Towards Cross-Media Feature Extraction

Thierry Declerck[1], Paul Buitelaar[1], Jan Nemrava[2], David Sadlier[3]

[1] German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
{declerck, Paul.Buitelaar}@dfki.de
[2] Engineering Group at the University of Economics in Prague (VSE)
W. Churchill Sq 4, 130 67 Prague, Czech Republic
nemrava@vse.cz
[3] Centre for Digital Video Processing at Dublin City University (DCU)
sadlierd@eeng.dcu.ie

Abstract

In this paper we describe past and present work dealing with the use of textual resources, out of which semantic information can be extracted in order to provide for semantic annotation and indexing of associated image or video material. Since the emergence of semantic web technologies and resources, entities, relations and events extracted from textual resources by means of Information Extraction (IE) can now be marked up with semantic classes derived from ontologies, and those classes can be used for the semantic annotation and indexing of related image and video material. More recently our work aims additionally at taking into account extracted Audio-Video (A/V) features (such as motion, audio-pitch, close-up, etc.) to be combined with the results of Ontology-Based Information Extraction for the annotation and indexing of specific event types. As extraction of A/V features is then supported by textual evidence, and possibly also the other way around, our work can be considered as going towards a "cross-media feature extraction", which can be guided by shared ontologies (Multimedia, Linguistic and Domain ontologies).

Introduction

The research presented in this paper is primarily concerned with the use of complementary textual resources in video and image analysis for supporting a higher level of automatic (semantic) annotation and indexing of images and videos. While in the past various projects (like the MUMIS project, see below for more details) used Information Extraction as the main mean for extracting relevant entities, relation and events from text that could be used for the indexing of images and videos in a specific domain, nowadays we can build on Semantic Web technologies and resources for detecting instances of semantic classes and relations in textual documents, and use those for supporting the semantic annotation and indexing of audio-visual content. So for example in the actual project K-Space[1] we used the SWIntO ontology [1] and the SmartWeb soccer ontology [2], for extracting relevant instances from various types of textual sources reporting on soccer games. Information Extraction systems, in our case the SProUT [3] tool, are then in fact instantiating the semantic classes derived from the ontologies used in the particular applications. In this we speak nowadays of Ontology-Based Information Extraction (OBIE).

This represents a progress for the "external" semantic annotation and indexing of audio-video, since the generated semantic annotation is more likely to be interoperable with other annotation generated in other context (by the use for example of the same or sharable upper ontologies), and is also more likely to be combined with recently emerging multimedia ontologies, which have also been developed in the context of various European projects, like aceMedia[2] and K-Space.

We aim specially at combining the semantic annotations extracted from text with the results of Multimedia analysis, which are given to us in the form of extracted audio/visual (A/V) features - such as motion, audio-pitch, field-line, close-up, etc. – and which are relevant for specific event types. In the particular context of the K-Space project we

[1] http://www.k-space.eu/
[2] http://www.acemedia.org/aceMedia

are working on the integration of semantic features extracted from both text and audio-video content, in the domain of sports, and more specifically soccer. This work is a collaboration activity between several partners: the DFKI Language Technology Lab[3], the Knowledge Engineering Group at the University of Economics in Prague[4] and the Centre for Digital Video Processing at Dublin City University (DCU)[5]. In this we try to respond to one of the main objective of K-Space: to narrow the 'Semantic Gap' between content descriptors that can be computed automatically by current algorithms, and the richness and subjectivity of semantics in high-level human interpretations of audiovisual media.

Related Work

The work we report here in the context of the K-Space project builds on related work in using textual data for the semantic annotation and indexing of soccer video as done for example in the past within the MUMIS project [5]. There, the main goal was to provide for a possibility to analyze textual sources related to soccer video. The textual base was given mainly by short structured summaries of a game (name of competition, names of teams, list of players for each team, name of the arena, name of the referee, and the listing of some central events, like goals, substitutions, and cardings, together with a time code and the actors involved, etc.). Another textual source consisted of on-line tickers, which give a more complex textual description of relevant soccer events, associated with temporal information that corresponds to a minute in the soccer game. An example of such a ticker information can have the following form:

Minute 13: A free-kick by Michael Ballack misses the goal of the English team

From this kind of text, the MUMIS information extraction systems [9, 10] were able to extract the relevant event - FreeKick - the entity associated with it - the soccer player "Michael Ballack", belonging to the German team - the opponent English team - and the time - minute - in which this event happened. The recognition of players and teams was facilitated by the use of structured summaries, from which gazetteers could be derived for the IE engines.
The information extraction was supported by a shallow semantic resource on soccer developed in the project. At this time we did not use the then emerging Semantic Web representation languages, but only XML. The MUMIS soccer resource has however been used as input for the SmartWeb project [4], both at the level of content and at the representational level [11]. This extended soccer ontology is also used in the K-Space project, as described below.
While the information extraction systems performed quite well, also being "corrected" by a merging tool [10], which checked the consistency and the completeness of annotations generated from the tools when applied on different documents reporting on the same game, there was a general bottleneck for using this information directly for the semantic annotation of soccer video: the temporal dimension. Events detected in MUMIS were encoded with minutes. And therefore it was difficult to align this semantic event with the video stream, where we have to deal with much shorter temporal units. Also the fact that video and audio analysis was not performed at all in this project, we could not take into account "repetitions" or longer breaks, etc. As a consequence, the detected and extracted events in the soccer game had to be manually aligned with the video stream. This was done for two games, but was very time consuming. However, in doing so, the MUMIS project could show the validity of the basic approach in extracting semantic information from complementary texts that could be used for indexing the corresponding video. Optimally, information extracted from complementary sources should be combined with the features extracted directly from the A/V stream. Most 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'. But there are vast textual data that can serve as a valuable source for more fine-grained event recognition and classification, along the line of a soccer ontology.

K-Space Approach to Cross-media Features Extraction

The K-Space project deals among others with the integration of knowledge structures, as encoded in high-level representation languages, and low-level descriptors for audio-video content, taking also into account knowledge that can be extracted from sources that are complementary to the audio/video stream, mainly speech transcripts and text surrounding images or textual metadata describing a video or images, or even text included in the images. These complementary resources typically fall into two groups: primary resources, which are directly attached to multimedia, and secondary resources, which are more loosely coupled with the audio-video material.

Combining features from video analysis and secondary external textual sources

As already mentioned above, we were able to build in K-Space on some results of the German project SmartWeb [4], and so to reuse a data-set that consists of an extended

[3] http://www.dfki.de/lt/
[4] http://keg.vse.cz/
[5] http://www.cdvp.dcu.ie/

soccer ontology (with bi-lingual labels in German and English) and the results of ontology-driven information extraction applied to structured and unstructured documents on games of the FIFA Soccer World Championship in Germany 2006[6].

A main advance in K-Space consisted in combining this resource on semantics and text analysis with work by DCU on the analysis of sports videos [12]. DCU applied its technology to video of soccer games, for which we had semantic data extracted from related texts in the SmartWeb corpus. This analysis is resulting in 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).

For the text analysis we used minute-by-minute reports from the following web sites: ARD, bild.de, LigaLive (in German) and Guardian, DW-World, DFB.de (in English). By use of the information extraction system SProUT in combination with the SmartWeb soccer ontology we were able to derive a domain knowledge base from these resources, containing 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 (scored goals, penalties, headers, etc.).

An important problem we had to solve in using this combined data was to know if the video time (the length of the video) corresponds to the length of the soccer game. 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 the content analysis of the video (video time) we need to know the actual difference between them.

The approach from DCU for analysing the content of the soccer videos includes a pre-processing phase with a focus on detection and elimination of irrelevant time periods. But still 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 can be 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, concentrating on detecting the time differences based on on-screen information such as scoreboard. However, with other K-Space partners we did some initial work on an automatic approach, based on OCR [6]. We can keep the presentation of this research aspect short, since in the same proceedings containing this paper, we have a short paper describing a demonstrator showing this kind of integration activity in action (the short paper with the title "Text Mining Support in Semantic Annotation and Indexing of Multimedia Data").

[6] SmartWeb used also official web pages of soccer organizations and teams.

Combining features from video analysis and secondary external textual sources

We concentrate in this section on preliminary work on the use of primary complementary textual resources, here text extracted from images by means of detection of textual regions in images and optical character recognition (OCR) for adding semantics to images and also to support audio/video analysis. We describe a first experiment on text extraction from news videos (news programmes of the German broadcaster "ARD"), for which we identified a list of relevant patterns of textual information appearing on the TV screen during those news programmes, and their relations to displayed images, which are particular in the sense that text belonging to one semantic unit might be distributed around an image. The use of the text for supporting the semantic annotation of their containing images implies an appropriate detection of the textual regions in the image, a high quality OCR (not at our disposal yet) and the applications of linguistic analysis of the extracted text.

In the following we just present 2 typical examples of presenting news information to the TV public, showing how the broadcaster (here the German public broadcaster ARD) combines image and text[7] to convey information.

[7] **We identified more patterns, which cannot be described here due to space limitation.**

Fig. 1. In this pattern, we can see the speaker and a background image, directly surrounded by two textual contributions (ignoring here the name of the News programme and the date)

In Figure 1, we consider the two textual regions being close to the background image, just above and below it. The text analysis tool applied to the extracted text can detect the topic ("decision of the Parliament about election") and also where it takes place (in "Kiew"). Interesting here: there is no linguistic hint, that this "decision" is being discussed in Kiew: We can infer this only on the base of heuristics applied to the distribution of words around the image. On the base of world knowledge, we can also infer that the Parliament presented here is the Ukrainian one (this information being most probably given by the speaker). Other information we can recognize: the voice to be heard in the audio streaming is in this pattern belonging to the news speaker. In case we know her name, this information can help in improving speaker recognition. In other patterns of information display, we can assume that the voice being heard is not from the speaker, but from someone presented in the image (or video).
It is still unclear in the first case we present, what kind of features of image analysis could be used for combining image and text analysis in one semantic annotation structure (maybe the detection of an "indoor scene"?).
A more complex pattern, with respect to semantic interpretation is shown in Fig. 2:

Fig. 2. We can see above the background picture a short phrase ("Allegations against son") and below the picture the name of a person. The text should be interpreted as "Allegations against the son of Annan".

In Fig. 2 the person name below the image is pointing to the content of the image. But the information on the top of the image is mentioning: "Allegation against son". So here we need also some inferences to get the point that the accusations are addressed against the son of the person shown in the image, but that the son is not shown here. An feature from image analysis that can certainly help here, is face detection (and better till, face recognition, while this is naturally very ambitious.
While the textual region detection tools used in our experiment perform quite good[8], but the OCR tools applied have still to be improved, and below we can see one error generated by the OCR procedure: the name of Kiew being represented as "Krew". But here when we know that we deal with Named Entities, a list (or gazetteer) of such Entities can be given to the OCR mechanisms for matching/correcting their results against it.
We can detect out of this extracted text from the image the different textual contributions. We cannot propose for an analysis of the string "Krew" for the time being (unknown word). The (automatic) linguistic dependency analysis of the textual contribution at the top of the image is giving:

[NP Parlamentsbeschluss *(head_noun)* [PP zur Wahl] *(noun_modifier)*]

The identified head noun is the main topic in text. And in fact this corresponds to the image, showing a parliament. The key frame is not about the „election", but about a parliament decision about the election. The dependency analysis allows to reduce considerably the number of key words that can be used for indexing, replacing them by structured textual fragment. We can also map the head noun onto ontology classes, so that the image might be also annotated with concepts like "political institutions" or the like. More detail on this preliminary work is given in [13].

A Model for the Integration of the Features of different Media

In the former sections of this paper, we have been presenting two approaches for combining features extracted from text related to images and videos with features

[8] This information is given on the base of a small informal evaluation done on the data. A formal evaluation is still to be proposed.

that can be extracted from images and videos by multimedia analysis methods. While the soccer use case is quite advanced and is showing already demonstrable results, the other use case is still in a preliminary investigation state.

The lesson that could be drawn for us from those studies, is the fact that the emergence of ontologies and the possibilities to organize them in a multi-layered manner, is offering the possibility to reach for a better integration of the features that are media specific. So in the K-Space project a Multimedia Ontology has been designed, in which an ontology of linguistic descriptors has been integrated. Work in the LSCOM project has been generating an expanded multimedia concept lexicon of more than 2000 concepts, of which slightly over 400 have been annotated in 80 hours of video (see http://www.lscom.org/concept.htm). A next step for us would be to integrate those concepts in domain ontologies that can also be "attached" to the Multimedia ontology, like we did in K-Space with the soccer ontology. A candidate model for the ontology-driven integration of cross-media, cross-domain and cross-lingual features, taken from [14] is graphically displayed shown below:

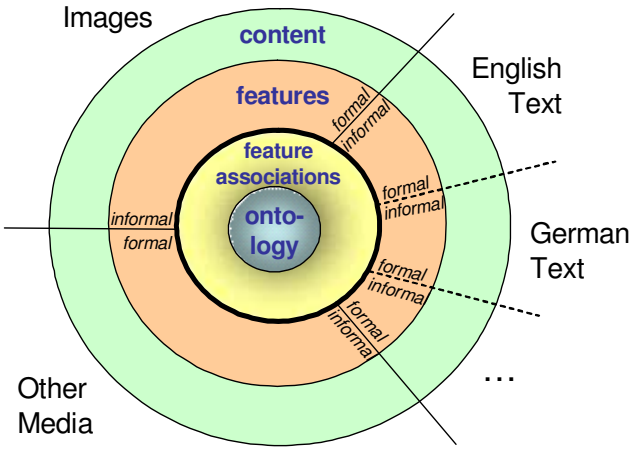

Figure 3: A candidate model for the integration of features to be detected in various media, in different domains and languages

Conclusions

We presented an approach to the use of resources that are complementary to soccer A/V streams, in such a manner that they can be used for the semantic indexing of such streams. We further presented event detection in the video based on general A/V detectors. Features extracted from both analysis types are aligned, and so the video indexing and retrieval can be refined with semantics extracted for the complementary resources. We also discussed a first experiment for using textual regions extracted from an image, for using them in combination of image features for the automatic annotation of images in news programme. Actual research is on the refinement of the model for the integration of cross-media features in an ontological framework that can support more effectively cross-media semantic extraction and annotation/indexing.

Acknowledgements

The research presented in the paper was supported by the European Commission under contract FP6-027026 for the K-Space project.

We would like to thank specially Andreas Cobet (Technische Universität of Berlin) for his help on the experiments with the detection and interpretation of textual regions in still images.

References

[1] http://www.smartweb-project.org/ontology_en.html
[2] http://www.dfki.de/sw-lt/olp2_dataset/
[3] http://sprout.dfki.de/
[4] http://smartweb.dfki.de/
[5] http://www.dfki.de/pas/f2w.cgi?ltc/mumis-e

[6] Nemrava, J., Svatek, V., Declerck, T., Buitelaar, P., Zeiner, H., Alcantara, M.: Report on algorithms for mining complementary sources. K-Space Deliverable D5.4.
[7] Wu, Y-F. Ma, H-J. Zhang, and Y-Z. Zhong, "Events recognition by semantic inference for sports video," Proc. *IEEE ICME 2002*, Lausanne, Switzerland.
[8] Zhong and S-F. Chang, "Structure analysis of sports video using domain models," Proc. *IEEE ICME 2001*, Japan.
[9] H. Saggion, H. Cunningham, K. Bontcheva, D. Maynard, O. Hamza, and Y. Wilks. Multimedia Indexing through Multisource and Multilingual Information Extraction; the MUMIS project. *Data and Knowledge Engineering, 48:247–264, 2003.*
[10] J. Kuper, H. Saggion, H. Cunningham, T. Declerck, F.M.G. de Jong, D. Reidsma, Y. Wilks, and P. Wittenburg. Intelligent multimedia indexing and retrieval through multi-source nformation extraction and merging. In 18th International Joint Conference of Artificial Intelligence (IJCAI), pages 409–414, Acapulco, Mexico, 2003.
[11] D. Oberle, A. Ankolekar, P. Hitzler, P. Cimiano, C. Schmidt, M. Weiten, B. Loos, R. Porzel, H.-P. Zorn, V. Micelli, M. Sintek, M. Kiesel, B. Mougouie, S. Vembu, S. Baumann, M. Romanelli, P. Buitelaar, R. Engel, D. Sonntag, N. Reithinger, F. Burkhardt, J. Zhou *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.
[12] 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
[13] Thierry Declerck, Andreas Cobet: Towards a Cross-Media Analysis of Spatially Co-located Image and Text Regions in TV-News. SAMT 2007: 288-291
[14] Buitelaar P., Sintek M., Kiesel M (2005). FeatureRepresentation for Cross-Lingual, Cross-Media Semantic Web Applications. In: *Proceedings of the ISWC 2005 Workshop "SemAnnot"*.
[15] Athanasiadis T., Tzouvaras V., Petridis K., Precioso F., Avrithis Y. and Kompatsiaris Y. (2005). Using aMultimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content. *In proceedings of the ISWC 2005 Workshop "SemAnnot"*.
[16] Jane Hunter: Enhancing the semantic interoperability of multimedia through a core ontology. IEEE Trans. Circuits Syst. Video Techn. 13(1): 49-58 (2003)