structured data in the analysis of video descriptors as described here.

6. CONCLUSION

Our work focused on aligning events extracted from semi-structured and textual data with extracted video features. The results of the presented work can hopefully be used as a feedback to the process of video analysis. In future work we hope to develop integrated software tools for carrying out the event extraction in text, tables and video, the event mapping and the cross-media feature extraction.

7. ACKNOWLEDGMENT

This research was supported by the European Commission under contract FP6-027026 for the K-Space project. Paul Buitelaar was also partially supported by the SmartWeb project, funded by the German Ministry of Education and Research under grant 01 IMD01. We would like to thank Noel O'Connor and James Lanagan (DCU) for providing the video data and analysis results.

8. REFERENCES

[1] Bild.de: http://wm2006.sportbild.de/index.html?spielplan
[2] Barnard, K. and Forsyth, D. (2000). Learning the semantics of words and pictures, CS Division, UC Berkeley.
[3] Nemrava, J., Svatek, V., Declerck, T., Buitelaar, P., Zeiner, H., Alcantara, M.: Report on algorithms for mining complementary sources. K-Space Deliverable D5.4
[4] Sadlier D and 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
[5] Saggion, H., et al: Event-coreference across multiple, multi-lingual sources in the Mumis project. In Proceedings of EACL. Budapest, Hungary, April 12 - 17, 2003.
[6] Srihari R. K.and Burhans D. T.. Visual Semantics: Extracting Visual Information From Text Accompanying Pictures. In Proc. of AAAI'94, pages 793-798
[7] SProUT website: http://sprout.dfki.de/
[8] SmartWeb website: http://smartweb.dfki.de/
[9] SmartWeb Corpus: http://www2.dfki.de/sw-lt/olp2_dataset/
[10] ARD Live Ticker: http://sport.ard.de/wm2006/wm/
[11] WM 2006 Live Ticker: http://www.ligalive.net/ticker_wm