An Architecture for Mining Resources Complementary to Audio-Visual Streams

J. Nemrava[1,2], P. Buitelaar[2], N. Simou[6], D. Sadlier[5], V. Svátek[1], T. Declerck[2], A. Cobet[3], T. Sikora[3], N. O'Connor[5], V. Tzouvaras[6], H. Zeiner[4], J. Petrák[1]

[1] University of Economics, Prague, CZ, {nemrava,svatek,petrak}@vse.cz
[2] DFKI Saarbrücken, DE, {paulb,declerck}@dfki.de
[3] Technical University Berlin, DE, {sikora,cobet}@nue.tu-berlin.de
[4] JOANNEUM RESEARCH, Graz, AT, herwig.zeiner@joanneum.at
[5] Dublin City University, IR, {sadlierd,oconnorn}@eeng.dcu.ie
[6] National Technical University of Athens, GR
{nsimou,tzouvaras}@image.ece.ntua.gr

Abstract. In this paper we attempt to characterize resources 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 detectors 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 resource analysis and discuss domain dependency aspects of this approach related to our domain of soccer broadcasts.

1 Introduction

Despite the advances in content-based video analysis techniques, the quality of video 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 within 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, 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 for retrieving relevant video segments.

In Section 2.1 we introduce a generic architecture for complementary resource exploitation. In the following Sections 2.2 and 2.3 we 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 possible ways they can be merged together to create a coherent textual match description. In the Section 2.4 we relate these information to the core video analysis framework. The Section 3 describes the work on reasoning over complementary resources, ending with a discussion of the domain dependency issues and plans for future work.

2 Architecture for Exploiting Complementary Resources

2.1 Overview of the Architecture

Figure 1 shows the proposed framework [9]. The process of gathering, aligning and mapping the complementary data to the video material can be divided into four different phases:

1. a) Gathering and preprocessing the textual sources,
   b) Building the database of video analysis results
2. Mutual synchronization of textual events based on temporal information.
3. Alignment of video analysis results with textual data and creating a knowledge database.
   a) Video feature extraction.
   c) Reasoning over complementary resources
   b) Event type recognition
4. 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

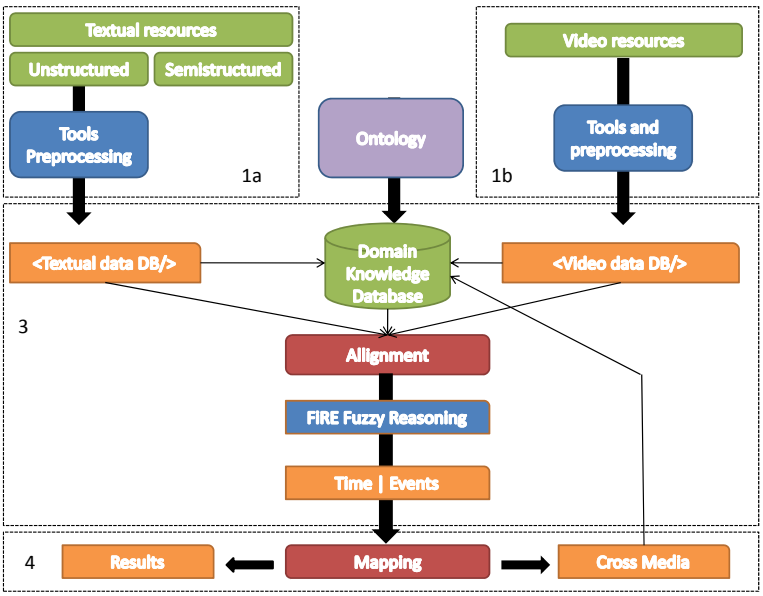

Fig 1.:Architecture for Complementary Resources

Textual resources are the main semantics carrier within this architecture. We use them to enrich the low-level audio-visual data with semantics because low-level video

detectors themselves are not able to provide this information. Web-based textual and semi-structured (tabular) match reports can be viewed as 'secondary' complementary resources, as they are not directly connected to the video while information available in the video file such as overlay text and spoken commentaries are a typical example of 'primary' complementary resources. These resources are described more in detail in the following sections together with the audio-video analysis detectors.

2.2 Primary Complementary Resources

Text present in videos in the form of overlays presents the main source of primary complementary resource. The region of such textual resources in video has first to be detected by means of image segmentation and then to be processed by OCR. General text detection in video is complex, since text may appear for example on signs (shop name, city name, street names), on non-rigid objects (e.g. a T-Shirt of a person with text) and so on. Therefore we will focus on overlay text in this section. The OCR will be applied on text detected in key frames of soccer matches videos. This promising source of information not only provides us with very valuable textual information about what is actually happening in the game, but it also - even more importantly - provides a way to synchronize video file time with the actual time of the events in the match, using the overlay time counter analysis.

A sequence of 16 frames is used for the 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 in Figure 2.

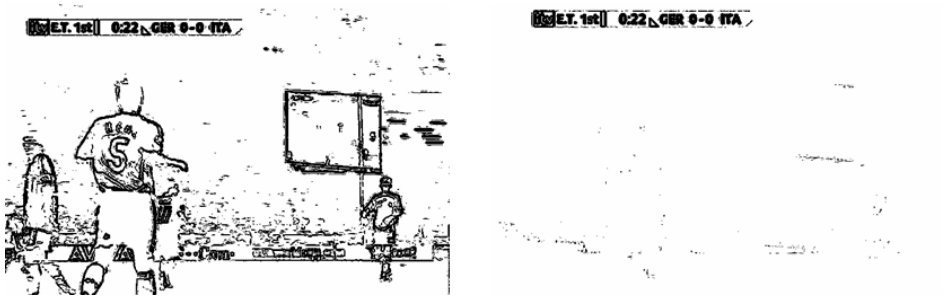

Fig 2.:Detection of moving objects in soccer broadcasts. On the right image above, all the moving objects have been removed

Captions are text fields occurring on the screen during the game. We can get valuable information about the shot on the player, current score, and basic events during the game (penalty, substitutions, corner-kicks etc.)

Other kinds of 'primary' complementary resources are speech transcripts of soccer commentaries. Their analysis in the soccer domain has been described in the MUMIS project [15]. 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. However automatic transcription of soccer commentaries, and all other broadcasts in which noise is mixed with the clean speech signals, could be improved substantially if the clean signals were available.

2.3 Secondary Complementary Resources

(Semi-)Structured, database-like *textual resources* contain the summary of statistical, numerical and categorical data connected with 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 contain detailed descriptions about particular events in the game including time point information. These reports however require more sophisticated techniques; for example, they can be extracted from web sites using wrappers and then NLP-based information extraction tools, like the one described in [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 and containing events that are not covered by structured 'protocols'. Combining several of these reports [10] can increase the probability of covering unidentified concepts from the video analysis.

A domain knowledge base can be built up from these resources. In the soccer domain, it contains information about *players* (a list of 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* (list of occurring events).

2.4 Audio-Video Analysis

The scope of **audio-video analysis tools** and **feature detection** is very broad and differs from one domain to another. This step can thus employ one-purpose AV analysis tool or general purpose annotation and video processing tool, such as the K-Space Annotation Tool (KAT) in order to obtain AV analysis results. KAT is a multimedia annotation tool and framework with an open architecture enabling customized plugins and unified output through an RDF repository. The first version of this tool is currently being developed in the context of the K-Space project (see www.k-space.eu). In our experiments, we relied on an event model that is inferred from evidence from feature detectors, which are chosen such that they are re-usable across multiple sports genres within the field sport domain [11]. Those techniques have been applied and tested generically across four distinct genres of field sport video. In particular we rely on the following six low-level feature generic sports A/V detectors: crowd/spectator detection, audio energy envelope, close-up detection, scoreboard activity measure, motion activity quantification, field-line extraction. More details can be found in [11].

3 Reasoning over complementary resources

In this section we will present a proposal for fuzzy reasoning over complementary resources. During last decade a substantial amount of work has been carried out in the context of Ontologies and Description Logics (DLs). DLs are a logical reconstruction of the so called frame-based knowledge representation languages, with the aim of providing a simple well-established declarative semantics to capture the meaning of the most prominent features in structured representation of knowledge [2]. On the other hand despite the rich expressiveness of classical DLs, they are insufficient to deal with vague and uncertain information which is commonly found in many real-world applications such as multimedia content. For that purpose a variety of DLs that can handle imprecise information in many flavors like probabilistic [6, 8] and fuzzy [13] have been proposed.

Our proposal for fuzzy reasoning over complementary resources is based on DL f-$\mathcal{SHIN}$ [12] which is the fuzzy extension of expressive $\mathcal{SHIN}$ [7]. In contrast to crisp DLs, the semantics of fuzzy DLs are provided by a *fuzzy interpretation* [14]. A fuzzy interpretation is a pair $\mathcal{I} = \langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$ where the domain $\Delta^{\mathcal{I}}$ is a non-empty set of objects and $\cdot^{\mathcal{I}}$ is a fuzzy interpretation function, which maps an individual name $\mathtt{a}$ to elements of $\mathtt{a}^{\mathcal{I}} \in \Delta^{\mathcal{I}}$ and a concept name A (role name R) to a membership function $\mathsf{A}^{\mathcal{I}} : \Delta^{\mathcal{I}} \rightarrow [0, 1]$

By reasoning over complementary resources, we refer to the automatic derivation of high-level semantic annotations from primary and secondary complementary resources associated with low-level multimedia data through the utilization of the provided domain knowledge that is soccer in this case.

A fuzzy knowledge base Σ consists of a fuzzy $TBox$, a fuzzy $RBox$ and a fuzzy $ABox$. $TBox$ and $RBox$ introduce the terminology, i.e the vocabulary of the application domain, while $ABox$ contains the assertions about named individuals in terms of this vocabulary. The alphabet of the concepts used in soccer domain is the following set and consist of the data that are extracted by the video and text analysis for a soccer game.

$Concepts = \{$*Scoringopportunity Outofplay Handball Kick Scoregoal Cross Foul Clear Cornerkick Dribble Freekick Header Trap Shot Throw Pass Ballpossession Offside SubstitutionAn Tackle Double Challenge Charge Lob Nutmeg GoalKeeperDive Block Save Booked* ***EndOfField MiddleField Crowd Motion CloseUp Audio*** $\}$

Since the features extracted from the text on a soccer game are characterized by coarse-grained minute information, while on the other hand video analysis features characterize every second of the game (bold in the concepts list above), the individuals set consist of the minutes and seconds of the game.

$Individuals$={min0 sec20 sec40 sec60 min1 sec80 sec100 sec120...},

where min0 corresponds to 1st minute in the game. Each minute is connected by the $consistOf$ role with four periods of 20sec (the time window consists of 80 seconds, from which 20 seconds are from the previous minute and 60 are from the minute described by the textual data) for which the main video features have been extracted using role assertion like these

$$(\langle min1, sec60 \rangle : consistOf) \geq 1$$
$$(\langle min1, sec80 \rangle : consistOf) \geq 1$$

$(\langle min1, sec100\rangle : consistOf) \geq 1$
$(\langle min1, sec120\rangle : consistOf) \geq 1$

The effective extraction of implicit knowledge from the explicit one requires an expressive terminology, which is able to define higher concepts. The definition of some representative event axioms for soccer domain are presented in the next table.

$$
\begin{aligned}
\mathcal{T} = \{\mathsf{Goal} &\equiv \mathsf{Scoregoal} \sqcap (\exists \mathsf{consistOfAudio}),\\
\mathsf{LongPass} &\equiv (\mathsf{Pass} \sqcup \mathsf{Kick} \sqcup \mathsf{Shot}) \sqcap (\exists \mathsf{consistOfMotion}),\\
\mathsf{CornerKick} &\equiv \mathsf{Cornerkick} \sqcap (\exists \mathsf{consistOfMotion}) \sqcap (\exists \mathsf{consistOfEndOfField})),\\
\mathsf{SubstitutionD} &\equiv \mathsf{Substitution} \sqcap (\exists \mathsf{consistOfMotion}) \sqcap (\exists \mathsf{consistOfMiddleField}),\\
\mathsf{HardFoul} &\equiv \mathsf{Booked} \sqcap (\mathsf{Foul} \sqcup \mathsf{Tackle})\\
&\qquad (\exists \mathsf{consistOf} \sqcap (\mathsf{CloseUp} \sqcup \mathsf{Audio})),\\
\mathsf{OffSide} &\equiv \mathsf{Offside} \sqcap (\exists \mathsf{consistOfEndOfField}),\\
\mathsf{ScoringOpportunity} &\equiv \mathsf{Scoringopportunity} \sqcap (\exists \mathsf{consistOfEndOfField})\\
&\quad \sqcap(\mathsf{Clear} \sqcup \mathsf{Shot} \sqcup \mathsf{Kick} \sqcup \mathsf{GoalKeeperDive} \sqcup \mathsf{Block} \sqcup \mathsf{Save}),\\
\mathsf{ScoringOpportunityFoul} &\equiv \mathsf{ScoringOpportunity} \sqcap \mathsf{Foul},\\
\mathsf{ScoringOpportunityCornerKick} &\equiv \mathsf{ScoringOpportunity} \sqcap \mathsf{CornerKick}\}
\end{aligned}
$$

Table 1. Knowledge Base ($TBox$). One can see how the features from the text are combined with detectors from video

This modeling associates the domain events with the time in which they occurred. This can prove very useful for various reasons. Firstly, in sports domains like soccer the exact time in which an event take place is very important. A user for example, would be able to semantically browse the video, retrieving all the minutes of the game in which goals were scored or hard fouls were made. Additionally, a relation of small periods of times (e.g. 5 minutes) similarly to the way that seconds are related to minutes could produce higher implicit knowledge. Such a period for example, consisting of minutes with hard fouls and booking events would imply a tough game. Furthermore, this representation permits the modeling of a sequence of soccer events, since they are described together with the possible subsequent events.

4 Domain dependency discussion

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 aspects can roughly be considered at several levels:

1. Different categories of 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.

The reusability is, in turn, bound to the availability of resources similar to those we used in our experiments such as the audio/visual recordings of the events, online textual reports and structured data/knowledge listing the entities involved in the events.

The reusability within *field sports* has been the focus of research at the DCU [11]. These are characterised by the following features: a) two opposing teams, b) enclosed playing area, c) field lines, d) commentator voice, e) spectator cheering, f) on-screen video text (scoreboard) and g) three well-defined styles of camera shot: global, zoom-in, and extreme close-up. The experiments in [11] were carried out for sports where the *scoring point* (goal, try or the like) is bound to a specific, restricted sub-area of the field where the ball (or similar object) is placed by a player. This is, in addition to soccer, also the case for e.g. rugby or (field) hockey. Note however that some other sports that are also played on a grass field (such as baseball and similar sports, and not considering at all individual sports such as some variants of tennis or athletics) don't share this feature; on the other hand, numerous non-grass sports such as basketball or ice hockey do have it. The latter then could represent a bigger challenge (or generally alter the setting) than soccer especially due to the following:

- The *field lines* may be harder to distinguish on unpredictably coloured floor (or ice) than they are on the green grass
- The *pace* of the game is often faster, which prevents the online web reporter to describe individual events in enough detail for matching with video content
- Due to the abundance of *scoring events* (often above 10 in ice hockey, several tens in handball, many tens in basketball etc.), the individual events have lesser importance and the task of their ad hoc retrieval is of lower interest. The central point of interest is the evolution of the score in a longer time frame.

The crucial difference occurs between the sports with a match structured according to a pre-defined time frame and those completed when a certain score (e.g. in tennis) or other situation (e.g. finish in races) is attained. A combination of both appears in some sports (e.g. ice hockey extra time) as principle of 'sudden-death': the game ends either on the event of one of the teams scoring or when the given time elapses. A similar model can appear in individual sports with the character of fight.

For *non-temporal sports*, we would need another guiding structure for matching online reports with video. This could be the score, but also other aspects of game progress, e.g. holes in golf or increasing height in athletics (high or pole) jumps: these typically appear as prefixes to online report items similarly to time in temporal sports. Here, similar to the races, the sports event culminates towards the end but important events can still appear at any time of the competition.

A practically important difference is that between games played with *net time* vs. *gross time*. Net-time sports require much more advanced synchronisation than e.g. soccer; the OCR-based analysis of the score-board, which is only auxiliary in our approach, would probably be indispensable there. The difference between *collective* and *individual* sports is particularly relevant to video analysis, where the identification of a concrete

object in motion obviously becomes easier with their decreasing number. Apart from that, the other aspects mentioned above have most impact.

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 visualize events in more 'spiritual' domains such as politics. Temporal reports sometimes appear in the web news documents related to the development of e.g. natural disasters or terrorist events; however, relevant videos are typically much shorter than sports recordings and are mostly likely to be viewed as wholes rather than subject of internal retrieval that are focusing on highlights.

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, also in the soccer domain [16]. 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 way they used the video analysis results.

The EU IST Project BOEMIE [3] focuses on the use of multimedia analysis results for population and enrichment of ontologies, in the athletics domain. So far most published results of the project deal with still images.

6 Future Work

While the present phase of the research focused on resolving temporal issues related to different granularity and asynchronity of multimedia and complementary textual results, the upcoming phase will focus on conceptual matching issues. Different events will be described using a logical representation, expressing different kind of relationships among the entities involved. We expect to leverage on the outcomes from the MUMIS project (where soccer situations identified in text were analysed with respect to semantic constraints) and from the BOEMIE project, where descriptions of sports events are expressed in description logics [4]. As both the video and complementary resource analysis can typically be burdened with uncertainty (or fuzziness), we plan to work using the fuzzified representation described in Section 3. FiRE is a reasoning engine that is based on the description logic language f-$\mathcal{SHIN}$ and can use the uncertain data produced by the text and video analysis together with the knowledge base to extract the event instances.

7 Conclusions

We presented an architecture for the use of complementary semi-structured and unstructured (textual) resources in video analysis. The central purpose of the architecture is to match semantically annotated complementary resources to corresponding video

material, e.g. in the sports domain, and to subsequently extract semantically organized 'cross-media features' from this aligned data set. Extracted cross-media features describe the behaviour of video detectors (audio, motion, close-up, etc.) relative to semantically annotated sports event types (shot, header, tackle, etc.). In further work we described the representation of extracted cross-media features and their use in event type classification of video segments by use of fuzzy reasoning with the FiRe inference engine. Our research has as starting point experiments in the soccer domain. In future research we will focus on cross-media analysis with video detectors based on specific event types described in complementary textual and semi-structured data (e.g. object recognition of a corner flag in the soccer domain). We believe that empirical exploration of domain specificity of various detectors will help us better assess and possibly overcome the domain dependency of our approach.

8 Acknowledgment

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

References

1. Arndt, R. et al.: Architecture Specification of the K-Space Annotation Tool, K-Space Deliverable D5.11. Public
2. Baader, F., McGuinness, D., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook: Theory, implementation and applications. Cambridge University Press (2002)
3. 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 - 7 June - Innsbruck, Austria
4. Castano S., Ferrara A., Hess G.: "Discovery-Driven Ontology Evolution" at the Semantic Web Applications and Perspectives - 3rd Italian Semantic Web Workshop (SWAP 2006), PISA, Italy, 18-20 December, 2006
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. Heinsohn J. et al.: Probabilistic description logics. In Proceedings of UAI-94, pages 311-318, 1994.
7. Horrocks I., Sattler, U., Tobies, S.: Reasoning with Individuals for the Description Logic SHIQ. In MacAllester, D., ed.: CADE-2000. Number 1831 in LNAI, Springer-Verlag (2000) 482496
8. Jaeger M. et al.: Probabilistic reasoning in terminological logics. In Proceedings of KR-94, pages 305-316, 1994.
9. Nemrava J. et al.: Architecture for mapping between results of video analysis and complementary resource analysis., K-Space Public Deliverable 5.10
10. 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.
11. 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

12. Stoilos G., Stamou G., Tzouvaras J., Pan J., Horrocks I.:The Fuzzy Description Logic f-SHIN, International Workshop on Uncertainty Reasoning For the Semantic Web (2005)
13. Stoilos G., Simou N., Stamou G., Kollias S.: Uncertainty and the Semantic Web, IEEE Intelligent Systems, 21(5), p. 84-87, 2006
14. Stoilos G., Stamou G., Pan J., Tzouvaras J., Horrocks I.: Reasoning with Very Expressive Fuzzy Description Logics, Journal of Artificial Intelligence Research, 30, 8, 273-320, 2007.
15. Sturm J. et al.: Automatic Transcription of Football Commentaries in the MUMIS Project. In EUROSPEECH-2003, p 1853-1856.
16. 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