Architecture for Enhancing Video Analysis Results using Complementary Resources

J. Nemrava[1,2], P. Buitelaar[2], T. Declerck[2], V. Svátek[1], J. Petrák[1], A. Cobet[3], H. Zeiner[4], D. Sadlier[5], N. O'Connor[5]

[1)] Univ. of Economics, Prague, CZ [2)] DFKI Saarbruecken, DE [3)] TU Berlin, DE, [4)] Joanneum Research, Graz, AT [5)] Dublin City Univ., IR

***Abstract*— In this paper we present different sources of information complementary to audio-visual (A/V) streams and propose their usage for enriching A/V data with semantic concepts in order to bridge the gap between low-level video analysis and high-level analysis. Our aim is to extract cross-media feature descriptors from semantically enriched and aligned resources so as to detect finer-grained events in video. We introduce an architecture for complementary resources analysis and discuss domain dependency aspects of this approach connected to our initial domain of soccer broadcasts.**

***Index Terms*— Multimedia databases, Text processing**

I. Introduction

Current progress in technology together with changing lifestyle towards mobile devices and on-demand video delivery allows a user to view video content almost anywhere at any time. However people are often interested only in some kind of highlight events within a larger volume of audiovisual data. This counts for almost all video genres available. Despite the advances in content-based video analysis techniques, the quality of video retrieval would strongly benefit from exploitation of related (complementary) textual resources, especially if these are endowed with temporal references. Good examples can be found within the sports domain [1]. 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, vast textual data sources can serve as a valuable source for finer-grained event recognition and classification. In particular, textual data can be exploited as background knowledge in filtering the video analysis results and thus improve the corresponding algorithms.

We introduce a generic architecture for complementary resource exploitation, and discuss domain dependency issues. We further describe concrete sources of complementary information that we came across in our application domain (soccer), such as real-time game logs (minute-by-minute reports), textual summaries found on websites, OCR on the video content, speech transcripts and others, and we describe the ways they can be merged together to create a coherent textual match description. We also relate them to the core video analysis framework.

II. Proposed Architecture and Resources

Figure 1 shows the proposed framework [3]. The process of gathering, alignment and mapping of the complementary data to video can be divided into five different phases:

1. a) Gathering and preprocessing the textual sources,
   b) Building the database of video analysis results
2. Mutual synchronization of the textual sources and building database of ontological resources.
3. Alignment of video data, textual data and knowledge db.
4. a) Time synchronization of textual and video data.
   b) Event type recognition
5. Annotation results provision and cross-media feature extraction. The former directly link concrete events in video to semantic categories. The latter characterize semantic categories in terms of typical values of video-oriented descriptors.

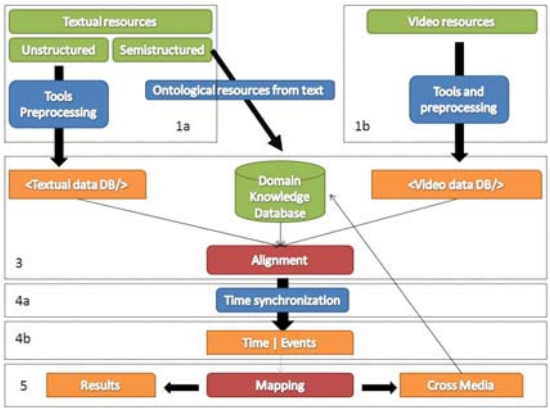

Figure 1. Proposed Architecture

Textual resources are the main semantics carrier within this architecture. *(Semi-)Structured*, database-like *textual resources* contain the summary of larger-scale events covered by individual video broadcasts (such as soccer matches), and, thanks to their clean structure, provide valuable and easily extractable information. On the other hand, *unstructured* event reports require more sophisticated techniques; for example, they can be extracted from web sites using wrappers and then information extraction tool [5] can be applied in order to extract domain-related ontology concepts.

Minute-by-minute reports are an example of free text information structured by the time points of finer-grained events not covered by structured 'protocols'. Combining

several of these reports [7] can increase the probability of covering unidentified highlights detected by the video analysis.

While web-based reports and semi-structured 'databases' can be viewed as 'secondary' complementary resources, as they are not directly connected to the video, **OCR** on text present in videos in the form of overlays represents a typical example of 'primary' complementary resources. Temporal alignment is not needed here. General text detection in video is more complex, because text may appear for example on signs (shop name, city name, street names…), on non-rigid objects (e.g. T-Shirt of a person with text) and so on. The OCR tests were done on key frames from soccer matches. This promising source of information will provide us with more textual information about what is actually happening in the game and – what's most important – *provides a way to synchronize video file time with the actual time of the match*, using the scoreboard analysis.

A sequence of 16 frames is used for detection of text as it carries information about moving objects and static text. For example in a soccer game the camera is moving all the time or the players are moving, but the text with the score, time, and teams are always on the same place in the frames. Thus most of the moving objects are removed and only the static edges of the text areas are left. In this way it is possible to detect the text regions as shown below on Figure 2.

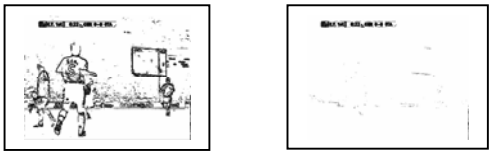

Figure 2. Fusion of frames for moving object detection

A different kind of 'primary' complementary resources are **speech transcripts.** Their analysis on the soccer domain is described in the MUMIS project [4]. The baseline automatic speech recognition system yields Word Error Rates (WERs) that varied from 84% to 94% for the different languages and test matches. Application specific words such as players' names are recognized correctly in about 50% of cases. The reason why the WERs are so high is because of the extremely high level of stadium noise present in the audio track.

A domain knowledge base serves as an ontological resource. In our soccer domain, it contains information such as *player database* (a list of players names, their numbers, substitutions etc), *match metadata* (basic information about the game such as date, place, referee name or attendance) and *event database* (list of possible event types and their cross-media feature descriptions).

The scope of **video analysis** tools is very broad and differs from one domain to another. In our experiments, we relied on low-level feature extraction from soccer matches as relying on six generic sports-video detectors (i.e. crowd/spectator detection, audio energy envelope, close-up detection, scoreboard activity measure, motion activity quantification, field-line extraction). More details may be found in [2].

III. Domain Dependency

A crucial question is whether our experiments are to some degree reusable in different domains and settings than the analysis of soccer videos. The reusability can roughly be considered at several levels:

1. Different categories of soccer (soccer) matches and/or different styles of audio/video data recording
2. Different groups of sports, which can be determined according to multiple facets
   a. field sports vs. others
   b. temporally structured (and in gross time vs. net time where breaks and pauses are not included) vs. those structured by score (or similar non-temporal aspect)
   c. collective vs. individual sports.
3. Beyond the sports domain.

When looking at *non-sports* events, the applicability of our approach is strongly limited by the unavailability of temporal online reports in many cases as well as limited possibility to visualise events in more 'spiritual' domains such as politics.

IV. Conclusions

We presented an architecture for complementary resource mining and for their mapping to video analysis results, which also yields cross-media descriptors. The research has as starting point our experiments in the soccer domain. In our future research we will to focus on creation of sports-genre-specific detectors (e.g. for corner flag) and compare the performance of adding extra detector with the performance when using just general detectors. We want to test a video detector based on specific event-type described in the textual data. We believe that this will help us better assess and possibly overcome the domain dependency of our approach.

Acknowledgment

This research was supported by the European Commission under contract FP6-027026 for the K-Space project.

References

[1] Xu, H. and Chua, T. 2004. 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

[2] 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

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

[4] Sturm J. et al.: Automatic Transcription of Football Commentaries in the MUMIS Project. In EUROSPEECH-2003, p 1853-1856.

[5] Drozdzynski W., Krieger H.-U., Piskorski J., Schäfer U., Xu F.. Shallow Processing with Unification and Typed Feature Structures - Foundations and Applications. In Künstliche Intelligenz 1/2004.

[6] Arndt, R. et al.: Architecture Specification of the K-Space Annotation Tool, K-Space Deliverable D5.11. Public

[7] Nemrava J., Buitelaar P., Svátek V., Declerck T.: Event Alignment For Cross-Media Feature Extraction In The Football Domain. WIAMIS Santorini : IEEE Computer Society, 2007, s. 1–3. ISBN 0-7695-2818-X.