The Workshop Programme

9:00 9:20 Introduction & Setting
Stelios Piperidis

Cross-media mechanisms

9:20 9:45 Multimedia Semantic Analysis in the PrestoSpace Project
Valentin Tablan, Hamish Cunningham, Cristian Ursu

9:45 10:10 Cross-Document Coreference for Cross-media Film Indexing
Eleftheria Tomadaki, Andrew Salway

10:10 10:35 Cross-media Indexing in the REVEAL-THIS system
Murat Yakici, Fabio Crestani

10:35 11:00 The iFinder audio-visual indexing framework for cross media applications
Joachim Koehler

11:00 11:30 Coffee Break

Cross-media applications

11:30 11:50 Cross Media Aspects in the Areas of Media Monitoring and Content Production
Herwig Rehatschek, Michael Hausenblas, Georg Thallinger, Werner Haas

11:50 12:10 Representation and Analysis of Multimedia Content: The BOEMIE Proposal
D.I. Kosmopoulos, V. Karkaletsis, C.D. Spyropoulos

12:10 12:30 X-Media: Large Scale Knowledge Acquisition, Sharing and Reuse across media
Fabio Ciravegna, Stephen Staab and X-media consortium

12:30 12:50 Cross-media summarisation in a retrieval setting
Byron Georgantopoulos, Toon Goedeme, Stavros Lounis, Harris Papageorgiou, Tinne Tuytelaars, Luc Van Gool

12:50 13:10 From Media Crossing to Media Mining
Franciska De Jong

13:10 13:30 Discussion - Conclusions

Workshop Organiser(s)

Stelios Piperidis, Institute for Language and Speech Processing

Hamish Cunningham, University of Sheffield

Valentin Tablan, University of Sheffield

Workshop Programme Committee

Kalina Bontcheva (University of Sheffield, UK)

Fabio Crestani (University of Strathclyde, UK)

Gabriela Csurka (Xerox Research Centre, FR)

Eric Gaussier (Xerox Research Centre, FR)

Gregory Grefenstette (Commissariat à l'Energie Atomique, FR)

Paola Hobson (Motorola, UK)

Franciska de Jong (University of Twente, NL)

Joachim Köhler (Fraunhofer Institute, DE)

Yannis Kompatsiaris (Informatics and Telematics Institute , GR)

Harris Papageorgiou (Institute for Language and Speech Processing, GR)

Katerina Pastra (Institute for Language and Speech Processing, GR)

Stelios Piperidis (Institute for Language and Speech Processing, GR)

Laurent Romary (Loria, FR)

Tinne Tuytelaars (Katholieke Universiteit Leuven, BE)

Table of Contents

Author Index

Multimedia Semantic Analysis in the PrestoSpace Project

Valentin Tablan, Hamish Cunningham, Cristian Ursu

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, S1 4DP
Sheffield, UK

Abstract

PrestoSpace is a European-funded research project that aims at addressing the problem of decaying audio-visual archives throughout Europe by means of digitisation for preservation and access. One of the work areas within the project is Metadata Access and Delivery (MAD) which employs innovative methods of generating metadata for the digitised media in order to enhance the resulting archives and to ease access to the stored material. One such method is the use of automatic semantic analysis using natural language processing techniques in the process of creating analytical metadata for the preserved essence.

1. Introduction

Europe has a long-standing tradition of museums, archives and libraries for preserving its cultural heritage represented by paintings, sculptures, printed material or photographs. The 20th Century, through the advent of audio-visual technology, has started producing new types of media that need to be preserved films and several types of magnetic tapes for both audio and video material. Key events were recorded, and audio-visual media became a new form of cultural expression. These new types of material have also started to be preserved using traditional methods, by storing copies on shelves in large preservation facilities. The size of these archives is considerable The UNESCO estimates the size of the world audio-visual holdings to about 200 million hours, out of which around 50 million are in Europe. It has soon become apparent that this solution is not ideal because these new types of media suffer from chemical and physical decay (some films produce acetic acid vinegar syndrome, while all types of magnetic tapes become demagnetised over time). Another problem faced by the archives is technical obsolescence, there are fewer and fewer machines still capable of playing the older formats and keeping those functioning is becoming more and more expensive. In some cases even finding operators who are still qualified to operate those machines is becoming a problem as older personnel retires and new one is only trained for newer types of devices.

Although one possible solution would be to copy the legacy material onto newer storage formats, these operations would lead to loss of quality which is inherent to analogue processes. It is now widely accepted that the best available solution given the technical possibilities of today is to digitise the contents of the archives thus stopping the process of deterioration and freezing the quality levels at their current state. Starting from the digital copy, further transfers to new types of media will be possible with no loss of quality.

Throughout Europe large audio-visual archives, such as those managing the holdings of large broadcasting organisations, have already started the process of digitisation for preservation. This is an expensive process, the average cost for transfer from old to new media using the most cost-effective current technology is around 500 euros/hour a finding of the now ended Presto project. Budgetary restrictions mean that the current rate of transfer to digital for the most archives is not fast enough to ensure the preservation of the entire back-catalogue before it falls prey to decay. While an increase in budget would solve the problem, expecting that would be unrealistic. This is why the PrestoSpace project is addressing the issue starting from the other end by finding a way to lower the costs associated with the preservation process.

Better preservation and access also leads to increased reuse potential for the legacy audio-visual material, enabling media organisations to

extract more value from their holdings. This extra value can be returned as extra investment for preservation activities speeding up the digitisation effort and thus helping to save even more material from being deprecated and forgotten.

The next sections of the paper provide an overall view of the organisation of the PrestoSpace project, a more detailed view of the Metadata Access and Delivery work-area of the project and then it centres on the work done for automatic semantic analysis.

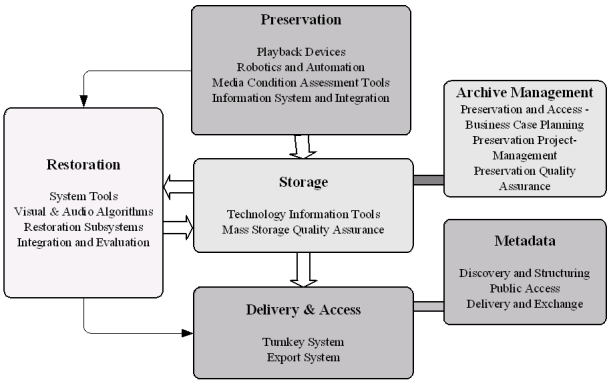

Figure 1: Overall organisation of the PrestoSpace project

2. The PrestoSpace Project

Audiovisual archiving is a complex and multidisciplinary domain spanning such diverse fields as chemistry, physics, signal processing, robotics and artificial intelligence. The challenge is to integrate partners of all domains representing the variety of competencies needed. The Project therefore brings together participants including 8 archive institutions, most of them representing the archives as well as their R&D departments, 3 applied R&D institutions, 6 academic institutes and 15 industrial partners.

The partners have analysed the different steps of preservation work towards access according to archives practises and to the required skills and technologies. The main production chain is the migration from analogue to digital material, including stock evaluation, identification and selection, the digitisation process and its control, the restoration, the storage and the production of content information (metadata) allowing for access and delivery.

Figure 1 depicts the projects work areas as well as the way they interact through the general work-flow. The Preservation work area is at the start of the chain and deals with the digitisation of the analogue media. All further processing is then performed on the digital copy. This area is concerned with robotics, hardware and software facilities dedicated to automating the process of digitisation to the highest possible level with a view to reducing the associated costs.

The next work area is Storage and Archive Management which aims to supply archives of all sizes with the required information and management tools so they can plan their own preservation process and keep track of their assets and the costs involved in moving from an analogue to a digital storage solution.

The Restoration work area provides an integrated restoration system that will be capable of analysing the digitised material, identify defects and apply the most appropriate software algorithms for correction. This will be a scalable system aimed at high throughput for a good enough quality at a low cost.

Metadata Access and Delivery MAD provides solutions to the problem of finding and making accessible the material preserved in the archives. This entails first generating metadata information describing the audio-visual items, by transferring the existing legacy metadata from the old analogue archives and by generating new information as a result of various content analysis processes and semantic analysis. Once the metadata exists, efficient retrieval methods are provided that combine the power of traditional information retrieval techniques with novel search methods based on conceptual search over the semantic metadata.

In order to help reducing the preservation costs, a factory approach is taken when the overall work-flow is designed. The various work areas interact creating a preservation chain that provides high throughput and good quality at a cost as low as possible. Human interaction is avoided wherever it can, being replaced by robotics and algorithms that can take decisions based on the setup of the system and the set of requirements.

2.1. The MAD Documentation Platform

Digital material can only be effectively accessed if metadata describing it is available in some sort of cataloguing system. Production of such metadata currently requires manual annotation by an archivist, a time consuming and hence costly task. The MAD platform is responsible for automating the documentation process as much as possible by employing state of the art algorithms for content analysis and semantic analysis based on human language technologies (HLT) in order to derive metadata. Depending on the level of detail required for the resulting metadata, some human intervention may still be necessary but that is kept to a minimum and the automated processing is still employed as a helping tool even when a human archivist is authoring the metadata.

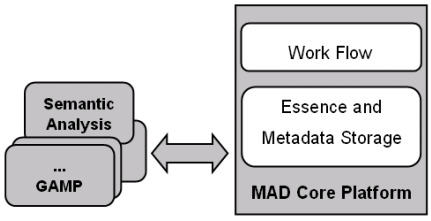

Figure 2: The MAD work-area architecture.

The architectural organisation of the MAD platform is illustrated in Figure 2. The system comprises a core element (the MAD core platform) and a set of configurable Generic Activity MAD Processors (GAMPs). The core platform handles the work and data flow through the system and provides services for storage of the essence and metadata files. The essence is stored as a file containing the digitised version of the audio-visual item. Several other representations such as a low-resolution preview version or separate audio channels or video track can be derived as required by the processes applied. The metadata is stored as XML files using a schema centred on the concept of Editorial Object (EDOB) which can represent either a programme or a unitary section of one. All temporal decompositions of EDOBs such as time-aligned speech transcripts or visual analysis metadata are represented using MPEG7. The storage for metadata files provides versioning support through SourceJammer, a CVS-like, open-source Java system, wrapped up as a web service. This provides some sort of transactional support by allowing rollbacks for failed operations that need to be rerun.

All the processing within the MAD platform is performed by the various GAMPs which implement algorithms for metadata creation or provide services to the other GAMPs such as multimedia de-multiplexing or the generation of automated speech-to-text transcripts. The two main metadata creating GAMPs are the Audio-Visual Content Analysis one which identifies keyframes, scene or shot boundaries and produces other technical metadata and the Semantic Analysis GAMP which generates conceptual metadata starting from the speech transcript or other textual sources available (such as subtitles or closed captions). A web-based interface allows the operator to configure the work-flow and the individual GAMPs as well as to monitor the state of the system at any point and to intervene for solving any problems arising.

3. The semantic analysis process

The Semantic Analysis GAMP uses textual sources such as automatic speech recognition or subtitles or the output of a video OCR process running over the titles or the credits section in order to derive conceptual information about a multimedia item. This metadata can then be used to perform new types of searches within the archives allowing the retrieval of material based on conceptual queries using semantic entities like person names, geographical locations or commercial organisations and the relations between them.

The main challenge that needs to be addressed is posed by the poor quality of the text that results form speech recognition of OCR. The process of automatic extraction of semantic meta-data from text is a complex one and needs good quality text as input in order to provide usable results. While some analysis can be performed on the type of text that is obtained from the multimedia file, the amount of information that can be extracted is rather limited. One way to solve this problem is to try and find better textual sources relating to the multimedia material. One source of good quality text is the Internet and, especially in the

case of news broadcasts, one could hope to find web pages that are closely related to the events mentioned in the media item. This is the hypothesis that our system is based on.

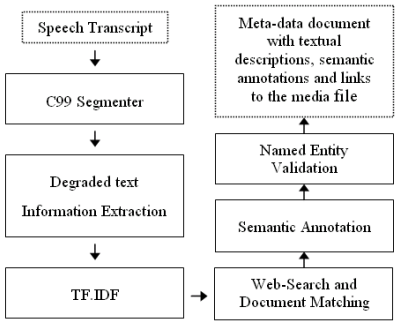

Figure 3: Architecture of the semantic analysis pipeline.

Figure 3 shows the pipeline used for performing semantic analysis. The speech transcript is first segmented using the C99 segmenter (Chaisorn et al., 2003) which uses lexical similarity to decide where a segment split should be inserted. For each of the resulting segments a customised Information Extraction system is used that is capable of extracting some entities from degraded text. Next keyphrases are extracted from each of the segments using the TF.IDF technique. These are then used to perform web searches that find pages that might be related to the content in the multimedia file. From all the candidate pages, the system selects the ones that exhibit a level of similarity with the transcript that is higher than a pre-defined threshold. The pages thus obtained represent high quality text associated with the multimedia item being analysed and can be used to extract semantic meta-data in the form of named entities and relations. This is done using the KIM platform (Popov et al., 2004) that was designed especially for processing web pages.

Once the semantic entities in the related web pages have been detected, a method for merging and assigning confidence scores for these results back in the transcribed text is required. The idea is to augment the entities found in the ASR transcript with the information extracted from the corresponding entities identified by KIM. Firstly, the stemmed entities from the ASR transcription are matched against the stemmed content of the ones in the related web document. If more than half of their content is found among the one of the entities found by KIM, the highest confidence score is assigned to both entities. The semantic information carried by the web entity, is then transferred to the one in the transcript obtaining both temporal and conceptual accuracy. Secondly, the remaining KIM entities are matched against the stemmed content of the ASR transcript and for every match, the semantic content of the KIM entity is transferred to the topical segment containing the text region of the match.

In order to browse and validate the results, a simple user interface has been implemented that can display the media content, the ASR transcript, the links to the related web pages and details about the entities discovered. A screenshot of that interface is presented in Figure 4.

Figure 4 shows the results of running the system over a news broadcast from 2002. One of the leading stories at the time involved a person named “Paul Burrell”. What is interesting about him is that because of the non-standard surname, the speech system failed consistently to recognise the name – despite the fact that it is mentioned 4 times in the story it is never recognised correctly. The semantic analysis system however, manages to get the correct named entity because it is extracted from the web page where it is spelt correctly and is then matched to the partial entity *paul* extracted from the transcript. The interface shows the transcript containing the text “paul bar all” while the entity details pane shows the spelling “Paul Burrell”.

4. Acknowledgements

The research for this paper was conducted as part of the European Union Sixth Framework Program project PrestoSpace (FP6-507336). We would like to thank the BBC archives for providing information about their annotation process and for making broadcast material available to us.

5. References

Chaisorn, T., L.and Chua, C. Koh, Y. Zhao, H. Xu, H. Feng, and Q. Tian, 2003. A two-level multi-modal approach for story segmentation of large news video corpus. In *Proceedings of the TRECVID Conference*.

Figure 4: User interface used to validate results.

Popov, B., A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov, 2004. KIM – Semantic Annotation Platform. *Natural Language Engineering*.

Cross-Document Coreference for Cross-Media Film Indexing

Eleftheria Tomadaki and Andrew Salway

Department of Computing, University of Surrey

Guildford, GU2 7XH, United Kingdom

[e.tomadaki, a.salway]@surrey.ac.uk

Abstract

Potentially, rich representations of film content could be extracted and merged from various texts, such as screenplays, audio description and plot summaries, in order to improve video indexing. As a first step, this requires solving the cross-document coreference (CDCR) task. The CDCR task is difficult in this new scenario because the texts each select and present information about film events very differently; furthermore, the set of possible events is relatively unconstrained. In order to propose new solutions for CDCR we first analysed how two different text types select and present information about the same film events. We present a corpus based analysis of the language used in plot summaries and in audio description, which suggests that while both use similar words to refer to entities, they use very different words to refer to events; there is little systematic relation between the words each use to refer to events. Based on our results, we propose and evaluate four heuristics for the CDCR task that match nouns, functional roles, some verbs, and take into account the number of expected matches according to event aspect. At best we achieved Precision of 49% and Recall of 32% based on 375 CDCR instances between plot summaries and audio descriptions. These figures are low compared to many information retrieval and extraction tasks but we believe that: (i) they may be close to the best possible given the differences between the text types and that they refer to an unconstrained set of events; (ii) they are high enough to start leveraging the information in the texts for video indexing purposes.

1. Introduction

Amongst the mass of multimedia information available today are many artifacts that in some way tell the same story, be it news or fiction. This fact can be exploited by multimedia indexing systems, for example, to generate rich descriptions of semantic video content from texts related to the video. An interesting and challenging scenario is film and the wide range of texts describing its content, such as novels, screenplays, audio description, plot summaries and reviews. These texts complement each other in providing different kinds of information about the story told by a film. It would be difficult, if not impossible, to extract much of this information directly from video data alone. However, before information extracted from each text type can be combined into a rich representation of semantic video content, it is necessary to solve the *Cross-Document Coreference* (CDCR) task. This is the task of deciding whether two linguistic descriptions refer to the same entity or event (Bagga, 1999).

Stories can be defined as comprising a sequence of causally connected events, where the causal connections typically relate to the goals and beliefs of characters (Bordwell and Thomson 1997). This paper focuses on CDCR between two interestingly different types of text that tell the stories of films – plot summaries and audio description. Plot summaries are written by film viewers to summarise the major events of the story. Audio description is mostly produced by trained experts to narrate what is happening on screen for visually impaired and blind film viewers. For more about audio description and its use for indexing narrative aspects of film video data, see (Salway and Graham 2003). Each text, along with the film itself, can be considered as a telling of the same story (Chatman 1978): each text selects and presents information about the events of a story differently. These differences mean that taken together the texts could provide richer indices for film video data, however the differences make the task of merging information from the two text types difficult.

2. Cross-Document Coreference for Video Indexing

Video can be indexed by applying a variety of natural language processing techniques on a

range of texts relating to its content. Systems such as WEBSeek (Smith and Chang, 1997), PopEye (Netter, 1998), Informedia (Hauptmann, 2005), REVEAL THIS (Piperidis and Papageorgiou, 2005) and Google Video (2005) between them process HTML tags, closed captions and speech transcriptions. Other systems extract information from multiple texts related to restricted sets of events in a specialist domain. The KAB system processes dance descriptions and interpretations produced by experts (Salway, 1999). In order to index television coverage of soccer matches, the MUMIS system (Kuper et al, 2003) merges information about 31 football events, such as *goal*, *free kick* and *substitution,* from football transcriptions, tickers and news reports. The automatic identification of cross-document coreference is a crucial step in this merging. To date, CDCR algorithms have typically been used as part of cross-document summarisation systems, matching information in news articles. N-gram algorithms match two or more consecutive words, scored according to statistical methods, such as the document term frequency (TF) and the inverse document frequency (IDF), and are used in systems such as News Story Gisting (Doran et al, 2004). The 'event centric' algorithm is part of the MSR – NLP system (Vanderwende et al, 2004) and suggests a verb centred method, by matching the verb plus another word in a functional role – subject or object. The Boosting algorithm (Zhang et al, 2003) matches open class words and their synonyms, according to WordNet.

3. Corpus-based Analyses of Audio Description and Plot Summaries

Our initial observations about why plot summaries and audio descriptions are written suggest that they will select and present information about film events very differently. The corpus analysis reported here investigated: (i) the frequent open-class words in corpora of plot summaries and audio description (Section 3.1); (ii) whether there are regularities between the words used to refer to events in plot summaries and audio description (Section 3.2). We gathered a corpus of audio description scripts for 45 films, totalling 356,394 words spread across 9 categories and a corpus of 111 plot summaries spread across the same categories, totalling 13,761 words. For more information about the corpora and how they were gathered, see (Tomadaki 2006).

3.1. Frequent Open-Class Words

An analysis of the 100 most frequent words in the corpora shows that both contain an unusually high number of frequent open-class words (nouns, verbs, adverbs and adjectives). The audio description corpus includes 41 open class words in the top 100 and the plot summary corpus includes 27; note, only 6 appear in the in the top 100 words of the general language, British National Corpus, see the list in (Kilgarrif 2002).

Kinds of words	**Audio description frequent OCW**	**Plot summary frequent OCW**
Words that refer to characters, human body	man, head, eyes, hand, face, hands, Tom, men, John, woman	man, family, wife, men, father, son, woman, Harry, Tom
Words that refer to events or states	looks, turns, takes, walks, sits, stands, open, pulls, smiles, stares, goes, look, puts, steps, watches, opens, runs, stops, go	get, help, love, finds, discovers, named
Words that refer to objects and space	room, car, side, table, window, bed, door, way	-
Colours	white, black, red	-
Words that refer to time	-	now, time, new, years, day, young
Miscellaneous	water	life, world, way, war, story, earth

Table 1: Open class words (OCW) in the top 100 words of the corpora of audio description and plot summaries (the lists were not lemmatised). Underlined words are common to the top 100 of both corpora.

The corpora are similar in the words referring to characters words - *man, woman, men,* Table 1. The rest of the frequent open class words are different, implying significant differences between the kinds of information given by both texts. In audio description frequent words refer to the human body (*head, eyes, hand, face*), objects (*door, window, table*) and colours. In plot summaries, frequent words refer to time

(*day, now*). Most interesting to us, the words referring to events are completely different. This contrast is apparent in a classification of the most frequent verbs in each corpus, according to the types of process (Halliday 1994) that they refer to, (Tomadaki and Salway, 2005), Table 2.

Borrowing terms from a set of generic actions for human movement (Gavrila 1999), audio description includes words referring to *stand-alone actions* (look, stare, watch, turn, walk, go, step) and *interactions with objects*, (open, put). Plot summaries describe processes involving complex sequences of human movement and implying something about characters' goals and beliefs (helps, love, discovers, escape, murder). This contrast can be explained by the functions of each text type: audio description is intended to communicate what is being depicted on-screen whereas a plot summary must condense the essential elements of the story. It is further indicated by the differences in mental processes.

Process	**Verbs in audio description**	**Verbs in plot summaries**
Material	turn, take, walk, sit, stand, open, go, put, step, hold, close, wear, carry, run, fall, lift, throw, kiss, lead, get, give, cross	get, help, find, go, take, meet, become, kill, make, destroy, play, save, come, escape, move, lose, try, murder, die, leave
Relational	be	be, have
Mental	watch, see	love, discover, want, know, see, decide, seem
Verbal	-	tell
Behavioural	look, smile, stare, glance, nod	-

Table 2: The 30 most frequent verbs in the audio description and plot summaries in lemmatised corpora wordlists, categorised according to functional grammar processes (Halliday, 1994)

3.2. Cross-Document Coreference

The contrasts between the verbs used to refer to events in plot summaries and audio description seem to present challenges for the CDCR task; it is not possible to match two linguistic descriptions according to whether they contain the same verb. However, it might be reasonable to think that even if two descriptions of the same event do not use the same verb, they will at least use verbs that share some lexical relation, e.g. synonymy, troponymy, entailment (Fellbaum 1998). Thus we created a data set comprising 355 instances of CDCR between plot summaries and audio description. In summary, our analysis showed, perhaps surprisingly, that there is very little systematic relation between the words used to refer to events in plot summaries and audio description. However, we were able to note some degree of regularity in the number of audio description utterances co-referring with a plot summary clause according to the aspect of the verb included in the plot summary.

The 10 most frequent events were identified in the plot summary corpus after lemmatising the corpus wordlist: *help* (12 instances), *meet* (11), *kill* (10), *bring* (8), *tell* (8), *force* (8), *find* (7), *discover* (7), *love* (6) and *murder* (6). Each plot summary instance, including a clause with one of these words (appearing mostly as verbs and on a very few occasions as nouns), was correlated (by one of the authors) to one or more audio description utterances, which could be consecutive or scattered in different parts of the script. Table 3 shows examples of CDCR relating to plot summary instances of *murder*.

Film	**Plot summary clause**	**Audio description utterance/s**
One Hot summer night	When the businessman is murdered the police naturally eye the woman as the top suspect.	[00.02.44] The driver pulls a gun. [00.02.57] The van driver shoots the middle aged man and he slumps back…
See No Evil, Hear No Evil	A man is murdered.	[00.10.19] The woman pulls a gun. She shoots him
Midnight in the Garden of Good and Evil	*The mysteries surrounding Billy's murder…*	[00.10.33] Billy is lying face down, blood on his back. ... [53.05.58] Jim shoots Billy in the back

Table 3: Example CDCR pairs for *murder*

We analysed the audio description utterances associated with each of the 10 frequent plot summary events to identify words that occurred unusually frequently in the co-referring audio description. Examples we found included, for the plot summary event *kill – body, gun, falls, shoots* and *police*, and for the plot summary event *love – kiss, kisses, gaze, gently, lips*. However, such examples were rare, and could only be detected for some of the very most commonly occurring plot summary events. Unless the corpora were much larger, or the CDCR task was reduced to a small set of common events, then it seems unlikely that a CDCR algorithm can rely on matching verbs either directly or via any kind of lexical relation; there is a large tail of infrequent events in the plot summary corpus – for more details see (Tomadaki 2006). Only in 3.2% of the CDCR instances did the audio description refer to the event expressed in the plot summary with the same word or with a synonym. Instead, sometimes, the audio description describes one or a series of related actions where there is, in common with the plot summary clause, at least one entity in the same functional role or a combination of entities.

One regularity we observed that might be helpful for CDCR solutions is a correspondence between the aspect of a verb – punctual or durative (Comrie 1976) – in the plot summary utterance, and the number of matching plot summary utterances. The events *discover*, *meet*, *bring*, *murder*, *kill*, *find* and *tell* are considered to be punctual and they co-refer with a mean average of 5 audio description utterances. By contrast, the durative *help*, *love* and *force* co-refer with a mean average of 29 audio description utterances. Punctual events usually occur in one part of the film and consequently in one part of the audio description, e.g. a murder usually happens quite quickly in one scene and is thus expressed in a few audio description utterances. 'Durative' events are expressed in multiple parts of the film and dispersed in the audio description, e.g. an event where the characters fall in love can be shown in different scenes throughout a film.

4. Proposed CDCR Solutions

We argue that the characteristics of CDCR in this scenario present new challenges for the task of automatically identifying instances of CDCR, and so previous approaches need to be adapted accordingly. In particular, it seems that we will have to rely on matching words referring to entities and their functional roles, because there is such little correspondence between the words used to refer to film events in audio description and plot summaries. Here we propose and evaluate four heuristics for the CDCR task, geared towards the film scenario.

4.1. Creating a Gold Standard Dataset

To create a gold standard dataset, five volunteers identified 375 CDCR instances between the plot summary and the audio description for the films 'Spiderman' and 'Chocolat'. The resulting data set will be made available on the web, see (Tomadaki 2006).

First they were asked to identify the events expressed by the plot summaries and then identify them in the audio descriptions. The identification of events in the plot summary was straightforward, as the pairwise agreement between the annotators was 90%; some volunteers annotated as events a few sentences including more than one verbs conveying different actions with different participants, while others annotated as events only clauses including one verb, which makes the task more focused. The annotators consolidated their answers, annotating plot summary clauses including one verb. The task is time-consuming and challenging when it comes to the event identification in the audio description, totalling four to six hours. The pairwise agreement between all annotators was quite low, totalling 62% in both films. The answers were quite different in events referred to in multiple utterances or because some utterances referred to more than one event, not all being annotated. All annotators noted that after the first couple of hours the task of annotating the audio description became laborious as they did not allow themselves to have multiple breaks. The annotators have finally concluded that the task was subjective due to the different inferences made by each person. They were then asked to

reassess their annotations, considering annotations detected by the others and deciding whether to include them in their answers. After the data reconciliation the pairwise agreement increased to 95% for both films. For more details see (Tomadaki 2006).

4.2. Heuristics for Cross-Document Coreference

We propose four heuristics for the cross-document coreference task, and evaluate each on the plot summary/audio description data set. For each heuristic we first identify the events in the plot summary, following an algorithm which adds grammatical and functional roles (subject, object), deletes sentences having the verbs *be* and *have* as main verbs (as they normally denote states), resolves pronouns and finally separates clauses giving them a unique identification number. The words are matched in their base form. The parsing was realised using the Connexor tagger (www.connexor.com) and the pronoun resolution using ANNIE in GATE (www.gate.ac.uk). We will show how each heuristic is applied for the first clause referring to an event in the plot summary for the film 'Spiderman': *A rather odd thing happened to the life of nerdy high-school student Peter Parker: after being bitten by a radioactive spider…* Note that the Precision and Recall statistics relate to 375 instances of CDCR between 21 different plot summary events in the two films and their audio descriptions.

The first heuristic concentrates on matching entities, which tend to be characters or occasionally objects and locations:

Heuristic 1: If at least two head nouns in the plot summary clause appear in the audio description utterance (in any form), then MATCH = TRUE, else MATCH = FALSE

In the first plot summary event, the words to be matched according to heuristic 1 are: *Peter Parker / Peter/ Parker + thing +/ life +/ student +/ spider*. Eleven audio description utterances, including at least two head nouns or proper nouns were matched. Only three out of ten matches of the combination *Peter* and *spider* were correct, e.g. *[09.56.00] the spider inches its way down towards Peter*, as the word *spider* referred to other spiders as well as to the radioactive spider which bit *Peter*, e.g. *[03.02.57] Peter, with a spider and web emblazoned on his sweatshirt…* A problem arises when two nouns referring to the same entity are matched in the two texts, e.g. *Peter* and *student*, reducing Precision. Overall, Heuristic 1 achieved Recall of 23.4% and Precision of 30.4%, identifying 83/375 CDCR instances.

Heuristic 2 adds verbs and all the nouns to the keyword list to be matched:

Heuristic 2: If at least two nouns or one noun and one verb in the plot summary clause appear in the audio description utterance, then MATCH = TRUE, else MATCH = FALSE

In our example, the words to be matched are: *thing – occur - life - school - student - Peter Parker / Peter/ Parker – bite – spider*. Heuristic 2 retrieved fifteen utterances in total, including eleven spurious, while the gold standard includes five. Four utterances were matched including keywords such as *Peter, spider* and *bite* and two of them were correct, whereas the rest refer to another event, e.g. *[24.07.10] He zooms in on the Daily Bugle front page: Big Apple dreads Spider bite*.

Both Recall (26.2%) and Precision (32.4%) improve slightly on Heuristic 1.

Heuristic 3 has stricter matching criteria in order to increase Precision, by requiring that a noun appears in the same functional role in both plot summary clause and audio description.

Heuristic 3: If at least one noun in the plot summary clause appears in the audio description utterance in the same functional role AND at least one other noun or a verb in the plot summary clause appears in the audio description utterance then MATCH = TRUE, else = FALSE.

We match words with other words in the same functional role, logical subject/agent or object, following the terms in the Connexor tagger: *[Thing: subj] +/ occur +/ Peter Parker / Peter/ Parker + [spider: agt] +/ bite*. Heuristic 3 detected three matches in event 1 of the film

'Spiderman'. Two out of five utterances have been retrieved, including *spider* in the role of subject and *Peter,* or *Peter* in the role of object and *spider,* e.g. *[10.05.00] the spider bites Peter*. All matches are correct, whereas another two utterances have not been detected, as they either include the words *Peter* and *spider* in different functional roles, or because the pronoun resolution failed in the previous step. Heuristic 3 achieved the lowest Recall (21.5%) but the highest Precision (49.4%).

Heuristic 4 was designed to balance Precision and Recall by combining heuristic 2, and heuristic 3. It checks the event aspect (punctual or durative), according to an index of all plot summary events that we have created, and if punctual it retrieves the 5 highest ranked matches according to the following match score algorithm, if durative it retrieves all candidate matches. Heuristic 4 achieved Recall of 32.9% and Precision of 47.8%. One limiting factor is that not all references to the same entity can be resolved automatically, e.g. *Peter Parker* and *Spiderman,* and *chocolate shop*, *patisserie* and *chocolaterie*. When Heuristic 4 was evaluated after manual entity resolution Recall increased to 46.2% and Precision to 50.1%.

1st: Match according to heuristic 3 with two keywords in the same roles

2nd: Match according to heuristic 1 and 2 with three keywords

3rd: Match according to heuristic 3 with one keyword in the same role and another keyword

4th: Match according to heuristic 1, 2 and 3 with two keywords in utterances appearing within or close to the estimated temporal interval

5th: Match according to heuristic 1 and 2 with two keywords

6th: Match utterances appearing within or close to the estimated temporal interval including one keyword

7th: the rest

5. Conclusions

Though these figures are low compared to many information retrieval and extraction tasks we believe that: (i) they may be close to the best possible given the differences between the text types and that they refer to an uconstrained set of events; (ii) they are high enough to start leveraging the information in the texts for video indexing purposes. Until now, researchers have focussed on CDCR between texts of the same type, or texts referring to a restricted set of events. We showed that with perfect entity resolution we could get both Precision and Recall to around 50%. Further improvements, for a small set of common plot summary events, should be possible by identifying correspondences between individual plot summary words (love, murder, etc) and words commonly used to describe the events in audio description.

6. Acknowledgments

This research was carried out as part of the Television in Words project (TIWO) supported by an Engineering and Physical Sciences Research Council (EPSRC) grant, GR/R67194/01: http://www.computing.surrey.ac.uk/personal/pg/A.Salway/tiwo/TIWO.htm.

The authors thank the members of the TIWO Round Table for sharing their knowledge of audio description and providing samples for our corpus.

7. References

Bagga A. and Baldwin B. (1999). Cross-Document Event Coreference: Annotations, Experiments, and Observations. In: *Proceeding of the ACL99 Workshop on Coreference and its Applications* (pp. 1-8).

Bordwell D. and Thompson M.K. (1997). *Film Art: An Introduction*. New York: McGraw-Hill.

Chatman S. (1978). *Story and Discourse: Narrative Structure in Fiction and Film*. NY: Cornell University Press.

Comrie B. (1976). *Aspect: An Intoduction to the Study of Verbal Aspect and Related Problems*, Cambridge: Cambridge University Press.

Doran W., Stokes N., Newman E., Dunnion J., Carthy J., and Toolan F. (2004). News Story Gisting at University College Dublin. In

Proceedings of Document Understanding Conference 2004, Boston, USA.

Fellbaum C. (1998). *An Electronic Lexical Database*. Cambridge: The MIT Press.

Gavrila D. (1999). The Visual Analysis of Human Movement. *Computer Vision and Image Understanding* 73 (1), (pp. 82-98).

Google (2005). *About Google Video,* http://video.google.com/video_about.html

Halliday M.A.K. (1994). *Introduction to Functional Grammar.* 2nd Edition, London: Edward Arnold.

Hauptmann A. (2005). Lessons for the Future from a Decade of Informedia Video Analysis Research. International Conference on Image and Video Retrieval. *Lecture Notes in Computer Science*, Volume 3568, August 2005, pp. 1-10.

Kilgariff A. (2002). *BNC wordlist*. ftp://ftp.itri.bton.ac.uk/bnc/all.num.o5

Kuper J., Saggion H., Cunningham H., Declerck T., de Jong F., Reidsma D., Wilks C. and Wittenburg P. (2003). Intelligent Multimedia Indexing and Retrieval through Multi-source Information Extraction and Merging. *Proceedings of International Joint Conference of Artificial Intelligence-2003 Workshop on Information Integration on the Web (*IJCAI *03),* pp. 409-414.

Netter K. (1998). 'POP-EYE and OLIVE - Human Language as the Medium for Cross-lingual Multimedia Information Retrieval.' The ELRA Newsletter (European Language Resources Association) November 1998, pp. 5-6.

Pickering M.J. and Rüger S.M. (2003). ANSES: Summarisation of news video. *Lecture Notes in Computer Science* 2728 Springer-Verlag, pp. 425-434.

Piperidis S. and Papageorgiou X. (2005). REVEAL-THIS: Retrieval of Multimedia and Multilingual Content for the Home User in an Information Society. In *Proceedings of the 2nd European Workshop on the Integration of Knowledge,Semantic and Digital Media Technologies*, London, U.K.

Salway A.J. (1998). Video Annotation: The Role of Specialist Text. PhD thesis, University of Surrey.

Salway A.J. and Graham M. (2003). Extracting Information about Emotions in Films. In *Procs. 11th ACM Conference on Multimedia 2003*, 4th-6th Nov. 2003, pp. 299-302.

Smith J.R. and Chang S.F. (1997). Visually Searching the Web for Content. *IEEE Multimedia* July-September, pp. 12-20.

Tomadaki E. (2006). Cross-Document Coreference between Different Types of Collateral Texts for Films. PhD thesis, University of Surrey.

Tomadaki E. and Salway A. (2005). Matching Verb Attributes for Cross-Document Event Coreference. In Erk, Melinger and Schulte im Walde (eds.) *Proceeidings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pp. 127-132.

Vanderwende L., Banko M., and Menezes A. (2004). Event-Centric Summary Generation. In *Proceedings of Document Understanding Conference 2004,* Boston, USA.

Zhang Z., Otterbacher J., and Radev D.R. (2004). Learning Crossdocument Structural Relationships using Boosting. In *Proceedings of Document Understanding Conference 2004,* Boston, USA.

Cross-Media Indexing in Reveal-This System

Murat Yakici, Fabio Crestani

I-Lab
Department of Computer & Information Sciences
University of Strathclyde
26 Richmond Street, Livingstone Tower, Glasgow, UK
{murat.yakici,fabio.crestani}@cis.strath.ac.uk

Abstract

In the tomorrow's world, people should be spending most of their leisure time enjoying the content, not searching for it. There is a lack of an integrated technology that will increase the effective usage of multi-lingual and multimedia digital content. The EU-IST Reveal-This (R-T) project aims at developing a *complete* and *integrated* content programming technology. The proposed system is able to capture, semantically index and categorise multimedia and multilingual digital content, whilst providing search, summarisation and translation functionalities. In order to fulfill these requirements, the project proposes an architectural unit called Cross-Media Indexing Component. The component leverages the individual potential of each indexing information generated by the analyzers of diverse modalities such as speech, text and image. The initial prototype utilises the *Multiple Evidence* approach by establishing links among the modality specific descriptions in order to depict topical similarity in the textual space. This information is then used to enrich the original index and is transformed into an MPEG-7 representation. This paper gives an overview of the project, the component's enrichment approach and its support for retrieval.

1. The Reveal-This project

The main objective of the R-T project is to design, develop and test a complete and integrated infrastructure that will allow the user to store, categorize and retrieve multimedia and multi-lingual digital content across different sources (TV, radio, music, Web), with a view to personalize the user experience with these sources. The system will be used by content providers, to add value to their content, restructure and re-purpose it and offer their subscribers, individual or corporate users a personalised content. For the end users, the system also offers tools for gathering, filtering and categorizing information collected from wide variety of sources in accordance with user preferences. The project integrates a whole range of information access technologies across media and languages.

However, a major challenge lies in developing suitable representations for crossing media [1] and languages in the processes of *retrieval*, *categorization* and *summarization*. Rather than looking into single modalities, individual potential of each modality can be exploited and cross analysed in order to improve the effectiveness of the system. This is the fundamental argument behind the efforts in developing a cross-media indexing system which is the main concern of this paper.

2. Cross-Media Indexing

The process of cross-media indexing consists of building relationships among concepts extracted from different modalities such as image, speech and text analysis which together constitute the multimedia knowledge for an audio-visual segment. Given the use of advanced audio-visual content analysis technologies, a unique and complete description of the topical content can be reconciled.

To this respect, cross-media indexing is an ambitious task. Despite the pre-enrichment of digi-

[1] Here, the term *media* refers to the format in which information on a topic is conveyed by one source (e.g. the audio, the image and the text of some video segment).

tal content with meta-data, there is a set of issues there exists in the process of cross analysis and establishing links between modalities. The task entails dealing with the uncertainty, imprecision and, often, the inconsistency of each module descriptions to provide a more complete indexing that is processable for retrieval and filtering purposes.

In the R-T project, these issues are addressed by the architectural unit called Cross-Media Indexing Component (CMIC). CMIC is a standalone and independent server which implements a standard way of identifying, accessing, manipulating, storing and retrieving links between media. It constitutes the final layer of the Cross-Media Analysis and Indexing Subsystem (CAIS) (see Figure 1).

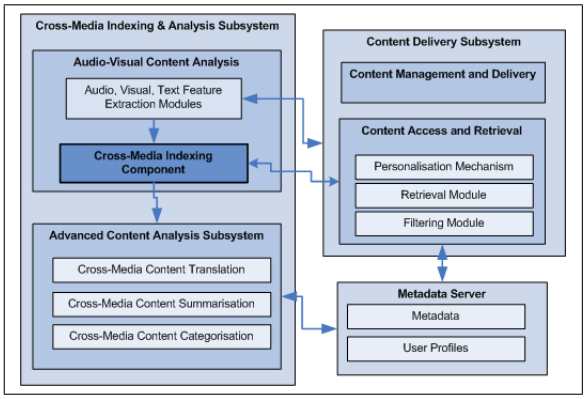

Figure 1: Reveal-This system architecture

CAIS supplies rich and integrated indexing of multimedia content through processes that cross the media borders by exploiting the links between different modalities. The richness of the indexing derives from identification of semantic concepts, entities, facts in speech and text in addition to images. CMIC receives this merged feature set produced by the speech processing component, the text-processing component, the image processing component, the image and the text categorizers after the segments are produced and aligned. Then the system analyses what each modality tells about the content. As a result, CMIC devises a unified view of the content with some measure of uncertainty/confidence attached. The index is transformed into MPEG-7 (Martìnez, 2002) to represent the content. All modality specific descriptions are mapped to their MPEG-7 corresponding elements, however where necessary, MPEG-7's extension mechanism is used to define new descriptors in order to meet the requirements. The output of this unit is consumed by Cross-Lingual Translation and Cross-Media Summarization subsystems.

2.1. Process Model

The indexing task requires merging and processing of digital content streams online and proactively searching for semantic links between features that give evidences to support a topical similarity. CMIC achieves this through several processes which can be generalised into *Analysis* and *Indexing* phases(see Figure 2a). The analysis phase comprises of parsing and transformation of an input stream, noise filtering and lexical analysis tasks (stop word removal, stemming etc.). There are two types of transformations:

- Bypass transform- in which the input features are mapped into an MPEG-7 representation without any further processing.
- Progressive transform- in which the input is processed and a cross-media index is devised. Later, this index is added to the MPEG-7 representation.

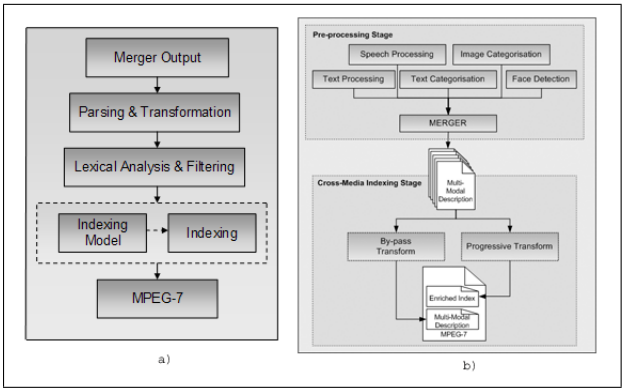

Figure 2: a)CMIC process model and b)transform model

The progressive transform enriches the information; cross links are established between each media description on the same story. This is done by an indexing model and especially, this part is expected to be utilised by the retrieval or filtering engine (See Figure 2b).

2.2. Indexing Models

The indexing approach chosen should avoid, otherwise minimise uncertainty, imprecision and the inconsistency across modalities. CMIC is a tool where various approaches can be deployed and tested to overcome these issues mentioned. From the simplest to more advanced, the current prototype supports several indexing approaches:

- Term frequency: Frequency of a term irrespective of which modality produces it used to give a weight for that specific term. It is assumed that the higher the frequency is, the more important/representative the information conveyed by the term.

- Inverse document frequency (Jones, 1979; Salton, 1989): Inverse document frequency (IDF) is a well known metric in Information Retrieval. It takes the possible misleading consequences of the preceding approach into account by utilising the inter-document term frequency information. This approach gives less importance to a term, if it appears in high number of stories.

- Modality frequency: In this approach, the term frequency provided from each processing module (such as image analysis, text processing etc.) is incorporated to the previous approaches.

- Dempster-Shafer: Also known as Theory of Evidence (Shafer, 1976) is extensively studied in Image retrieval (Jose and Harper, 1997; Aslandogan and Yu, 2000; Jeon et al., 2003), structured document retrieval (Lalmas, 1997; Lalmas, 1998; Ruthven and Lalmas, 2002; Graves and Lalmas, 2002), and web user log analysis (He et al., 2002) research.

2.2.1. Dempster-Shafer in Detail

Rather than formulating our approach as a structured document retrieval problem, we assume that a single segment is independent of the rest of the audio-visual stream irrespective of its later representation in MPEG-7. Thus every modality individually gives support for a single segment, unlike in structured document retrieval approach where evidence in leaf nodes propagates to root. Briefly, Dempster-Shafer combines two or more bodies of evidence defined with in the same *frame of discernment* T into one body of evidence. T contains all the possible hypotheses in the domain of concern. According to the theory, it is possible to assign probabilities to each subset of T. In our approach, T contains all the terms generated by all modalities. This is due to the computational cost where we restrain the frame of discernment by only considering singleton subsets of T. Hence, a document d (segments or stories in our case) containing a term t's existence in one modality m is counted as an *evidence* to support the *topical similarity* hypothesis. Each processing module is treated as a probability density function also called as *Source of Evidence or Base Probability Assignment (BPA)*. BPA contains a set of normalised term frequencies summing to a confidence for each modality m.

$$m(\emptyset)_d = 0 \text{ and } \quad 1 = \sum_{t \in d} m(\{t\}) \ .$$

where

$$\mathrm{m}_d(\{t\}) = \begin{cases} tf(d,t).\log_N(\frac{N}{n(t)}) & \text{if} \quad t \in d, \\ 0 & \text{otherwise.} \end{cases}$$

$$m_d(T) = 1 - \sum_{t \in d} m_d(\{t\}) \ .$$

$tf(d,t)$ stands for the term t's frequency in document d. $\log_N(\frac{N}{n(t)})$ is a variant of inverse document frequency where N represents the number of documents in the collection and $n(t)$ represents the number of documents that contain t. Normalisation can be a problem in some circumstances when applying Dempster-Shafer. However, the preceding formula gracefully overcomes this and adheres to the theory where $\sum_{t \in d} m_d(\{t\}) \leq 1$. $m(T)$ is equal to the unassigned BPA to any preposition set or the hypothetical uncertainty. In order to combine the evidences, we adapted the following combination rule (Lalmas and Moutogianni, 2000):

$$m(\{t\}) = m_1 \otimes m_2(\{t\}) = \frac{1}{K}(m_1(T).m_2(T))$$

where

$$\mathrm{K} = \left(\sum_t m_1(\{t\}).m_2(\{t\}) \right) + m_1(T).m_2(T) \ .$$

such that $m_1(\{t\}) > 0 \; and \; m_2(\{t\}) > 0$ conditions are satisfied.

2.3. Related Work

Cross-media indexing and retrieval is a product the recent advances in each speech, image and text retrieval research. In multimedia retrieval research, it has been long studied that, a single modality can be analyzed and indexed in a certain way which would effectively increase retrieval system's performance [2].

The research in speech retrieval (Allan, 2001; Crestani, 2003a; Crestani, 2002) has shown that, even under high levels of word recognition error rates, it is possible to get reasonable retrieval performance using classical information retrieval techniques. In (Singhal and Pereira, 1999), the authors focus on expanding documents to alleviate the effect of transcription mistakes on speech retrieval. Discussion of the effects of out of vocabulary items in spoken document retrieval is given in (Woodland et al., 2000). Experiments on TREC-8 (1998) audio collection using various retrieval setups suggests that it is possible to demonstrate moderate retrieval performance when advanced IR techniques (a combination of Query and Document expansion method) are used to compensate for recognition errors caused by out of vocabulary words. Experiments include query expansion and document expansion against a baseline retrieval system that uses Okapi variant.

More recently, research has been moving forward to cross modal analysis of multimedia. Recent advances in the image retrieval report that text (in the form of annotations or speech transcripts) and image (histogram, texture, colour frequencies etc.) can be combined successfully (Barnard et al., 2003; de Vries et al., 2004; Westerveld, 2004). In (Barnard et al., 2003), the authors suggest an approach for predicting words that are associated to whole images and corresponding to particular image regions. The process is regarded as a problem of translation, such as translation of image regions to words. Using multi-modal and correspondence extensions to Hofmann's (Hofmann, 1998) hierarchical clustering/aspect model (which is a model adapted from statistical machine translation and a multi-modal extension to Latent Dirichlet allocation) is proposed. In (Duygulu et al., 2003), Duygulu extends the previous work and studies videos rather than single images. This is a difficult problem since, text is not associated with a single frame. The authors report reasonable success over TREC2001 collection. In another study (Adams et al., 2003), semantic labeling is formulated as a machine learning problem. Concept representations are modeled using Gaussian mixture models(GMM), Hidden Markov Models(HMM), and support vector machines(SVM) and were evaluated against TrecVID(2001) corpus. Study reports that fusion scheme achieves more than %10 relative improvement over the best single modal concept detectors.

[2] An up-to-date survey of the research fields can be found in (Snoek and Worring, 2005)

3. Evaluation

In order to validate our arguments and thus the first prototype, we intent to evaluate the performance of different cross-media indexing models. This will enable us to explore the pros and cons of various indexing models for the combination of evidence. In this context, we are currently pursuing two different strategies *known item search* and *task oriented user test*, both involving the construction of a test collection using real users which is still in progress. The collection we are using for both evaluation strategies is three-hour multimedia collection. The collection covers politics, travel and news domains in English, French and Greek languages.

The availability of this test collection will enable to test not only CMIC, but also the reliability of the single modality indexing modules. CMIC will make use of the results of the evaluation of the single modality indexing module for weighting the reliability of the evidence provided by each of them. It should also be noted that the final evaluation of the R-T system will be carried out using a user and task oriented evaluation involving home user and TV broadcast professionals.

4. Conclusions and Future Work

In this paper, we gave an overview of a work in progress which is part of the EU-IST Reveal-This project. Our approach to cross-media indexing

was presented. We consider that the utilisation of different advanced technologies of audio-visual content analysis and indexing by multiple evidences approach can be employed to robustly reconcile a unique and complete description of the topical content. The hypothesis of cross-media indexing and the models in use are still an open research area. We will be considering implementation and evaluation toward more complicated approaches such as Bayesian Networks, Kernel Canonical Analysis, Gaussian Mixture models.

5. References

Adams, W. H., G. Iyengar, C.Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith, 2003. Semantic indexing of multimedia content using visual, audio, and text cues. *Journal on Applied Signal Processing*, 2003(2):170185.

Allan, J., 2001. Perspectives on information retrieval and speech. In *SIGIR Workshop: Information Retrieval Techniques for Speech Applications*. Springer.

Aslandogan, Y. A. and C. T. Yu, 2000. Multiple evidence combination in image retrieval: Diogenes searches for people on the web. In *Proceedings of the ACM SIGIR*. Athens, Greece: ACM Press.

Barnard, K., P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan, 2003. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

Crestani, F., 2002. Spoken query processing for interactive information retrieval. *Data and Knowledge Engineering*, 41(1):105–124.

Crestani, F., 2003a. Combination of similarity measures for effective spoken document retrieval. *Journal of Information Science*, 29(2):87–96.

de Vries, A. P., T. Westerveld, and T. Ianeva, 2004. Combining multiple representations on the trecvid search task. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*.

Duygulu, P., D. Ng, N. Papernick, and H. Wactlar, 2003. Linking visual and textual data on video. In *Workshop on Multimedia Contents in Digital Libraries*. Crete, Greece.

Graves, A. and M. Lalmas, 2002. Video retrieval using an mpeg-7 based inference network. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. Tampere, Finland: ACM Press.

He, D., A. Göker, and D. J. Harper, 2002. Combining evidence for automatic web session identification. *Inf. Process. Manage.*, 38(5):727–742.

Hofmann, T., 1998. Learning and representing topic. a hierarchical mixture model for word occurrence in document databases. In *Workshop on learning from text and the web*. CMU.

Jeon, J., V. Lavrenko, and R. Manmatha, 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. Toronto, Canada: ACM Press.

Jones, K. Spärck, 1979. Experiments in relevance weighting of search terms. *Information Processing and Management*, 15:133144.

Jose, J. M. and D. J. Harper, 1997. A retrieval mechanism for semi-structured photographic collections. In *Proceedings of the DEXA*. Springer-Verlag.

Lalmas, M., 1997. Dempster-shafer's theory of evidence applied to structured documents: modelling uncertainty. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*. Philadelphia, Pennsylvania, United States: ACM Press.

Lalmas, M., 1998. Information retrieval and dempster-shafer's theory of evidence. In *Applications of Uncertainty Formalisms*. Springer-Verlag.

Lalmas, M. and E. Moutogianni, 2000. A dempster-shafer indexing for the focussed retrieval of a hierarchically structured document space: Implementation and experiments on a web museum collection. In *Proceedings of RIAO*. Paris, France.

Martìnez, J. M., 2002. Mpeg-7: Overview of mpeg-7 description tools. *IEEE Multimedia*, 9(3):83–93.

Ruthven, I. and M. Lalmas, 2002. Using dempster-shafer's theory of evidence to combine aspects of information use. *Journal of In-*

telligent Information Systems, 19(3):267–301.

Salton, G., 1989. *Automatic Text Processing*. Massachusetts: Addison Wesley.

Shafer, G., 1976. *A Mathematical Theory of Evidence*. Princeton University Press.

Singhal, A. and F. Pereira, 1999. Document expansion for speech retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. Berkeley, California, United States: ACM Press.

Snoek, C.G.M. and M. Worring, 2005. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35.

Westerveld, T., 2004. *Using generative probabilistic models for multimedia retrieval*. Ph.D. thesis.

Woodland, P. C., S. E. Johnson, P. Jourlin, and K. Spärck Jones, 2000. Effects of out of vocabulary words in spoken document retrieval (poster session). In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. Athens, Greece: ACM Press.

The iFinder Audio-Visual Indexing Framework for Cross Media Applications

Joachim Koehler

Fraunhofer IMK
Schloss Birlinghoven, 53754 St. Augustin, Germany
Joachim.Koehler@imk.fraunhofer.de

Abstract

This paper describes the multimedia indexing system iFinder and its usage in several research and development projects and applications. The iFinder system is a development of the Fraunhofer IMK and is a result of several national and European research projects and internal development activities. The main idea of iFinder is to integrate different multimedia extraction methods for the automatic generation of metadata of audio-visual content and to support international meta data standards, like MPEG-7. The extraction methods are based on statistical pattern recognition methods and require sufficient training material. iFinder supports different media types, like speech and video, and can process each single medium separately from each other. However, the full benefit of iFinder is the combination of audio and video processing algorithms. The fusion of metadata from different modalities is still a research challenge, because in many applications the streams are more or less independent from each other. The iFinder system is mainly a result of two already completed national research projects. In the project AGMA the video recordings of the German Parliament are indexed and analyzed. In the project Piavida broadcast data from the Deutsche Welle is segmented and transcribed automatically. In new IST projects of the 6th IST framework iFinder is applied to index and annotate multimedia content for cross media applicatons. The paper describes the methods, approaches and usage of the iFinder technology in detail.

1. Introduction

Research tools and solutions for media analysis and media retrieval are currently gaining increasing influence. In almost all IST projects of the action line for Semantic based knowledge and content systems activities on multimedia indexing are included. The challenge to generate metadata directly from the audio-visual sources and to create cross-media applications is still high. Especially when multimedia retrieval applications should move from research systems to industrial applications additional requirements on robustness and system performance have to be considered. Much research work has investigated methods to carry out the indexing process on low-level audio/visual features. However, for many cross media application scenarios the need for high or at least medium feature extraction methods is obvious.

To provide a framework to index and retrieve multimedia documents the iFinder system is developed. With the realisation of two national German research projects and the effort to standardize multimedia metadata descriptions in MPEG research work was carried out to investigate methods for automatic metadata generation and to integrate them in a media indexing framework. After the finalization of the national research projects AGMA and Piavida, which are described later in the paper, additional work has been invested to integrate the algorithms and methods into one framework with a common application interface. The core technologies rest on pattern recognition methods which are adapted to index multimedia data.

The iFinder framework consists of a toolbox to index multimedia data sources and a retrieval engine to build distributed media retrieval applications. The following figure shows the components and processing steps in the iFinder environment.

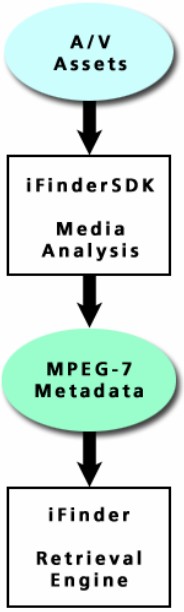

Figure 1: iFinderSDK and iFinder

The audio visual media assets are indexed by the different modules of the iFinderSDK. As output of the metadata generation process a MPEG-7 XML description is stored in a file. Based on this metadata description a retrieval engine is built.

The paper is organized as follows. In section 2 the iFinderSDK is presented including the different media indexing methods. Section 3 shows the principles of the iFinder retrieval environment. The usage of iFinder technology in different research and development projects is given in section 4.

2. The iFinderSDK Media Indexing Toolbox

2.1. Principles and Architecture of iFinderSDK

The iFinderSDK is a collection of media indexing methods with an open application programming interface [LOEAEM04], [LOEAES04]. Other applications can include the software libraries for the operating systems Windows and Linux. The interfaces are defined in C++ header files and all media indexing modules have the same class function calls. First, the *init* procedure starts the initialisation process and loads the media data into the media extraction modules. The *process* function performs the features extraction. Finally the *exit* function releases the allocated computational resources. Each module can be configured by an XML based parameter file which also contains the information of the knowledge sources (e.g. acoustic models of the speech recognition module). The architecture of the iFinderSDK is shown in the following figure. For several routines open source SW is applied, like the Xerces XML. The XML parser processes the MPEG-7 output of the features extraction modules which can be used as input for other processing modules or finally for the export to an XML database.

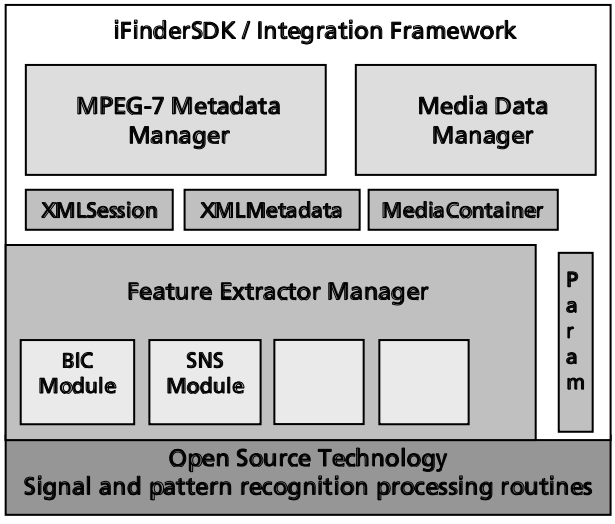

Figure 2: Sketch of the iFinderSDK architecture

To set up the workflow desired the developer only needs to edit the parameter file defining all modules and the order in which they should be applied to the input media data. When starting the application the so-called Feature-ExtractorManager will then instantiate and deploy the objects of the feature extractors as they are specified in the parameter file. In this way the implementation effort is reduced to a minimum.

2.2. Speech/Audio Analyses Modules

The iFinderSDK contains several modules to carry out a feature extraction and analysis of audio signals. Here a list and a short description of the indexing tools are given.

2.2.1. Speech Recognition

For automatic speech recognition the ISIP speech toolkit is used and adapted for the needs of media retrieval applications [PICISI98]. The main issue to adapt the system is the training of the knowledge sources, like the acoustic models, language models and the definition of the lexicon. The core system of the decoder is not modified and it can be assumed that most of the available speech recognition systems are based on very similar technologies, like HMM-modelling, Viterbi-decoding and MFCC feature extraction. Hence, the focus of adapting a speech recognition system is to provide domain dependent databases to train the knowledge bases. Each of the media retrieval projects provides specific databases and requirements to achieve an optimal recognition process. A variety of different vocabulary sizes are supported in the projects and ranges from keyword recognition to medium size vocabulary for alignment tasks to large vocabulary recognition tasks with up to 100K lexicon entries. A special attention is given to subword unit recognition tasks, like syllable based recognition [LAREUR03]. Fraunhofer IMK has contributed a syllable based recognition and retrieval strategy to the MPEG-7 standardization effort [MPE7AUD01].

2.2.2. Speech/Non-Speech Detection

To filter out the parts of the audio signal which contains pure speech or speech mixed with background music, a speech/non-speech detection module is developed. First, the speech frames are classified into the given classes, like speech, music or other segment classes. Each frame gets an unique label. Second, the labelled frames are grouped into homogeneous groups of labels to smooth misclassified frames. For the first processing step 12 different features are used, like zero crossing rate or average energy in a frame. The label decision is performed by a multivariate Gaussian classifier which is trained before on hand labelled training samples. In the second stage of the processing an optimal segmentation algorithms is applied using dynamic programming techniques which find the optimal boundaries for homogeneous segments [BIAACM02].

2.2.3. Speaker Segmentation

This speaker segmentation method is based on the Bayesian Information Criterion approach (BIC) [TRIEUR99]. The main idea of this algorithm is to estimate three Gaussian distributions of the preprocessed speech frames. Therefore the speech signal is divided in frames of 20 ms duration with an overlap of 10 ms. Each frame is transformed from the time domain into the mel frequency domain by computing the mel frequency cepstral coefficients (MFCC) which are well known from the area of speech recognition. The observation window for a speaker change is increased frame by frame. One Gaussian distribution (probability density function PDF) is calculated from the left part (PDF-Left) of the observation window, the other Gaussian distribution from the right part of the observation window (PDF-Right). For each possible speaker change point (i.e. frame), which defines the left and the right observation window the two Gaussian distributions are calculated. Additionally a third Gaussian distribution is estimated for the whole observation window (PDF-Whole). The idea of the BIC algorithm is now to determine whether it is more likely that the observation window is modeled by one Gaussian distribution (PDF-Whole) or by two single Gaussians, namely PDF-Right and PDF-Left. If the BIC-threshold is below zero it is assumed that a speaker change occurred. As soon as a speaker change point is found the calculation restarts as described before. The BIC speaker change detection delivers important information to generate a structure of an audio document.

2.2.4. Speaker Clustering and Jingle Detection

After the speaker segmentation process based on the BIC method a clustering procedure is developed which groups similar segments. Every cluster is attached with a speaker identification label. To group the speaker segments a global similarity measure is calculated. Details of the speaker clustering methods and the clustering performance are given in [LARSPE05].
Also other audio segmentation and identification methods are included in the iFinderSDK. A jingle detection module is able to create an audio fingerprint of a audio segment. This fingerprint can be used to identify known material (e.g. jingles) in a large audio collection.

2.3. Video Analyses Modules

The iFinderSDK also includes methods to analyse and index image and video data.

2.3.1. Cut Detection

For the segmentation of videos into segments a video cut detection algorithm is developed. Based on frame based differences of a colour histogram a decision is made, if a scene change has occurred.

2.3.2. Face Detection and Recognition

For the face detection three processing steps are carried out. First, the sub images are computed and these sub images are transformed by a robust feature extraction method. In the second step the features of the sub images are sent to the input of a trained neural network which decides if the sub image contains a face. Finally, the different face hypotheses are merged.
The face recognition is based on a 2 dimensonal Hidden-Markov-Modell. For every face a reference model is trained with several examples of a specific face. The decoding is realized with a Viterbi algorithm finding the optimal path between the test pattern and the reference model.

3. The iFinder Media Asset Management Framework

The iFinder multimedia retrieval system builds the application on top of the media indexing toolbox iFinderSDK. The iFinder search engine enables to search in large audio/video archives. The MPEG-7 metadata descriptions are stored in either in XML or relational databases (e.g. Oracle). The iFinder system includes also components for video streaming and Web services. The system architecture is based on open standards and allows the integration in existing IT infrastructures and multimedia processing workflows. The client application is written in JAVA and support a direct user interaction with the retrieved content. The usage of the iFinder retrieval application consists of three steps. First the user defines a query input. This is usually a definition of a search term (e.g. person name), a keyword or the date of the media production. The GUI for the input of query parameters of the iFinder client in the application for the German Parliament video search is shown in figure 3.

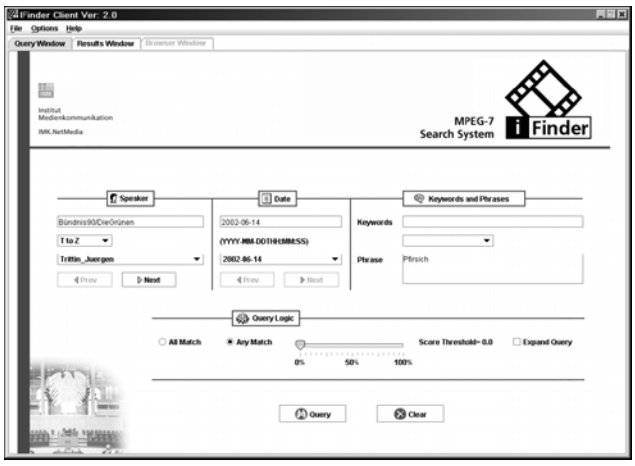

Figure 3: iFinder client window to define input queries

This query is send to the retrieval application which presents a list of possible hits. All possible result items are attached with additional metadata to give the user a first overview about the different search results. Finally the user selects one of the search results to get the full multimedia document. With the Java client application including a text/video browser the user can navigate and browse through the search result.

4. Cross media applications with iFinder

The development of the iFinder framework is based on several research and development projects. All projects have the common goal to structure and index audio-visual unstructured data material for later retrieval, search and navigation applications. Although the projects differ in the composition of different media types and the search functionalities, the general idea to realize an automatic indexing functionality is the same. Because the indexing methods are based on statistical pattern recognition methods the availability of speech, audio and video resources and data material are very important to achieve an optimal indexing and retrieval rate. In most cases the data material can be collected and stored easily. The main effort is to annotate and prepare sufficient training and evaluation material.

4.1. BMBF Project: AGMA

The national BMBF (Federal Ministry of Education and Research) project AGMA (Automatic Generation of Metadata based on MPEG-7) started in November 1999 and last until October 2003 [KOECAS01]. The main objective of AGMA was to investigate multimedia indexing methods in context of the upcoming MPEG-7 standard. As application domain the speech and video recordings of the German Parliament are used. The data is directly recorded from the DVB signal of the German broadcaster Phoenix TV in MPEG-2 quality. A typical session in the German Parliament has an average duration of six hours. More than 100 hours of audio-visual data are collected and a few hours are transcribed on a detailed level. The domain of the parliament has several advantages. First, the German parliament is a naturally delimited domain. Domain limitation is an important aspect to build a system that can be used for real world applications. Second, the high volume of speech and video with the corresponding transcriptions and annotations necessary for system training is available. Third, the data is publicly available, and its use does not give rise to intellectual property issues.
Another speciality of this domain is the availability of the stenographic transcriptions. They are intensively used to develop an automatic alignment procedure using a HMM based state-of-the-art speech recognition system with specialized grammars and language models [BIALRE03]. The aligned speech data is evaluated if the reliability of the transcriptions are high enough to use for later training and adaptation processes. Therefore the a-posterior probabilities are estimated and chunks of a pre-defined size are extracted automatically. With the speech alignment methods it is possible to generate an automatic synchronisation of the audio-video signal with the corresponding stenographic text. For every spoken word a time position is generated [BIATSD03].
The video analysis is composed from three successive processing steps: temporal segmentation, face detection and tracking and face recognition. The temporal segmentation uses the difference image and the difference of the histograms of two consecutive frames to find the scene boundaries (cuts and dissolves) in the video sequence. A frontal face detector is used to find the faces in the scenes. The gaps between the detected frontal views of the face, where the person looks to the left or right, are filled using a simple histogram based tracking algorithm. The recognition of the detected faces is based on pseudo 2D Hidden Markov Models. To build face models for the politicians 50 persons are annotated manually. With the face database a training and evaluation is carried out. Also other video processing algorithms are investigated and developed, like object segmentation and clustering of different scene types.
The generated metadata descriptions from the speech and video analysis are merged into a MPEG-7 description. The following diagram shows the data flow of the metadata storage.

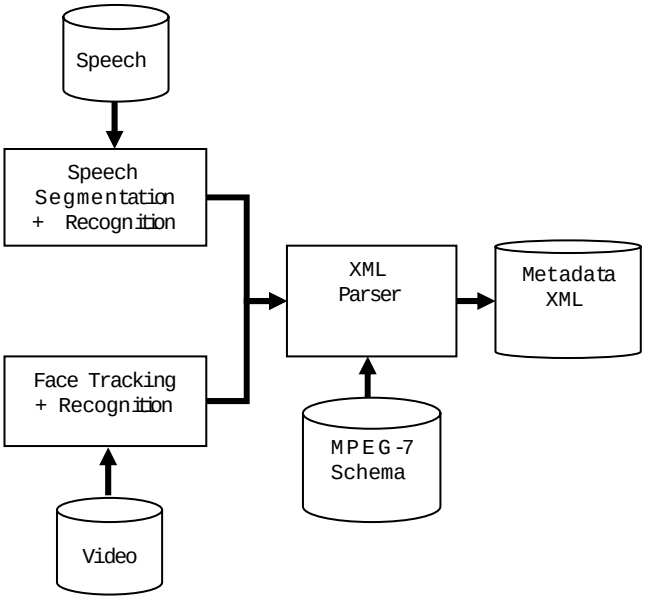

Figure 4: Content extraction architecture in AGMA

The retrieval client application is realized as JAVA GUI program with a user-friendly search and browsing functionality. The text-video browser is shown in figure 5.

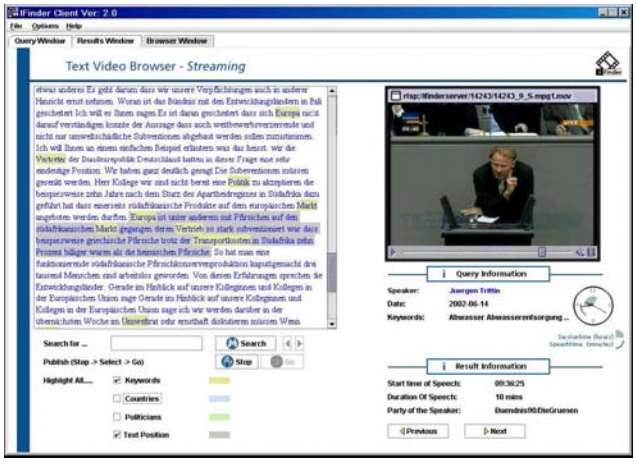

Figure 5: Text/Video Browser of the AGMA project

The text window contains the stenographic transcriptions. The search results are high lighted on the sentence level. Also named entities are marked automatically to give the user a fast overview about the thematic topic. The video signal is streamed using the Quicktime streaming server. The video and the text are synchronised and the user can browse through the video or the text and the positions of the different synchronized media types are changed simultaneously. This text/video browser is one the most advanced cross media applications because the synchronisation is realized automatically.
This audio and video analysis methods and the retrieval application build the basics of the iFinder system. More information can be found in [LOEACM02].

BMBF Project: Piavida

In the project Personalized Interactive Audio, Voice and Video Information Portals and Applications (Piavida) six Fraunhofer institutes developed a system to extract relevant information automatically from large audio-visual corpora. This project ended in December 2003 after a project duration of 3 years. The Fraunhofer IMK has investigates methods for spoken document retrieval on subword unit, like syllables and phonemes [EICNOE02]. Based on the automatically generated transcriptions a text classification module analyzes the speech documents.
As speech training and retrieval database a speech collection from the German broadcaster Deutsche Welle is used [EICLRE02]. Also here the training material is aligned automatically based on the calculation of a-posterior probabilities. The transcriptions are manually evaluated and corrected with the Transcriber Toolkit. The training set of 6 hours of speech is used to build cross-word triphone models. These models are still used in the iFinder framework.
The project has shown the importance of large multimedia databases for training and evaluation purposes.

4.2. Audiomining for broadcasters

In this project the iFinder system is adapted to build a pilot speech and audiomining application for two German broadcast stations [LARGIW05]. The media documentary departments of broadcast stations have the challenge to archive and describe huge amounts of the radio program every day. With the support of an automatic indexing system this work effort can be reduced and radio material which could not be documented can be described automatically using an audiomining system. The term audiomining reflects the similarity to available text mining solutions which are already on the market. To realize a solution to index the radio program of broadcasters the iFinder system is integrated in a complete application with a user-friendly web-based frontend. This pilot system includes methods for speech/nonspeech detection, speaker segmentation and speaker clustering. The speaker algorithms generated for each radio show an identification number for each speaker cluster. The speech recognition process is carried out with a syllable based speech recognizer integrated in the iFinderSDK. The syllable transcriptions are stored in a file system and the related search module translated the word and phrase based queries to a syllable and a phonetic transcription which is compared to the recognized transcripts. This approach has the advantage to be vocabulary independent and to be very robust against variations of the domains.
The speech database for this pilot application consists of 160 hours of recordings from the broadcast stations. The DW provided two different radio formats which have a high speech portion. The radio shows from WDR has much more music included and one of the shows has a more dialogue based structure. The recordings are available in the MPEG-1 layer II format and are re-sampled to linear 16 KHz sampling rate.

4.3. EU-IST Project Boemie

The objective of the IST project Boemie is to investigate the creation, evaluation and exploitation of multimedia ontologies [SPYEWI05]. The project has started recently in March 2006 and the first requirements of the projects are defined. One important component of this research project is the automatic extraction of audio-visual content. Therefore the iFinder framework should be used, adapted and extended. As application domain kind of sports are already defined and the content collection phase will be started in the next phase of the project. This means that iFinder will be used in

another domain (i.e. sports domain) which requires new or at least modified extraction algorithms. Also the aspect of modelling multimedia knowledge will be an important aspect of future multimedia retrieval systems, like iFinder. Here the challenge is to apply and exploit approaches from ontology based technologies, like reasoning methods, to multimedia retrieval applications.

4.4. EU-IST Project LIVE

In the research project LIVE of the IST program on knowledge modelling and intelligent content processing the objective is to investigate new formats of interactive TV productions for live sport content [WILEWI05]. The central idea is to provide the viewers of TV different channels for a sport event which are linked with each other, so that the viewer will have the opportunity to consume the sport event to his personal interest. To achieve this goal a semi-automatic authoring of the content must be achieved. Therefore the production and availability of metadata from live and archived content are very important. The iFinder framework should be adapted to generate metadata, like keywords, in real-time. Further the archive of the ORF will be indexed by the iFinder solution.

5. Summary and Outlook

It has be shown that as result of several research projects an integrated framework for multimedia retrieval, called iFinder, is developed. It already covers many different modules for media indexing and provides retrieval functionality. The main issue for industrial projects is to achieve a robust system performance regarding the required recognition and retrieval rates. Therefore domain dependent data collections have to be carried out to adapt and optimize the indexing modules. In further research projects new features, like knowledge modelling and online processing, will be extend the iFinder framework.

6. References

[BIAACM02] K. Biatov, J. Köhler: An audio stream classification and optimal segmentation for multimedia application, Proceedings of the 11th ACM International Conference on Multimedia, Berkeley, CA, USA, Nov. 2003

[BIALRE02] K. Biatov, J. Köhler: Methods and Tools for Speech Data Acquisition exploiting a Database of German Parliamentary Speeches and Transcripts from the Internet, LREC-2002, Las Palmas, Spain, June 2002

[BIATSD03] K. Biatov: Large Text and Audio Data Alignment for Multimedia Applications, TSD 2003, Proceedings: Lecture Notes in Computer Science 2807 Springer, ISBN 3-540-20024-X, pp. 349-356, 2003.

[EICNOE02] S. Eickeler, K. Biatov, M. Larson J. Köhler*:* Two novel applications of Speech Recognition methods for robust Spoken Document Retrieval, DELOS Network of Excellence Workshop, Crete, 2002

[EICLRE02] S. Eickeler, M. Larson, W. Rüter, J. Köhler*:* Creation of an Annotated German Broadcast Speech Database for Spoken Document Retrieval, LREC-2002, Las Palmas, Spain, June 2002

[KOECAS01] J. Köhler, K. Biatov, M. Larson und C. Eckes: AGMA: Automatic Generation of Meta data for Audio-Visual Content in the Context of MPEG-7, CAST01, Sankt Augustin, 2001

[MP7AUD01] MPEG-7 Audio – ISO/IES CD 15934-4 Multimedia Content Description Interface – Part 4: Audio, 2001

[LAREUR03] M. Larson and S. Eickeler: Using syllable-based indexing features and language models to improve german spoken document retrieval. In European Conference on Speech Communication and Technology, Geneva, Switzerland, 2003

[LARGIW05] M. Larson, T. Beckers, V. Schlögel: Structuring and Indexing Digital Archives of Radio Broadcasters, GI Workshop on Digital Media Archives, Bonn, 2005

[LARSPE05] M. Larson, K. Biatov: Speaker Clustering via Bayesian Information Criterion using a Global Similarity Constraint, SPECOM, 2005

[LOEACM02] J. Löffler, K. Biatov, C. Eckes, J. Köhler: iFinder: An MPEG-7-based Retrieval System for Distributed Multimedia Content, ACM Multimedia Conference 2002, 431-435, 2002.

[LOEAEM04] J. Löffler, K. Biatov, J. Köhler: Automatic Extraction of MPEG-7 Audio Metadata Using the Media Asset Management System IFinder, 25th International AES Conference, London, June 2004

[LOEAES04] J. Löffler, J. Köhler, H. Blohmer, K. Kaup: Archiving of Radio Broadcast Data using Automatic Metadata Generation Methods within MediaFabric Framework, 116th AES convention, Berlin, May 2004

[PICISI98] J. Picone et al.: A public domain decoder for large vocabulary conversational speech recognition, Mississippi State University, 1998

[SPYEWI05] C.D. Spyropoulos, G. Paliouras, V. Karkaletsis: BOEMIE: Bootstrapping ontology evolution with multimedia information extraction, EWIMT 2005, London

[TRIEUR99] A. Tritschler, R. Gopinath: Improved Speaker Segmentation and Segments Clustering Using the Bayesian Information Criterion. In Proc. EUROSPEECH, vol 2, 261–264, 1999.

[WILEWI05] C. Mac Williams, R. Wages: Virtual personalised channels: video conducting of future TV broadcasting, EWIMT 2005, London

Cross Media Aspects in the Areas of Media Monitoring and Content Production

Herwig Rehatschek, Michael Hausenblas, Georg Thallinger, Werner Haas

JOANNEUM RESEARCH, Institute of Information Systems & Information Management
Steyrergasse 17, A-8010 Graz, Austria
{Herwig.Rehatschek, Michael.Hausenblas, Georg.Thallinger, Werner.Haas}@joanneum.at

Abstract

Cross media tools and multi-modal analysis are crucial technologies for automizing media monitoring and advancing content production. We discuss relevant project results from the projects DIRECT-INFO and NM2 where cross media tools and multi-modal analysis were developed and applied.

1. Introduction

Media monitoring, and specifically global advertisement expenditure measurement, is a huge world-wide market. In 2005 the global advertising market was more than 400 billion US $ [Zenith Optimedia, (2006)]. Main goal of media monitoring companies is to calculate advertisement expenditure on all kind of products and deliver to their customers numbers on specific products. Customers of media monitoring companies are executives, policy and decision makers, who are interested to receive data on how much money one company spent on a specific advertisement campaign for one product. Currently most of the work of media monitoring companies is performed manually, resulting in enormous personnel efforts. The introduction of semi-automatic tools requires multi-modal analysis and cross media capability, which makes the task from a technical point of view very challenging. Within DIRECT-INFO project we targeted a specific area of media monitoring, sponsorship tracking, and created a first prototype system for context aware multi-modal analysis with cross media functionality.

NM2 aims at creating a variety of new media genres using all of the facilities of modern broadband communication and interactive terminals. New production tools for the media industry are created within the project that allow the easy production of non-linear broadband media which in turn can be personalised by the Viewers to interact directly with the medium and influence what they see and hear according to their personal preferences.

2. Sponsorship tracking in Direct-Info

DIRECT-INFO primarily targeted the media monitoring sector and specifically the needs of Media Information Firms that capture, monitor, archive, and analyze media information. As a concrete business case the project focused on sport sponsorship monitoring which practically means that a sponsor wants to know how often his brand appears in connection with the sponsored organisation. Knowledge about how often a sponsor is mentioned in connection with the sponsored party is a direct indicator for executive managers to estimate whether to continue sponsorship or invest e.g. in direct advertising activities. The sponsored party can use this information in order to further motivate the sponsor to invest. End-users of DIRECT-INFO are customers of media information companies, respectively top managers, communication managers or PR managers. For this reason an easy to use web based user interface was implemented which promptly presents all the relevant information in a summary page, allowing further “drill down” only if requested.

2.1. Technical overview

Technically the project covers cross media aspects and a multi-modal analysis. Relevant parts of TV streams and electronic press feeds are automatically selected and subsequently monitored to find appearances of the name or logo of a sponsoring company in connection with the sponsored party. For this purpose basic features are fully automatically extracted from TV and press and thereafter fused to semantically meaningful reports. Extracted features include logos, positive & negative mentions of a brand or product, multimodal video segmentation, speech-to-text transcripts, detected topics and genre classification.

From a technical point of view sponsorship tracking is a very complex task. The simple detection of e.g. a brand in one modality (e.g. video) is not sufficient in order to meet the requirements. In praxis a sponsor very often sponsors more than one party hence the context information is needed as well in order to filter relevant appearances.

A multi-modal analysis and fusion, which relates information from different modalities was needed in order to get this context information. Within DIRECT-INFO the multi-modal analysis covers four modalities (video, audio, text and images) and two media

(TV and press) [Rehatschek, H. (2004).] [Kienast, G. Stiegler H., Bailer W., Rehatschek H., Busemann St. Declerck Th. (2005)].

2.2. Workflow of the system

Main project result is a pilot system which has been installed at the premises of partner Nielsen Media Research, Italy. The functionality of the DIRECT-INFO pilot system can be best explained when looking at the workflow which is depicted in Figure 1 and consists of the following main steps

1. Acquisition records video chunks of constant length & Electronic Program guide (EPG) information and notifies the central content analysis controller (CAC) on their availability.
2. Based on EPG information the CAC prepares semantic blocks (represented as MPEG-7 documents) i.e. per sport event, TV show etc.
3. CAC starts an automatic genre classification subsystem on this semantic block in order to get another indicator – next to the EPG information - if the semantic block is relevant for analysis.
4. Based on a condensed result of the genre classification and the EPG information the CAC decides if the corresponding semantic block shall be analyzed or not.
5. If a semantic block is relevant for analysis, the CAC passes the block now according the user defined workflow to the corresponding analysis subsystems
6. After analysis a "Quality Check" is performed by the user (MPEG-7 result editor/viewer). The user can change begin/end of a semantic block and/or modify parameters, go to step 5.
7. After the quality check the results are passed to the fusion component.
8. Fusion component first automatically reduces the different results of the analysis subsystems according to user defined rules. Then based on user interaction the data will be classified.
9. The fused classified semantic blocks are stored in a local database of the fusion component.
10. If a specific customer request comes in the database can be queried via a set-up application for fused classified semantic blocks. Specific customer relevant data will be put together.
11. The delivery / Push system visualizes the output of this set-up application via a web interface and/or immediately alerts (via SMS, MMS or email) the end user in case of important events.

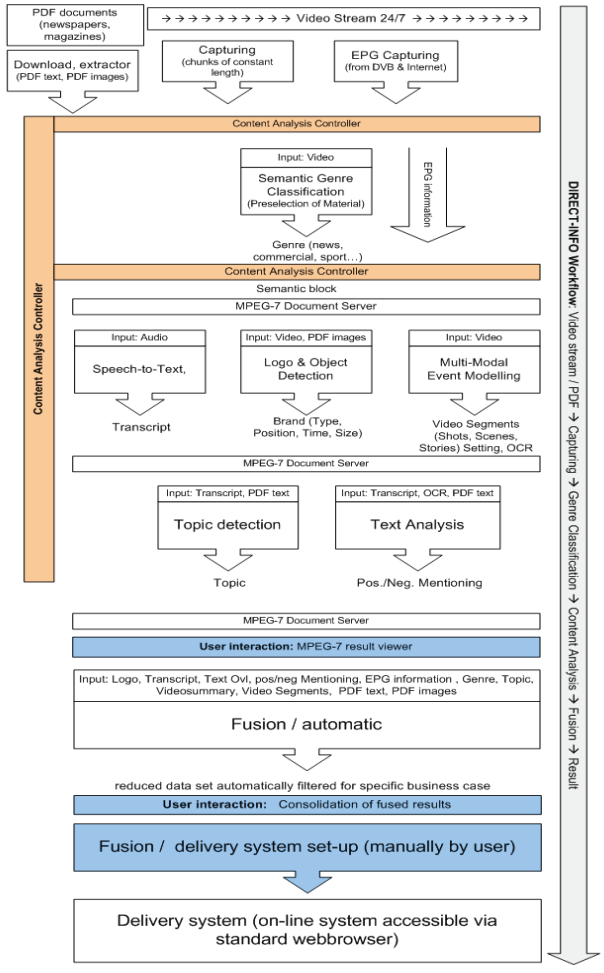

Figure 1: DIRECT-INFO system workflow

Next to the integrated system also standalone components of several meaningful subsystems have been developed. The scientifically most innovative are briefly discussed in the following sections.

2.2.1. Genre classification

This module analyses the video stream in real-time and decides on a shot basis the following semantic genres: commercials, sports, speaker, trailers and comics. The component is based on a generic approach and requires proper training on the genres of interest. As the relatively unreliable single shot classification is not directly used within DIRECT-INFO, the component summarizes the shot information and provides a highly reliable classification of semantic blocks.

The approach is based on low-level feature extraction and their combination in feature vectors that are related to the feature vectors of the training data according to the Bayesian decision theory and a Markov Model [Weiß J., (2005)].

2.2.2. Logo detection

The logo detection module detects visual appearances of logos in the video stream. The task is closely related to detecting known planar objects in still and moving images, with some special requirements. Logos vary in size, can be rotated and are subject to different lightning conditions. The SIFT (scale invariant feature

transformation) algorithm [Lowe, D. 1999], [Lowe, D. 2001] chosen by us is invariant against all these factors. Furthermore logos may be partly occluded or on non rigid surfaces (a player's shirt) so a logo has still to be detected, even if only parts of it are visible/matched. Some logos may appear in any color hence the chosen algorithm does not rely only on color features. The output of this subsystem is a list of logo appearances including time, size and position on the screen. In addition to the original Lowe SIFT we significantly improved the matching part of the algorithm and adapted it especially on video content. In particular we added a tracker [Lucas, B. D., Kanade T. (1981)], [Bouguet, J.-Y. (2000)] in order to improve performance and stability. We currently reach a precision of approx. 35% and a recall of approx. 85%.

2.2.3. Detection of pos./neg./neutr. Mentions

One very important aspect of monitoring is being able to detect positive or negative mentions of a brand. Understandably, the busy executives of a company that pay dear money for sponsorships and advertising are highly interested in receiving such information in a prompt manner. The linguistic and semantic analysis of all textual documents relevant to DIRECT-INFO is delivered by the WebSCHUG system as XML-encoded dependency structures that comply with the MPEG-7 format for textual annotation (the Linguistic Description Scheme) [Kienast, G., András H., Rehatschek H., Busemann St., Declerck Th., Hahn V., Cavet R. (2005).].

The annotation structure has been augmented with a 'polarity' tag. Polarity information is associated with linguistic units (e.g., words). The predicate-argument relations WebSCHUG can analyse allow us to support the more complex linguistic and semantic detection of 'positive' and 'negative' mentions. The WebSCHUG system can be parameterized as to which entities should be assessed. Parameters may include a list of synonyms which supports the inclusion of other brands and other use cases.

The system analyses PDF documents in Italian as well as English plain text, as it could result from ASR or from capturing TV text captures.

2.2.4. MPEG-7 result viewer & editor

The MPEG-7 Result Viewer & Editor is used to visualize analysis results per relevant semantic block (job) of the TV workflow (for the PDF workflow appearances will be directly visualized and checked within the fusion component). The application can be started from the Content Analysis Controller monitoring GUI to perform manual quality checks of the analysis. It consists of independent GUI components (video player, keyframe viewer, timeline view and result editing area) which are synchronized with each other in a common GUI framework. The operator may change in this editor either parameterization of analysis modules and restart the analysis on this semantic block or manually edit specific results in order to get better results of the fusion component.

2.2.5. Semantic data fusion component

Data fusion using different resources is a challenging task. As only high-quality results are acceptable to end-users, DIRECT-INFO opted for an automatic fusion process complemented with a manual assessment and correction phase. Hence quality assurance remains with the human media analyst.

The level of granularity is the *appearance*, representing an occurrence of a logo, a topic or a mention of interest. Fusing appearances requires a homogeneous representation scheme, which is defined using archetypes.

The technology used in the Fusion component is based on the Zope [Zope (2006)] application server, the Plone [Plone (2006)] content management system (CMS), and the Archetypes package that allows the easy definition of new content types for the Plone CMS. This software is in the public domain.

The Fusion Component works on an MPEG-7 document which stores all analysis information of one semantic block. From MPEG-7 content basic appearances per logo, a topic or a mention of interest are derived and stored. Basic appearances and further MPEG-7 information such as EPG data are then used to form complex appearances through a set of fusion rules. These rules are parameterized with respect to sponsor name, company name, or date and time. Results are assessed for correctness by the media analyst through the Facts Assessment Interface and stored in the Zope Object Database. The Setup Application Q/R interface queries and retrieves application-specific appearances according to end-user requirements. The media analyst decides which ones to make available to the end-user and stores them in the database for delivery to the end-user [Declerck, Th., Busemann St., Rehatschek H., Kienast G. (2006)].

3. New media formats in NM2

The State-of-the-Art in media production is the creation of entire finished stories that get delivered through myriads of distribution channels (TV, Internet, DVD, etc) to the end-user. With NM2 media professionals can instead conceive story components that can be used in the production of many different stories (equivalent to the car industry reusing components in different models) and screenwriters are supported to think in "story worlds". It is then up to the end-user to actually create her very own story on-the-fly based on her personal preferences.

In a technological sense innovation in NM2 concerns the development of new frameworks, technologies, tools, methods and architectures for narrative-based annotation (description of content), content recognition and content delivery to support the re-engineering of the production value chain and simultaneously enable the creation of a range of new, profitable, entertaining and engaging media genres with mass-market appeal to Europeans with screen dependent devices such as televisions, computers, games consoles and DVD players.

New tools for personalised, interactive and reconfigurable media productions are created within the project that will be elaborated in seven audio-visual interactive and non-linear productions. The NM2-productions range from news reporting and documentaries through a quality drama serial to an experimental television production about love.

3.1. Production Workflow

The State-of-the-Art production workflow comprises the following phases:

- The **pre-production phase** where preliminary steps before the actual shooting starts are made, followed by
- the **production phase** where shooting takes place, and
- Finally the **post-production phase** where the shot material gets finalized.

Compared to the classical setup introduced above, the NM2-prodcution workflow as depicted in Figure 2 is non-linear and iterative.

The production workflow differs in detail depending on the kind of production (cf. section 3.3, though all three production phases are present in every production.

While the NM2 tools are mainly situated in the post-production phase – where the tools have to integrate with existing post-production tools like existing non-linear editing systems (NLE) – they must also be capable to capture information in the production and pre-production phases due to the fact of the potential non-linearity of NM2-workflow. A very important task is to support NM2-producers in testing and previewing parts of the production to validate the artistic and semantic aspects of the story.

As shown in Figure 3 the NM2 system can be divided into three main areas of functionality:

Production tools support the creators of NM2 productions to produce stories. The production tools cover the ingestion of essence, the manual and automatic description of media items, and the authoring, i.e. the construction of possible stories. These tools integrate seamlessly into the existing production environments.

The **delivery system** presents the output to the end-users. In some NM2-productions customized end-user applications are required, which enable media composition at a user terminal. In some NM2-productions, the end-users are able to interact with the media and also with other consumers within the production in order to exchange comments about the narrative.

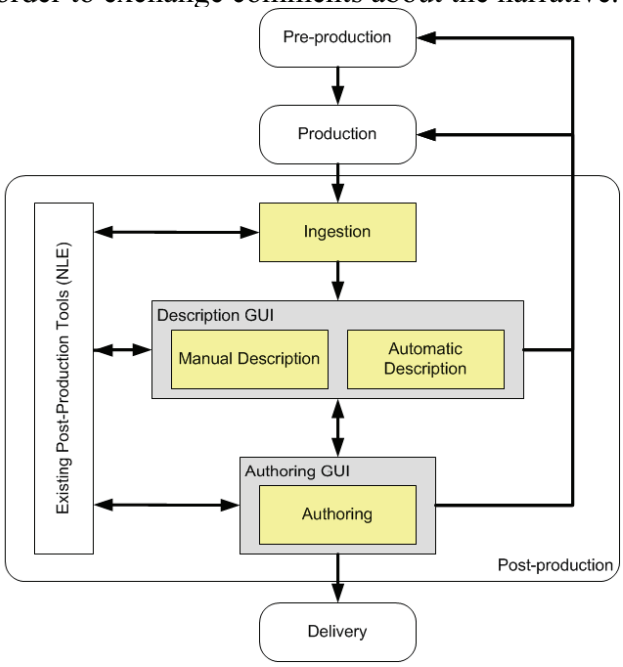

Figure 2: production workflow of NM2

3.2. Architecture

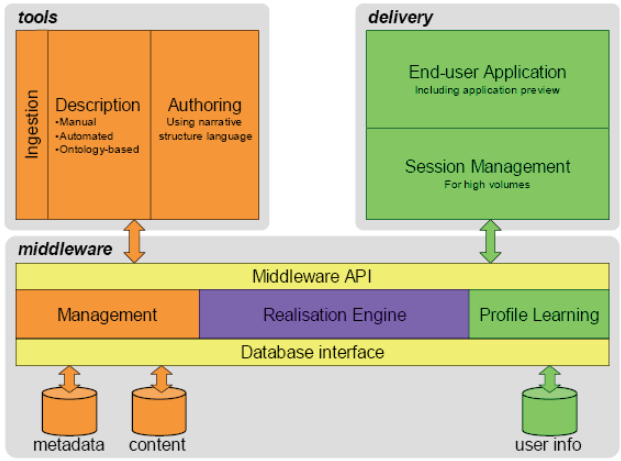

Figure 3: system architecture of NM2 system

This will contrast with the traditional broadcast structure of centralised media generation and delivery in a closed format. The delivery system is set up in a client-server model that is already supported in many popular domestic devices such as PCs, advanced set-top boxes and games consoles.

The **Middleware** mediates between the production tools and the delivery system by managing and interpreting the metadata and content. As such, the Middleware represents a core layer in which shared functionality is implemented. The middleware handles all data management tasks, including database management, automatic assembly of media essence based on metadata, interpretation and recording of user profiles as well as user

interaction. It finally includes the Realisation Engine, which is responsible for dynamically creating a user-specific story, based on a given story world and the interaction of a particular engager. A common API makes all of the above mentioned accessible to all instances of the production tools and delivery system.

3.3. Productions

The media productions are chosen to reflect a range of content genres. They are suited to a range of cross media publishing channels, including, broadcast (television), broadband delivery, and DVD. Each production is mentored by a target professional broadcaster/production company who are assessing the new media experiments:

"Gods In The Sky choice"

An interactive version of an imaginative and thought-provoking set of programmes exploring ancient myths in colourful and imaginative drama, dance and puppet theatre form, with a new astronomical interpretation.

City Symphonies

A new production in a traditional documentary genre. City Symphonies makes use of an old but recently revitalised screen language – montage – which has proved critical to the history of cinema, and is essential to any understanding of the relationship between cinema and the architecture of the city.

MyNews&SportsMyWay

A digital, interactive archive that makes it possible for engagers via broadband to discover, select and recombine news & sports items and stories according to their individual tastes.

Runecast

Runecast is inspired by the time-honoured oral-storytelling, performance-based structures, which contemporary interactive digital media re-enable in new forms. Engagers are enabled to compose their own coherent story constellations of songs, tales and images from mixed audio-visual media.

Gormenghast Explored

A fantastical, allegorical version of Mervyn Peake's great novel, originally produced by BBC Television. The content from the production will be developed to allow the story to be explored through new narrative paths, enabling flexible narrative structures in drama to be explored.

A Golden Age

An ambitious configurable documentary exploring the arts of the Renaissance in England, concentrating on the final two decades of Elizabeth I's rule. The engager determines the aspects of this subject which are of most interest, and the system produces in real-time a version which responds to these preferences.

Accidental Lovers

This production is a participatory black comedy about love for television, mobile phone and Internet using a generative narrative. The engager can affect in real-time the unfolding drama of the unlikely romantic couple, Juulia in her sixties and Roope in his thirties.

3.4. Market

European households are increasingly prepared to receive digital and interactive audio-visual media. At present up to 30 percent of households in the European Union can receive digital television, of which two-thirds receive digital TV via satellite. The next few years will see a substantial increase in digital broadcasting through all networks (DVB-T, DVB-C and DVB-S), but with national differences in the degree of household penetration and in which digital platform dominates the market. Estimations on future penetration vary substantially from 40 percent to 70 percent. Broadband Internet is also taking off rapidly in many European countries.

This distinction between the TV and the PC/Internet environment has consequences for the distribution platforms, devices and users at which the NM2 productions aim. At the same time the distinction between the two worlds seems to become less clear-cut. At present we see increasing convergence between both platforms and increasing possibilities for combining interactivity with attractive and entertaining viewer experiences.

3.5. End-user Devices

In the realm of NM2 end-user devices, we target at three platforms:

- **Barebone PC/Windows Media Centre and TV/Projector**. This is the combination used by most of the home theatre personal computer and media centres nowadays.
- **Game consoles**. This platform refers to the use of any of the most spread "living-room" game consoles available in the market nowadays or in the near future (such as PlayStation 2, XBox, and Nintendo Revolution)
- **Mobile phones**. Currently this platform is for possible spin-off applications, because the main goal of a living-room experience is not reached yet, the upcoming generation of mobile devices at least promises an comparable experience.

3.6. Understanding Visual Content

For non-linear interactive cinematic narrative, it is convenient to work with units, each of which constitutes a "micro-narrative"' contained within a video-clip (defined by its cut-in and cut-out points). This may consist of more than one shot and thus itself contain a number of internal cuts. These units are defined as narrative objects in NM2 and are the building blocks for

the interactive movies. The "glue" that keeps them together is represented by narrative structures using the Narrative Structure Language (NSL) as described below.

In NM2 the narrative objects reify as media objects that have metadata attached on different semantic levels w.r.t. formality and reusability:

- **AAF** (Advanced Authoring Format) [AAF, (2006)] is used to interface with existing NLE-systems and to integrate cutting metadata into the NM2-system.
- **MPEG-7** [MPEG-7, (2006)] is used to capture intrinsic low-level features of the content (colour descriptor, key frame, shot-border, etc.). NM2-tools target at extracting as much as possible automatically from the essence to produce sound MPEG-7 descriptions though some manual post-editing and/or validation is unavoidable in general. The so generated media objects are reusable, i.e. not production-dependent.
- **OWL-DL** [OWL, (2006)] is utilised to define production-specific characteristics for the multimedia objects, to add contextual information and interface to the NSL. Most of the high-level features are derived from MPEG-7 using domain specific mapping from features to semantic entities. A core ontology is defined that formalises all generic concepts and relations. In addition a production-specific ontology (based on the core ontology) is defined per production that describes concepts and relations depending on the domain of the production (news items, historical elements, etc.).

- The **NSL** – developed within NM2 – is a language for expressing non-linear narratives. In NM2 we distinct between specific narratives and global narratives. A specific narrative is a set of representations of media objects arranged into a play-list that is delivered to a NM2 end-user. A specific narrative can be regarded as being rendered as a number of layers playing in parallel, each playing a sequence of media objects. A global narrative contains the same references to media objects, but instead of fixed sequences, it specifies rules that are used to create a specific narrative on-the-fly based on context information. A specific narrative can insofar be regarded as an instantiation of a global narrative. The software that interprets a global narrative, producing a specific narrative, is referred to as the Inference Engine which is part of the above mentioned Realisation Engine (see section 3.2).

4. Conclusions and outlook

4.1. Conclusions DIRECT-INFO project

The main strength of the DIRECT-INFO system is in offering an integrated approach between several analysis components. Other systems on the market focus strictly on a single modality (e.g. brand recognition), whereas DIRECT-INFO provides a unique multi-modal approach which fuses information sources. Besides this the consortium has learned its lessons from the project, as given below.

Efforts for integration - especially for the set-up of a proper infrastructure, firewall configuration for remote access possibility and on-going administration - was higher than expected.

Overall system recall and precision was estimated on a worst case scenario by taking no human interaction into account (which is in praxis not the case) and on a specific use case as follows: rules assigned to a fusion use case: fr1, ..., frn; Subsystems involved in a fusion rule fr: s1(fr), ..., sm(fr);

then F value (fval) for fr = $\prod_k$ fval(sk(fr)), $1 \le k \le m$;

F value for fusion use case = $1/n \times \Sigma i \prod k$ fval(sk(fri)), $1 \le k \le m$, $1 \le i \le n$

For Juventus use case: assume we are interested in Tamoil logos, Tamoil logos during topics, and positive mentions of Juventus: Tamoil rule tr. Two subsystems involved (logo recognition, topic detection): fval(lr)= 52%; fval(td) = 72%; fval(tr) = 62%; Rules for pos mentions: fval(pm) = 63,5%; Juventus use case with lr, pm, tr: fval(Juv) = 59,2%

Even though Web Services are commonly seen as easy to use, it has to be stated, that it takes a considerable amount of time to get the necessary know how, even for very experienced programmers. The standardized SOAP protocol supports a broad functionality with numerous options, but not all tools have implemented all of them making interoperability sometimes cumbersome.

It was a good choice to define MPEG-7 as our general metadata exchange format within the system. The disadvantage of an initially higher learning effort is later gained multiple times by avoiding the time consuming phases of definition and continuous extension of a proprietary format.

4.2. Conclusions NM2 project

Currently prototype implementations for all parts of the system (production tools, delivery system, and middleware) exist and are elaborated in a number of productions within NM2.

From a technological point of view a document server is in use that allows transparent and efficient access to MPEG-7-related data either via an RDBMS or (if needed) on a file-system basis. The production information (described on project level) and the set of media object descriptions is managed using a newly

developed OWL Data Store on top of the Jena Framework [Jena, (2006)]. The NM2-system is written largely in C++ and Java – utilising XML-RPC [XML-RPC, (2006)] to ensure interoperability – and partly in Prolog to capture the NSL-rules used by the Inference Engine.

NM2 in the present setup focuses on moving image, but subsequent research projects could apply the methodologies developed in NM2 to produce new cost effective ways of creating multiple versions of content in other media types as well.

4.3. Outlook: MediaCampaign project

Seamlessly with the end of DIRECT-INFO a new R&D project – MediaCampaign - in the area of cross-media analysis is started. MediaCampaign's [Rehatschek, H. (2005)] scope is on discovering, inter-relating and navigating cross-media campaign knowledge. A media campaign is defined as the universe of measures in order to fulfill a specific objective. The project's main goal is to automate to a large degree the detection and tracking of media campaigns on television, Internet and in the press. This will lead to new research results in media monitoring and analysis, and we aim to positively impact the European scientific community.

5. Acknowledgements

The R&D work described in this paper and performed in the projects DIRECT-INFO (IST FP6-506898), „nm2 – New Media for a New Millennium“ (FP6-004124) and MediaCampaign (IST FP6-027413) is partially funded under the 6th Framework Programme of the European Commission of the IST Work Programme 2003 – 2005/06. More information about the projects can be found on the corresponding project public websites http://www.direct-info.net and http://www.ist-nm2.org.

6. References

AAF, (2006). AAF Association. Homepage. <URL: http://www.aafassociation.org/>

Bouguet, J.-Y. (2000). Pyramidal Implementation of the Lucas Kanade Feature Tracker: Description of the algorithm, *Technical Report, Intel Corporation*, Microprocessor Research Labs 2000.

Declerck, Th., Busemann St., Rehatschek H., Kienast G. (2006). Annotating Text Using the Linguistic Description Scheme of MPEG-7: The Direct-Info scenario. *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006), EACL (European Chapter of the Association for Computational Linguistics)*, Trento, Italy, April 2006.

Jena, (2006). The Jena Framework. Homepage. <URL: http://jena.sourceforge.net/>

Kienast, G. Stiegler H., Bailer W., Rehatschek H., Busemann St. Declerck Th. (2005). Sponsorship Tracking Using Distributed Multi-Modal Analysis (Direct-Info). *Proceedings of the International Workshop on the integration of knowledge, semantics and digital media technology (EWIMT)*, ISBN-0 86341 595 4 / 9780863415951, London, November 2004, pp. 341 - 348.

Kienast, G., András H., Rehatschek H., Busemann St., Declerck Th., Hahn V., Cavet R. (2005). DIRECT INFO: A Media Monitoring System for Sponsorship Tracking, *Proceedings of the Twenty-Eigth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Workshop on Multimedia Information Retrieval*, Aug. 2005, Salvador, Brasil.

Lowe, D. 1999. Object Recognition from Local Scale-Invariant Features. *In Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 1150-1157.

Lowe, D. 2001. Local Feature View Clustering for 3D Object Recognition. *In Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lucas, B. D., Kanade T. (1981). An Iterative Image Registration Technique with an Application to Stereo Vision. *International Joint Conference on Artificial Intelligence*, pages 674-679, 1981.

MPEG-7, (2006). Official Standard Site. Homepage. <URL: http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>

OWL, (2006). OWL Web Ontology Language. Homepage. <URL: http://www.w3.org/TR/owl-features/>

Plone (2006). Content Management System. Homepage. <URL: http://www.plone.org>

Rehatschek, H. (2004). DIRECT-INFO: Media monitoring and multimodal analysis for time critical decisions. *Proceedings of the 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, ISBN-972-98115-7-1, Lisbon, April 2004.

Rehatschek, H. (2005). MediaCampaign - Discovering, Inter-Relating And Navigating Cross-Media Campaign Knowledge. *Proceedings of the International Workshop on the integration of knowledge, semantics and digital media technology (EWIMT)*, ISBN-0 86341 595 4 / 9780863415951, London, November 2005, pp. 335 - 336.

Weiß J., (2005). Genre Classification: Semantic Interpretation of Video. *Diploma thesis Graz University of Technology, Austria.* <URL: http://hs-art.com/products/download/download.html>.

XML-RPC, (2006). XML Remote Procedure Call. Homepage. <URL: http://www.xmlrpc.com>

Zenith Optimedia, (2006). Homepage. <*URL: http://www.zenithoptimedia.com*>.

Zope (2006). Application Server. Homepage. <URL: http://www.zope.org>

Representation and Analysis of Multimedia Content: The BOEMIE Proposal

D.I. Kosmopoulos, V. Karkaletsis, S. Perantonis, G. Paliouras, C.D. Spyropoulos

Institute of Informatics and Telecommunications
National Centre for Scientific Research "Demokritos"
P.Grigoriou & Neapoleos str., 15310 Agia Paraskevi Attikis, Athens, Greece
{dkosmo,vangelis,sper,paliourg,costass}@iit.demokritos.gr

Abstract

We propose an approach to knowledge acquisition, which uses multimedia ontologies for fused extraction of semantics from multimedia content, and uses the extracted information to evolve the ontologies. We present the basic components of the proposed approach, describe an application scenario we currently examine, and discuss the open research issues focusing on knowledge representation and extraction techniques that will enable the development of scalable and precise knowledge acquisition technology.

1. Introduction

The objective of multimedia content analysis is the automated knowledge acquisition from various modalities, e.g., text, images, video etc. The high complexity that characterizes the multimedia content, along with the currently prevailing dearth of precise modeling for multimedia concepts, makes automatic semantics extraction a very challenging task.

Although latest advances in content analysis have improved capabilities for effective searching and filtering, a gap still remains between the low-level feature descriptions, and high-level semantic descriptions of concepts. A suitable approach to fill this gap is to use a semantic model in the extraction process. Moreover, the analysis of single modalities, in particular of visual content alone, is inadequate in all but a small number of restricted cases.

The proposed approach, which is envisaged in the framework of the IST project BOEMIE, is unique in that it links multimedia extraction with ontology evolution. This approach will be used to enrich digital maps with multimedia content related to city events. The content is collected from various proprietary or open sources and it becomes automatically semantically annotated. Driven by domain-specific multimedia ontologies, the information extraction systems implementing the proposed approach will be able to identify high-level semantic features in image, video, audio and text and fuse them for optimal extraction. The ontologies will be continuously populated and enriched using the extracted semantic content. This is a bootstrapping process since the enriched ontologies will in turn be used to drive the multimedia information extraction system.

This work provides the key ideas involved in the whole system and then focuses on the semantics extraction. Section 2 highlights the related research. Section 3 presents the main aspects of the proposed approach, the architecture and its basic components. Section 4 provides an application scenario we are currently examining for the evaluation of the proposed approach. Section 5 discusses some of the issues that arise under this bootstrapping framework. The paper concludes presenting our next steps for the implementation of the proposed approach.

2. State of the art

The involved technologies include the semantics extraction from multimedia content, the multimedia ontologies and techniques that exploit their synergy.

Semantics extraction from multimedia content is the process of assigning conceptual labels to either complete multimedia documents or entities identified therein. In general, extraction can be performed at the levels of l*ayout (structure), content* and *semantics* (intended meaning of the author).

In the case where content is available in multiple related modalities, these can be combined for the extraction of semantics. The combination of modalities may serve as a verification method, a method compensating for inaccuracies, or as an additional information source (Snoek and Worring 2005). The processing cycle of combination methods may be iterated allowing for incremental use of context. The major open issues in the combination approaches concern the efficient utilization of prior knowledge, the specification of open architecture for the integration of information from multiple sources and the use of inference tools for efficient retrieval.

Most of the extraction approaches encountered in the literature are based on learning methods, e.g., naive Bayes classifiers, decision tree induction, k-Nearest neighbour, Hidden Markov model (Manning and Schutze 1999, Rabiner 1989). However, with the advent of promising methodologies in multimedia ontology engineering, knowledge-based approaches are expected to gain in popularity and be combined with the machine learning methods. This is also the case we will study in the proposed approach.

Ontologies can play a major role in multimedia content interpretation because they can provide high-level semantic information that helps disambiguating the labels assigned to multimedia objects. Indicative approaches for constructing multimedia ontologies are the ones presented in Hunter 2001, Mezaris et al 2004, and Troncy 2003. The major open issues here concern the automatic mapping between low level audio-visual features and high level domain concepts, the automated population from unconstrained content and when there are no metadata attached to the content. In cases of complex domains, multiple ontologies may be present and ontology

coordination techniques have to be employed (e.g., Castano 2004, Kotis and Vouros 2004, Gomez 2002).

The interaction between information extraction and ontology learning has also been modelled at a methodological level as a bootstrapping process that aims to improve both the conceptual model and the extraction system through iterative refinement. In Maedche and Staab 2000, the bootstrapping process starts with an information extraction system that uses a domain ontology. The system is used to extract information from text. This information is examined by an expert, who may decide to modify the ontology accordingly. The new ontology is used for further information extraction and ontology enrichment. Brewster et al. 2002 propose a slightly different approach to the bootstrapping process. Starting with a seed ontology, usually small, a number of concept instances are identified in the text. An expert separates these as examples and counter-examples which are then used to learn extraction patterns. These patterns are used to extract new concept instances and the expert is asked to re-assess these. When no new instances can be identified, the expert examines the extracted information and may decide to update the ontology and restart the process.

3. Methodology and architecture

We advocate an ontology-driven multimedia content analysis (semantics extraction from images, video, text, audio/speech) through a novel synergistic method that combines multimedia extraction and ontology evolution in a bootstrapping fashion (see Figure 1). In the following sub-sections we describe the proposed components.

3.1. Semantics Extraction from Multimedia Content

A suitable approach to bridge the semantic gap is to use a semantic model in the extraction process. Moreover, the analysis of single modalities, in particular of visual content alone, is inadequate in all but a small number of restricted cases. The effort required to provide problem-specific extraction tools makes single-media solutions non-scalable, while their precision is also rarely adequate. In the proposed approach, on the level of individual modalities, particular emphasis will be given to visual content, from images and video, due to the richness of this source and corresponding difficulty of extracting useful information. Non-visual content, audio/speech and text, will provide supportive evidence, in order to improve extraction precision. Since no single modality is powerful enough to encompass all aspects of the content and identify concepts precisely, fusing information from multiple media sources is needed.

3.2. Multimedia Ontologies

In our approach, we propose the development of a unifying representation for multimedia ontologies and related knowledge. This "multimedia semantic model" will serve as an integrated model for the different ontologies that are necessary to support the semantics extraction process:

- Multimedia content ontology: It represents the structure of the content of the multimedia documents. The top level hierarchy of a multimedia document is classified into: Image, Video, Audio, Audiovisual and Multimedia. Each of these types

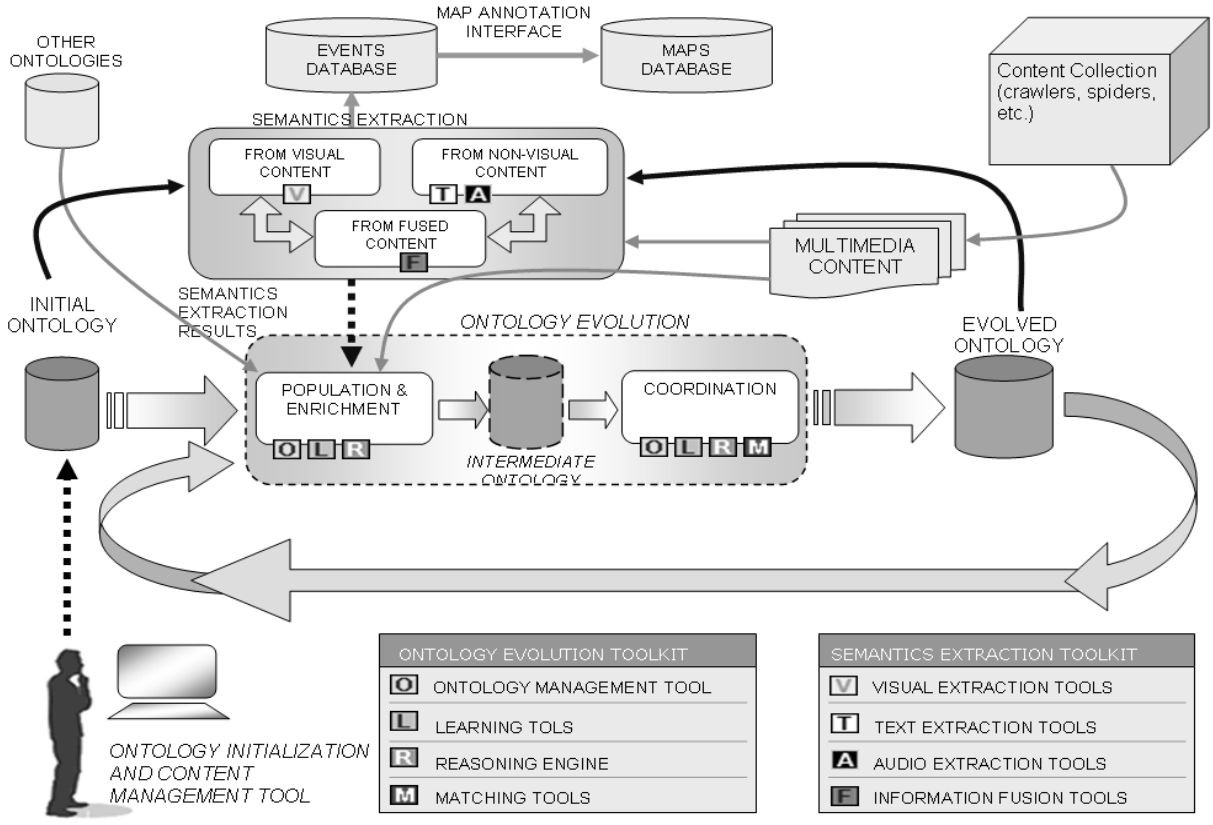

Figure 1: Architecture of the integrated system

has its own segment subclasses. These subclasses describe the specific types of multimedia segments, such as video segments, moving regions, still regions and mosaics.

- Multimedia descriptor ontologies: This ontology models concepts and properties that describe visual characteristics of objects in terms of low-level features and media structure descriptions. Sub-concepts will include MPEG-7 standard features like colour, shape, texture, motion, localization and basic descriptors. Separate descriptor ontologies apply to different modalities. Along with the multimedia descriptor ontologies is used the hierarchical "fusion model", where the significance (weight) of each modality is defined for each concept.
- Domain-specific ontologies: These ontologies contain concepts and properties related to the knowledge of the domain of interest. In these concepts we assign instances, which are used to recognize semantic objects using the results of the content analysis process. These ontologies also contain detailed descriptions of objects using spatiotemporal and partonomic relations defined in the multimedia semantic model.

3.3. Evolution of Multimedia Ontologies

According to Stojanovic 2004 ontology evolution is "the timely adaptation of an ontology to the arisen changes and the consistent propagation of these changes to dependent artefacts". Thus, ontology evolution is a complex process, involving the following sub-processes: ontology *population* and *enrichment*, i.e., addition and deletion of concepts, relations, properties and instances, *coordination* of homogeneous ontologies, e.g. when more than one ontologies for the same domain are available, and heterogeneous ontologies, e.g., updating the links between a modified domain ontology and a multimedia descriptor ontology, *maintenance* of semantic consistency, since any of the above changes may generate inconsistencies in other parts of the same ontology, in the linked ontologies or in the annotated content base.

Our approach for ontology population and enrichment will be based on machine learning techniques using the information from the semantics extraction process. More specifically, the extraction process will populate the ontologies with instances of the various concepts, together with their properties and will also provide unclassified entities extracted from the multimedia content which may lead to suggestions for the enrichment of the ontologies with new concepts and relations, through novelty detection. This novelty detection is based on information from all different types of media being processed.

Ontology coordination approaches will be devised to interlink ontologies with different levels of heterogeneity. Ontology coordination involves the use of matching techniques and tools for mapping, alignment and merging.

During ontology evolution, any of the changes may generate inconsistencies in other parts of the same ontology, in the linked ontologies or in the annotated resources. At the current state of the art, description logic reasoning systems (e.g., RACER[1]) are not tailored to these "incremental changes". We will investigate how such changes can be much more efficiently supported. The aim is the development of models, techniques, and tools for semantic consistency checking of ontology content throughout the evolution process.

4. Application scenario

The application we are currently examining for the evaluation of the proposed bootstrapping approach concerns the enrichment of digital maps with semantic information. In other words, the results of the semantics extraction process will be displayed to the end-user, through an interactive digital map.

The user is interested to find "what" happens "where" in the city. The possible queries can be, events of a particular type in a specific time frame, events in a venue – location, persons related to event (e.g., actors, players etc), events at specific dates, events similar to a given one, events at nearby venues. It is assumed that we have discrete locations in the map, where all possible events are allowed to take place. As a concrete example of the application scenario, we propose the domain of sports where the user asks to know about a specific sport event (e.g., football game) in the city. He receives a list of games and is able to browse multimedia content related to game type, league, previous games, comments-gossip-interviews on the game to be played, concerning team history or the football ground.

4.1. Initialization

We will start by collecting, extending and merging existing ontologies for sub-domains referring. These ontologies will also be linked to the appropriate multimedia descriptor ontologies. This process will be accomplished using the ontology initialization and content annotation tool and will result to the initial multimedia semantic model for the domain.

4.2. Training

The various semantics extraction and ontology evolution tools are trainable to the domain. Therefore, a training dataset needs to be collected and used to customise the system. This training set should contain representative and annotated multimedia content, as expected to be encountered by the system at run time.

4.3. Information gathering

Having customised the system, the first step of its run-time use is to collect content from various Web and proprietary sources. In the case of sports events, such sources may include TV and news programmes, on-line magazines, sports-related sites, specialized discussion fora and Weblogs, as well as generic content sources.

4.4. Semantics extraction

The trained semantics extraction tools will be applied at regular intervals to the incoming stream of multimedia content, performing extraction of the relevant information from each piece of content.

[1] http://www.sts.tu-harburg.de/~r.f.moeller/racer/index.html

We define some city-specific concepts in the multimedia ontologies. For each concept we should have already defined features (in the multimedia descriptor ontology) and some classification parameters that enable a decision about the content with a certain probability or certainty factor or fuzzy membership (hard rules are inappropriate due to uncertainties in processing). This applies to all available modalities (text, images, video, audio), which means that individual modality – specific features (and classifiers) are available and decoupled from other modalities.

There is a "fusion model" with a hierarchical structure, which receives a decision input from the individual modalities (see Figure 2). For each node it holds the information about the weight of the decision taken by each modality. Its role is to combine the decisions taken by each modality-specific processing by applying the respective weights.

For the sports scenario we assume that the multimedia ontology defines the following city-specific classes-concepts (from general to more specific): *Indoor-outdoor*. The outdoor may decompose to *concert*, *sports*, *theater*. Each of the above categories may decompose to relevant subcategories. For example, the sport decomposes to *football*, *swimming* in the initial ontology. More football-like sports may be identified progressively through the evolution of the initial ontology.

In the example scenario the system

(i) receives the input, which happens to be a video scene of football.

(ii) The multimodal information is separated and processed separately to the visual part, and the audio part.

(iii) The information is processed hierarchically assuming no prior knowledge.

(iv) Using the visual features we classify the content with respect to the highest concept in the hierarchy, i.e., the indoor – outdoor. The appropriate feature for this task is the color. We find that the percentage of "green" is high in the average image histogram so the classifier gives a probability of 0.7 for outdoor and 0.3 for indoor (P(outdoor)=0.7, P(indoor)=0.3). Other uncertainty representations instead of probabilities could be applicable too.

(v) Using the audio ontology we find that there are no sounds that are typical for outdoor environment, e.g., sounds of birds, waves, wind etc, so the classifier gives P(outdoor)=0.45 and P(indoor)=0.55.

(vi) The text processing (after OCR) does not provide any relevant info so the probability is shared between the two classes (P(outdoor)=0.50 and P(indoor)=0.50).

(vii) The "fusion model" has predefined through training that the weights for the visual, audio and text modalities (W_v, W_a, W_t, given in Figure 2) and based on them it decides that P(outdoor)=0.565. So we proceed to the next layer examining the "child" of the outdoor concept.

Similarly we examine the three modalities to classify to event categories.

(i) The visual modality finds very high motion and gives P(sport)=0.7, P(concert)=0.2, P(theatre)=0.1.

(ii) The audio detects speech and crowd sounds and gives P(sport)=0.5, P(theatre)=0.4, P(concert)=0.1.

(iii) The text does not find relevant features so the probability is shared to the three concepts.

(iv) The "fusion model" decides using the related weights that P(sport)=0.589.

The next level has to do with classification into football and swimming.

(i) The visual ontology defines human motion features, color histogram and we calculate P(football)=0.9.

(ii) The audio ontology identifies patterns related to football such as "goal", "corner", "foul" and therefore gives probability P(football)=0.85.

(iii) The text identified gives the score, the team names and thus P(footbal)=0.7.

(iv) The "fusion model" decides that P(football)=0.86. So the shot is classified as "football".

Additional information that could be used include team names, team order in the score board, player names etc, which are associated with a certain football stadium. We assume that the number of football stadiums is limited in a city. The information is localized in the map based on the known locations of the football stadiums and the teams that are associated with them.

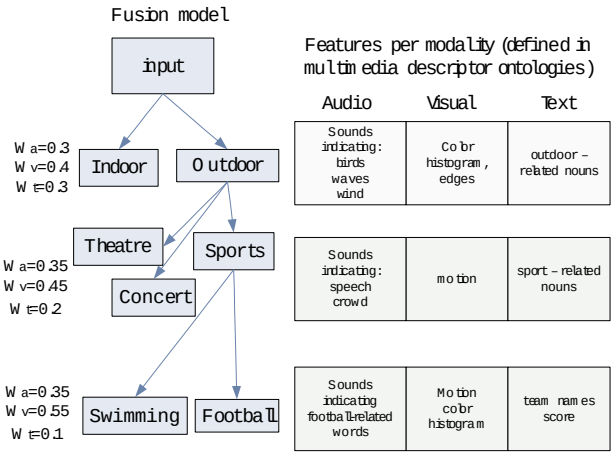

Figure 2: The fusion model for the described scenario, which defines the concept hierarchy and the modality weights per concept and the related multimedia features used.

4.5. Ontology evolution

The former, extraction task will populate the ontologies with instances of the various concepts, together with their properties. This process will also be accompanied by the appropriate annotation of content in the server, in order to provide semantic access to the content by the end-user. The latter, concept modelling task, performed by the extraction methods, will lead to suggestions for the enrichment of the ontologies, through novelty detection. The evolution can be performed through clustering with respect to specific features. As

new sport events may emerge (e.g., football on beaches) the knowledge representation has to evolve accordingly.

5. Discussion

In terms of semantics extraction from multimedia content, we propose the integration of an ontology-based approach with a probabilistic inference scheme. We need to examine carefully the role of the ontology in fusing information extracted from multiple media. We also have to examine new ways to fuse features derived from multimedia content.

Ontologies must be sufficiently expressive to describe the construction space for possible interpretations in general and for specific interpretation results in terms of a particular piece of media. Multimedia applications have highlighted the need to extend representation languages with capabilities which allow for the treatment of the inherent imprecision in multimedia object representation, matching, detection and retrieval. Existing standard web languages do not provide such capabilities. Therefore, considerable research effort needs to be directed towards representation and management of uncertainty, imprecision and vague knowledge in real life applications.

In terms of ontology population and enrichment, we will exploit the multimedia semantic model as well as current research on learning and inference techniques aiming to develop a generic framework for ontology learning and inference from multimedia content, due to the complexities introduced by the multimedia context. Addition of instances in the multimedia descriptor ontology may also require updating the corresponding link with the domain-specific ontology. The semantics extraction process will provide unclassified entities extracted from the multimedia content which may lead to the enrichment of the ontologies with new concepts and relations based on information from all different types of media being processed. Concerning inference techniques for ontology population and enrichment, we need to optimize and enhance description logic inference technology to support learning and retrieval requirements.

We also propose the use of machine learning techniques to assist ontology coordination in this context and we need to investigate the appropriate methods. This depends very much on the type of training data that is available. Supervised learning of complex representations requires data that may not be possible to acquire manually. Unsupervised or partially supervised methods may prove more useful in these cases.

Concerning semantic consistency checking in ontology evolution, there are two main problems. The first occurs at the instance level and requires techniques for efficiently handling incremental additions of instances, while checking integrity constraints. A second one occurs at the concept level and requires techniques for checking the consistency of new concepts against the current ontology, to choose a valid and consistent enrichment solution among a set of possible alternatives.

6. Concluding remarks

We propose a new approach towards automation of knowledge acquisition from multimedia content, by introducing the notion of evolving multimedia ontologies which will be used for the extraction of information from multimedia content. We have outlined the approach through a sports scenario. This is a synergistic approach since it combines multimedia extraction and ontology evolution in a bootstrapping process involving, on the one hand, the continuous extraction of semantic information from multimedia content in order to populate and enrich the ontologies and, on the other hand, the deployment of these ontologies to enhance the extraction robustness.

The main measurable objective of this initiative is to improve significantly the performance of existing single-modality approaches in terms of scalability and precision. Towards that goal, our aim is to develop a new methodology for extraction and evolution, using a rich multimedia semantic model, and realize it as an open architecture. The architecture will be coupled with the appropriate set of tools.

7. Acknowledgements

BOEMIE is an FP6-IST-Call 4 project (027538), to begin in March 2006. BOEMIE consortium consists of NCSR “Demokritos” (GR), Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung (DE), University of Milano (IT), Centre for Research and Technology Hellas (GR), Hamburg University of Technology (DE), TeleAtlas (BE).

8. References

Cees G.M. Snoek, M. Worring (2005). Multimodal Video Indexing: A Review of the State-of-the-art, *Multimedia Tools and Applications*, 25, pp. 5–35.

Manning C.D. and Schutze H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, USA.

Rabiner L.R. (1989). A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257–286.

Hunter J. (2001). "Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology", *International Semantic Web Working Symposium (SWWS)*, Stanford, July 30 - August 1,

Mezaris V., Kompatsiaris I., Boulgouris N.V. and Strintzis M.G. (2004). "Real-time compressed domain spatiotemporal segmentation and ontologies for video indexing and retrieval", *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Audio and Video Analysis for Multimedia Interactive Services*, vol. 14, no. 5, pp. 606-621.

Troncy R. (2003). “Integrating Structure and Semantics into Audio-Visual Documents”, In the *2nd International Semantic Web Conference, ISWC 2003*, LNCS 2870, pp. 566-581.

Castano S., Ferrara A., Montanelli S., and Racca G., (2004). “Matching Techniques for Resource Discovery in Distributed Systems Using Heterogeneous Ontology Descriptions”, *IEEE Proc. of the International Conference on Coding and Computing (ITCC04)*, Las Vegas, Nevada, USA

Kotis K., Vouros G. (2004). HCONE approach to Ontology Merging. *ESWS'04: The Semantic Web: Research and Applications,* LNCS, Vol. 3053, Springer-Verlag.

OntoWeb. Deliverable D1.3. (2002). A survey on ontology tools, May (ed. A. Gómez Pérez)

Maedche A. and Staab S. (2000). Mining ontologies from text. In R.Dieng and O.Corby, editors, Proceedings of EKAW-2000, LNCS, v.1937, pp. 189–202. Springer.
Brewster, Ciravegna F., and Wilks Y. (2002). User-centred ontology learning for knowledge management. In B. Andersson, M. Bergholtz, and P. Johannesson, editors, NLDB, volume 2553 of LNCS, pages 203–207. Springer.
Stojanovic L. (2004). Methods and Tools for Ontology Evolution. PhD thesis, University of Karlsruhe.

X-Media: Large Scale Knowledge Acquisition, Sharing and Reuse across-Media

F. Ciravegna*, S. Staab† and the X-Media Consortium

*Department of Computer Science, University of Sheffield, Sheffield S1 4DP UK,
†Department of Computer Science, University Koblenz-Landau, 56016 Koblenz, Germany,
f.ciravegna@dcs.shef.ac.uk staab@uni-koblenz.de

Abstract

X-Media is an integrated Project funded by the European Commission which addresses the issue of knowledge management in complex distributed environments. It will study, develop and implement large scale methodologies and techniques for knowledge management able to support sharing and reuse of knowledge that is distributed across different media (images, documents and data) and repositories (data bases, knowledge bases, document repositories, etc.). The project starts in March 2006 and will last for 4 years. In this paper we provide an overview of the project and outline the cross-media approach.

1. Introduction

While in the past, medium size, mainly textual, centralized archives used to be the only resources for knowledge management, nowadays large companies handle very large quantities of multimedia information in distributed archives. Their intranets connect thousands of computers and reach sizes of dozens of millions of documents. In addition, the increased use of the WWW as a source of information has made the boundary between intra- and inter-net very thin. This dramatically increases the size of the information space. Moreover, databases and archives are used to store huge amounts of information that is vital for the organization life, such as data on products, financial information, etc. Therefore Collecting and aggregating multimedia knowledge is of fundamental importance in order to gain competitiveness and to reduce costs. For example thousands of documents are produced during the design and manufacturing of a class of jet engines. During service, a single engine produces about 1Gbyte of vibration data per flight; if irregularities are found, part of the data is stored. Every time an engine is serviced, financial information is produced. If problems are found, pictures are taken, reports are written. Each individual engine has a potential "folder" of information describing the whole lifecycle of the engine that can easily sum up to several Gigabytes of information, potentially Terabytes, and contains highly interrelated information stored in different media. From a knowledge management point of view, information on engines belonging to the same class is to be collected and compared in order to:

Spot trends and problems common to the whole class, requiring actions in terms of design, manufacturing or organization of service; the sooner the action is taken, the lower is the cost in terms of maintenance and in terms of customer satisfaction

Identify rising problems on the individual and fix it before it is seriously damaged or breaks down; the latter point is particularly important for very expensive artefacts that also carry important safety constraints.

Compile, and then disseminate and share, best practices for design, manufacturing and service in order to minimize the downtime.

Such needs are common to a large part of the manufacturing industry, including car companies, aerospace companies, electronic appliances producers and many others. In each case it may change the scale of the production and the type of products, but the problem of dealing with large amount of multimedia information is common.

The growing size and the multi-media nature of the archives has serious implication on the way knowledge management can be implemented. There are a number of dimensions along which the complexity arises:

Focusing: large amount of information implies that managing knowledge becomes more complex and needs powerful focusing methodologies. Focus of searching changes in time and from user to user, and requires a balanced mixture of exploration and searching;

Knowledge integration: large distributed archives require the ability to map the distribution of information, to weight every single source and to distribute searches carefully; this is very difficult and often search is performed just in some of the archives, disregarding others that can bring very useful information;

Uncertainty and Dynamicity: information is often ambiguous, incomplete, or referring to a specific context - therefore archives can contain noise and imprecision, as well as obsolete information; each piece of knowledge must therefore be judged based on provenance, evidence, etc.

Cross-Media: evidence is often distributed in different media; it is possible that knowledge expressed in just one medium does not carry enough evidence. Connecting information in more than one medium is often required.

Infrastructure: different media cannot easily be shared. A folder of text documents may be sent via email, but a folder of images may not, and may instead require a shared image repository. For 10 GByte of data remote access to the underlying data base is to be considered.

Current knowledge management technologies and practises cannot cope with such new situation, as they mainly provide simple mechanisms (e.g. keyword searching) for supporting knowledge workers manually *pierce* together the information from different sources.

2. X-Media

X-Media addresses the issue of knowledge management in complex distributed environments. It will study, develop and implement large scale methodologies and techniques for knowledge management able to support sharing and reuse of knowledge that is distributed in different media (images, documents and data) and repositories (data bases, knowledge bases, document repositories, etc.). Technologies will be able to support knowledge workers in an effective way, (i) hiding the complexity of the underlying search/retrieval process, (ii) resulting in a natural access to knowledge, (iii) allowing interoperability between heterogeneous information resources and (iv) including heterogeneity of data type (data, image, texts). The expected impact on organizations is to dramatically improve access to, sharing of and use of information by humans and between machines. Expected benefits are a dramatic reduction of management costs and increasing feasibility of complex knowledge management tasks. The project plan is structured along the four areas listed below.

2.1. Area 1: knowledge sharing and reuse

X-Media will study and implement technologies and methodologies for easy and intelligent access to and reuse of formalised and non formalised knowledge. The reuse will take into consideration the user context to help focus searches and reuse. Reuse and sharing will be enabled via cross-media ontology supported automatic indexing. The technology will work in a largely automated way, but it will be centred on supporting users' work, rather than replacing them. This is because the activity of a knowledge worker is complex and humans are irreplaceable agents in this process. In this context, we will study, design and develop: (1) Effective and efficient new paradigms for knowledge retrieval, sharing and reuse which enable users to define and parameterize views on the available knowledge according to their needs. (2) Novel and cutting-edge knowledge fusion methods to support knowledge workers in making decisions when confronted with – possibly contradicting – knowledge derived from different resources. (3) techniques able to represent and manage (i) uncertainty, (ii) trust and provenance as well as (iii) dynamic aspects of knowledge. Usability will be a major concern together with ease of customisation for new applications.

2.2. Area 2: automated knowledge acquisition and extraction from documents, images and raw data.

Functional to the methodologies for knowledge sharing investigated in Area 1, is the ability to acquire knowledge across media in a rich, semantically-oriented way. X-Media will develop a set of tools able to support sharing methodologies in a seamless and automatic way. Media addressed are raw data, texts and images (e.g. results or parameters in experiments, raw images, textual documents, etc.). The outcome of the acquisition technologies will be a semantic representation of the content (conceptualization) to be used for knowledge management purposes. Enrichment of multimedia documents with additional layers of automatically generated annotation will be the main medium of associating conceptualizations to resources.

2.2.1. Limitations of Existing Technology

Current technology focuses on single medium technologies to acquire knowledge in multi media environments; this means that retrieval methods use mainly one medium (e.g. text) even in multimedia environments. This is the methodology adopted by some European projects such as IST-MUMIS (project providing indexes for multimedia material) or the IST-PrestoSpace, as well as services like Google Images. To our knowledge, there are no technologies available for information extraction that work truly cross media and that can be used in cases where information in one medium is necessary to understand the information in the other.

2.2.2. Innovation in X-Media: Multi and Cross Media Information Extraction

The strength of the approach will come from the flexible, deep combination of media. The project will provide technologies for two approaches to multi-media content.

On the one hand a **multi-media approach** will be pursued where information from each data source is initially extracted separately and – still separately - transformed into conceptual information. The mean for assigning conceptual information to the different sources is annotation with respect to an ontology as proposed by the Semantic Web community. The resulting pieces of conceptual information are joined to a common description via information fusion. Multimedia IE is expected to provide a richer way of extracting information and will enable access to knowledge using different media. Though multi media information extraction solves the problem of extraction information from several media, it has a severe limitation in that each IE process lives in a separate universe. This means that each process has to produce its own internal evidence in order to create conceptual information. In some cases, identifying such evidence can be a difficult or even an impossible task in one single medium, although plenty of additional evidence might be available across the media.

For this reason, X-Media will investigate **cross media information extraction (CM-IE)**, where evidence is considered across different media; this means that each extractor will be able to use evidence from different media in order to assign conceptual descriptions. The advantage of cross media IE will be more effective and robust extraction of information.

Evidence in different media will be represented in terms of conceptual information that will constitute also the output of the CM-IE process. So, both evidence and outputted conceptual information will be used by other modules as **background conceptual knowledge**, i.e. pre-existing knowledge.

The global process of CM-IE will be implemented by sequentially iterating single medium extraction where the output of one process is used as background knowledge of

the next one. The whole extraction process will be iterated until some convergence is reached.

Completely bridging the "semantic gap" among media is clearly beyond what is feasible in the next few years; however, we are confident to reach a considerable progress in this intricate and semantically ambiguous task and to advance the state of the art to the point that it can produce concrete results in real world applications.

2.2.3. Innovation in X-Media: Uncertainty handling in Information Extraction

A direct implication of cross media extraction is that different extractors will have to output evidence and conceptual information with different degrees of confidence. As a matter of fact evidence and extracted information are never completely clear cut and their successful integration requires that uncertainty is maintained and exploited. In X-Media uncertainty will be exploited in form of fuzzy or probabilistic results**.** Such probabilistic output will be managed by the probabilistic reasoning methods defined in Area 1.

2.2.4. Technological Need: Advancement in Single Medium IE

Although the focus of the project is cross media, we believe that further advancements are still necessary for single medium technologies in order to provide new and improved technologies that satisfy the requirements for future information extraction.

2.3. Area 3: infrastructure

A knowledge acquisition, integration and sharing environment will be defined. Since X-Media is an application-oriented integrated project, integration is required on the implementation as well as on the conceptual level. The main outcome of this area of activity will be a a methodology and a technical infrastructure able to deliver knowledge from across media to the knowledge workers, taking into account the complexity of managing different media with different size of data.

2.4. Area 4: application and testing

The technology above will be used to define showcases and prototype applications. Two main testbeds are defined by the two large industrial users (Rolls Royce and Fiat) and other supporting testbeds will be analysed by the technology providers (Quinary, Ontoprise, Solcara and CognIT). System trials with final users will showcase the technology and pave the way to further exploitation.

2.5. The role of ontologies

Ontologies will provide a common vocabulary and a common representation among modules in the architecture, supporting knowledge sharing and reuse and enabling sharing of information. Ontologies, and semantic annotations according to them, are expected to play the key role as glue behind sharing and reusing knowledge processes, supporting integration and fusion of information derived from different extraction systems working over different media. As a consequence, support for management and storage of ontologies will be a key infrastructure component, central in the architecture of the X-Media knowledge sharing platform.

3. Consortium

Partners: University of Sheffield (coordinator, UK), University of Koblenz (D), ITC-Irst (I), University of Ljubljana (Slovenia) University of Freiburg (D), CERTH (G), Labri (F), University of Karlsruhe (D) and the Open University (UK). Quinary (I), Ontoprise (D), Solcara (UK), CognIT (N), Rolls Royce (UK) and Fiat (I).

4. Acnowledgement

X-Media is funded by the European Commission as part of Framework 6 of IST, contract no FP6-26978. Project web page: `www.x-media-project.org`.
For information:
`xmedia-coordinator@dcs.shef.ac.uk`

5. References

S. Bloehdorn, N. Simou, V. Tzouvaras, K. Petridis, S. Handschuh, Y. Avrithis, I. Kompatsiaris, S. Staab and M. G. Strintzis 2004, "Knowledge Representation for Semantic Multimedia Content Analysis and Reasoning", in Proc. of European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT '04), London, U.K.

Ivan Bratko, Dorian Šuc 2003: Learning qualitative models. AI magazine, Vol 24(4), pp 107-119.

Fabio Ciravegna, Sam Chapman, Alexiei Dingli, Yorick Wilks 2004: Learning to Harvest Information for the Semantic Web, Proceedings of the 1st European Semantic Web Symposium, Heraklion, Greece.

Cross-media summarization in a retrieval setting

Byron Georgantopoulos[1,2], Toon Goedeme[3], Stavros Lounis[1],

Harris Papageorgiou[1], Tinne Tuytelaars[3], Luc Van Gool[3]

[1] Institute for Language and Speech Processing / IRIS
6 Artemidos & Epidavrou, Athens, Greece
{byron, slounis, xaris}@ilsp.gr
[2] University of Athens, Department of Informatics & Telecommunications
Panepistimiopoli, Athens, Greece
[3]ESAT-PSI, University of Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{Toon.Goedeme, Tinne.Tuytelaars, Luc.VanGool}@esat.kuleuven.be

Abstract

In this paper, we present work in cross-media summarization which is currently in progress under the REVEAL-THIS IST project. The proposed strategy emphasizes on exploring different ways of synthesizing the most salient elements of the constituent parts of a cross-media object. At the core of our work lies an open, adaptable architecture that decides the way the salient parts are fused in accordance with both the users' interests and digital equipment and the typology and semantic characteristics of the original information. We apply our methodology by showing specific examples of our notion of cross-media fusion and summarization in the domains of broadcast TV news and European Parliament sessions.

1. Introduction

Multimedia information fusion and presentation of personalized summaries in a range of different consumer devices is of growing interest today where available multimedia content increases exponentially. There is a basic need to provide multimedia content systems that will help users keep up with the explosion of digital content scattered over different platforms (radio, satellite TV, Web, etc), different media (speech, text, images, video) and different languages. Moreover, end users require intelligent filtering facilities, embedded in user-friendly, personalized interfaces that will allow them to distinguish crucial content from a plethora of irrelevant information. An answer to this need is provided by organizing, selecting and presenting summarized information in a personalized way.

In this context, state-of-the-art techniques of distributed information retrieval, related to multimedia selection, data fusion and presentation of results coupled with cross-media summarization and hierarchical categorization enable users to effectively search and browse the large amount of content gathered. The process of cross-media summarization consists in constructing summaries by exploiting and analyzing the different media (images, speech, audio, text, etc) that co-exist in the original data stream (Benitez et al., 2002), (Papageorgiou et al., 2005). Although research is still in its infancy, the growing availability of multimedia material along with the technological advances of display capabilities in consumer devices makes cross-media summarization a challenging, yet worthwhile task.

In this paper, we present work in cross-media summarization which is in progress under the REVEAL-THIS project. Our objective is to analyze the audiovisual content, retrieve and rank stories relevant to users' profiles and then provide a personalized multimedia summary highlighting the salient information of the original media.

Our cross-media summarization system combines visual findings (e.g. characteristic scenes extracted from video) with textual ones (summaries of the accompanied transcript along with facts, named entities) and pre-defined templates, which constitute the class of knowledge-intensive approaches that we believe to be more suitable for this task.

The structure of the paper is as follows: Section 2 describes the data collection and the pre-processing applied to it. Section 3 outlines the system architecture and methodology. Section 4 discusses in depth the two basic building blocks: the textual and visual summarization components. Section 5 summarizes what has been achieved so far with an outlook to future work.

2. Data collection & pre-processing

A great range of multimodal (video, audio, text), and multi-source (Sat, TV, Radio, Web) corpora have been collected in REVEAL THIS during the data collection phase. The collection will aid the development and evaluation of an

integrated infrastructure that will allow the user to store, categorize and retrieve multimedia and multi-lingual digital content across different sources with a view to personalize the user experience with these sources. The data collection pertains mainly to the domains of EU politics, News and Travel, and to a lesser extent to Health. The corpus consists of TV and radio recordings and of illustrated documents gathered from the web from February until May 2005. The corpus has the following general characteristics (Pastra & Piperidis 2006):

- It encompasses *multiple domains*, i.e., Politics (EU politics in particular), Travel (travel information for tourists), News (national news in Greece, Britain, Belgium/France and web news on politics, travel and health) and Health (pharmaceutical products, common health issues e.g. alcoholism).
- It is *multimodal*, i.e., it consists of video, audio, text and images.
- It is *multi-source*, i.e., it comes from a variety of sources such as TV channels (satellite and terrestrial), radio broadcasts, and the web.
- It is *multilingual*, i.e., it contains data in more than one language, and in particular in English (EN), Greek (EL) and French (FR).
- It contains multimodal documents of *different genres*, which guarantee a significant variety in modality-specific characteristics, e.g.,

⇨ European Parliament plenary speeches vs. Press-conferences, which stands –among others- as a *read speech vs. spontaneous speech* distinction,

⇨ European Parliament plenary session transcripts vs. web news and online travel guides, which stands as a *formal vs.colloquial* language distinction,

⇨ European Parliament plenary session and politics press conferences vs. travel documentaries, which stands as a *face-rich vs object/scene-rich* distinction for images.

The following table illustrates the size of data collected for the various domains:

Domain	**Size**
Video/Audio Politics	101.5h
News TV	17.5h
News Radio	17.5h
Web News	35K tokens

Table 1: Data collection size

2.1. Pre-processing stages

The audio data have been preprocessed via an existing pipeline of shallow processing tools for English and Greek. This processing infrastructure is based on both machine learning algorithms and rule-based approaches, together with language resources adapted to the needs of specific processing stages. Specifically, the processing tools include automatic speech recognition, tokenization and sentence boundary detection, part-of-speech tagging, lemmatization, chunk and clause recognition, and head identification modules. At a later stage, more sophisticated linguistic processing is performed in order to identify summary-worthy semantic units: terminology, named entities, speakers and facts inside utterances/sentences. The video data are preprocessed by a tool for automatic shot cut detection and keyframe extraction. This segments the video data into smaller units, and selects the most representative still images for further processing (image categorization, face detection and identification). Module inter-communication is facilitated via XML documents adhering to predefined DTDs.

2.2. Terminology

In this paper, the term *document* is used in its broader sense, denoting a video file, an audio stream or a Web document. A *frame* refers to a static image taken from the image track of a video. The term *shot* defines a single camera shot. A *scene* or *story unit* can be seen as a higher hierarchical unit of video data, situated above the frame and shot levels. Unfortunately, there is not a very precise general definition of a scene, the interpretation varies with the domain of video data. In a news bulletin, the parts where different news items are discussed form different stories. In a political debate video, a scene can be interpreted as the part where one speaker is talking.

3. Cross-media Summarization

3.1. Introduction

Cross-media summarization is the process of generating a summary by exploiting the media streams (images, text, audio etc), pertaining to a single document. It principally attempts to combine in a non-linear way different types of features: visual, aural and textual co-existing in the original document, resulting in a condensed representation that retains as much important information as possible. These cross-media summaries can be further personalized based on user profiles and devices. For instance, on mobile platforms like PDA's and mobile phones, the bandwidth and screen size is limited so that the visual part of a summary is restricted to a few representative images. On the contrary, in case of Web access video summaries in different configurations can be simultaneously broadcast to multiple users.

The architecture of our cross-media summarization [CSS] subsystem is depicted in the following diagram:

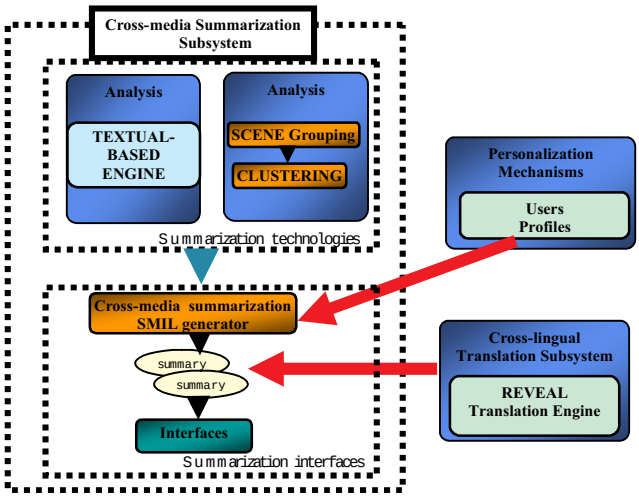

Figure 1: Cross-media Summarization System architecture

The Cross-media Summarization Subsystem consists of three major components: the textual-based summarization component, the visual-based summarization component, and the cross-media summarization component. The role of the textual engine is to produce summaries of speech transcriptions (video/audio sources) and html pages (Web data). The role of the visual engine is to select a variable number of characteristic frames (and corresponding shots) that are most representative for the video. Finally, the cross-media summarizer aims at the fusion of the two previous analyses and the dynamic generation of a self-contained SMIL object, based on predefined templates.

The proposed architecture is designed in order to cover:

- Domains under exploration
- Different channels across countries
- Different genres: documentaries, news
- Different sources: Radio, SAT/TV, Web
- Single and/or multi-document summarization

Our goal was to develop a unified architecture, adaptable to the different parameters stated above. Additionally, we target two types of users/audiences:

1. a generic audience who would like a summarized preview of the current stories (or the stories matching the user's profile)

2. specialized audiences (e.g. journalists) who usually seek for specific information, for instance a timeline of statements made by a particular politician.

In order to fulfill the above requirements, we have implemented two different types of summaries. The first is more stereotyped, focusing on neutral, objective ways of presenting a story, offering a gist. The second involves user interaction, since the user actively influences the direction of the summary, and the system has to respond to the users' interest.

With regard to layout, both types of summaries adopt the SMIL (Synchronized Multimedia Integration Language) markup language [W3C]. SMIL provides a framework for this type of collaboration and presentation of different media that can be properly time coordinated and synchronized. It also caters for customizing the presentation according to the display properties of the user device (mobile, settop box, desktop PC, etc.).

3.2. State-of-the-art

Research reported in the literature conceives the cross-media video summarization task as a collection of audio, visual, and text segments that preserve the content and the structure of the underlying video (e.g. pictorial summary, story boards, and surface summary). (Uchihashi et al., 1999) present methods for automatically creating pictorial summaries of videos using image and audio analysis to find keyframes of relative importance. The output consists of static images linked to the video and the users can interact with it. (Agnihotri et al., 2001) present a surface summarization method for talk shows. They process incoming video, extract and analyze closed caption text, determining the boundaries of program segments as well as commercial breaks. (Aner et al., 2002) construct a highly compact tree-like hierarchical representation of video called mosaic-based scene representation. They cluster scenes using a scene distance measure based on mosaic comparison.

The Informedia project (Christel et al., 2002) is focused on the automatic generation of video summarizations over very large archives of video segments. Their work utilizes the Informedia Project infrastructure, consisting of a huge collection of news, documentaries, lectures and other video genres along with multimedia content analysis tools. They have developed "video collages": presentations of text, images, audio, and video derived from multiple video sources. These collages are created through extracted textual and audio-visual metadata generated from the various Informedia modules (speech recognition, image and language processing) along with manually generated transcripts and closed-captioned text. A video collage, along with providing summaries of multiple video sources, serves as a navigation aid for further exploration.

The Video Scout project explored visualization and summarization of TV content (Zimmerman et al., 2001) in a user-centered approach. By analyzing the visual, audio, and transcript data, Scout can segment and index TV programs, finding and recording specific video clips that match requests in users' profiles. Each program segment offers an image of the

dominant host/guest for the segment, their name, and summarized text.

Kim et al (2003) produce a dynamic video summary by exploiting closed captions data to locate semantically meaningful highlights in a news video and speech signals in an audio stream to align the closed caption data with the video in a time-line.

3.3. Generic summary

Our cross-media summarization system produces two summary types - generic and dynamic - addressing two types of audience. A generic summary aims at encompassing the most salient parts of its input and presenting them in a suitable way. On each occasion, we exploit domain knowledge to guide the summarization process. Currently, two different domains have been focused on: TV news and European Parliament sessions.

3.3.1. TV News

We have been investigating several ways of creating cross-media summaries in the genre of TV broadcast news. The central idea is to unveil the structure of such a program and simulate the actions taken by a news editor when preparing each story. Several videos from the Greek NET (New Hellenic Television) channel have been studied with focus on discovering patterns of presenting a news story. The general structure of a story is as follows:

Initially the newscaster, looking directly at the viewing audience, introduces the topic with a headline along with some explanatory utterances. This usually has a small duration: about half a minute. Then follows a reportage that displays either the person or a location depending on what the story is about. The sentences uttered in the reportage might repeat the words of the newscaster, and/or elaborate on the story. The narration may be interrupted by interviews of people involved in the story or texts (e.g. announcements, statements, tables, charts) or reportage from the field. At the end, the newscaster closes the story with a single sentence and introduces the next one. The order and duration of the story are determined by its importance; top stories are covered first and occupy a bigger part of the news bulletin.

The following figure displays the four basic elements of a TV news story, and the way they interact with each other:

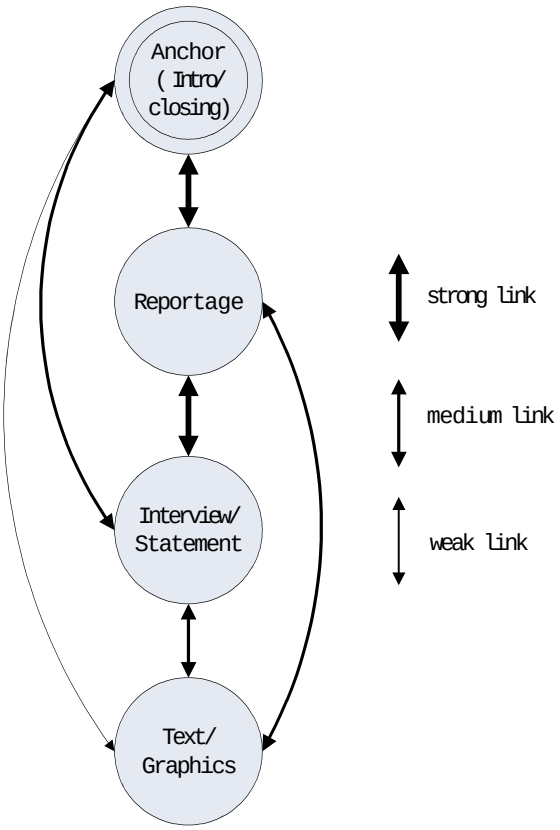

Figure 2: Elements of a TV news story: The stronger the link, the more frequent the transition between them is expected to happen

The study has shown that the essence of the story, with regard to textual elements, is captured by the transcript of the very first speech segments of the news anchor. The person who prepares the script aims at presenting the core without "exhausting" the story, leaving space for the visual reportage. In addition, he/she tries to be as objective as possible, so this introductory transcript fits best for an un-biased textual summary of the story. Moreover, the audio transcript of the news anchor is very reliable since it is read from the auto-cue, without any background noise. On the visual axis however, the image of the anchor does not carry any semantic load. The very first seconds (frames) of the reportage following the introduction are often the most content-bearing and in direct coherence with the newscaster's audio (while drawing the immediate attention of the watcher). The shots that follow tend to analyze the first sentences/images or present subjective views, and they should not contribute to a generic summary. Our hierarchical clustering tool is responsible for selecting the most representative frames of the reportage - possibly with an extra bias towards the first frames.

It has to be stressed that the story type is crucial in determining what elements should be chosen for the cross-media summary. News stories about politics focus on persons and statements, while news about disasters or crimes

emphasize the scene of the event. The system should therefore synthesize summaries according to the story type.

A typical example of a SMIL summary of a news story is illustrated in the following:

```
   <smil
xmlns="http://www.w3.org/2001/SMIL20/Language">
    <head>
     <layout>
      <root-layout width="400" height="300"
background-color="white" />
       <region id="layMainLogo" top="5"
left="108" height="50" width="185" fit="fill" />
       <region id="layDescription" top="60"
left="108" height="65" width="185" fit="fill" />
       <region id="layNewsVisuals" top="130"
left="10" height="165" width="185" fit="fill" />
       <region id="layNewsText" top="130"
left="205" height="165" width="185" />
     </layout>
    </head>
     <body>
     <par>
      <img src="RT-logo.jpg" region="layMainLogo"
dur="60s" />
      <text region="layDescription"
src="data:,Genre:%20News%0AChannel:%20Net%0ADate:
%2028/04/2005%0ATime:%2021:00" dur="60s" />
      <ref src="News_Text.txt"
region="layNewsText" >
       <param name="fontFace" value="arial"/>
       <param name="charset" value="iso-8859-7"/>
      </ref>
      <img src="thnail0385-001.jpg" region=
   "layNewsVisuals" id="NewsButton" dur="60s" />
       <excl dur="indefinite">
       <video src="News_Net_280405_2100.rm"
region="layNewsVisuals"
begin="NewsButton.activateEvent"
clipBegin="25:04.0" clipEnd="25:10.0" />
       </excl>
      </par>
     </body>
   </smil>
```

Figure 3: Code snippet and snapshot from a SMIL TV-News cross-media summary

3.3.2. Europarliament sessions

In the case of European Parliament sessions, most of the time the camera remains static; focused on the face of the specific member who has taken the floor. Occasionally, a general long shot of the chamber might be shown.

In CSS, a generic synopsis shows off the major topic/issues raised during the session presenting the different views/argumentation of the political parties involved in the discussion. Our implementation follows the above presented strategy. It opens with an image indicating the major topics discussed in the session. By clicking on a specific topic, users can watch a short video summary serializing the different political views expressed in the chamber. The significance of a topic/speaker is related to the time period allocated.

3.4. Dynamic summary

Generic summaries provide more or less entrenched representations that address the major topical issues; on the contrary, dynamic summaries offer an intriguing way of zooming at and scrutinising the various facets of the programme users are mostly interested in. In this respect, a dynamic summary is tailored to users' requests facilitating their exploration to an extended depth through alternative navigation paths in the audiovisual content. To this end, we have implemented a dynamic generation of SMIL presentations component providing the necessary functionality allowing users to interactively guide into the content through their own choices.

In case of TV news, each story is decomposed into semantic categories (news anchor, reportage, interview, graphics etc) allowing users to delve into the relevant subsections and distill the crucial points of the particular parts of a story. In case of Euro-parliament, a 3-tier template is foreseen. First, a synopsis of the topics addressed is presented. By clicking on a specific topic, a list of speakers/members with relevant footage is shown. By clicking on a specific speaker, users can hear or/and view a summary of his/her speech in their preferred language.

The following snapshot provides a preview of such a type of summary for the European parliament sessions:

Figure 4: Snapshot of a dynamic topic-based summary

4. Components description

4.1. Textual-based Summarization

4.1.1. Architecture-methodology

Our textual summarization component provides extract-based, single/multi document summaries. For each segment (sentence in case of written texts) a score, indicative of the segments's salience, is calculated as a weighted sum of several features we believe to be important. To this end, we exploit the MEAD summarization environment (Radev et al., 2004) which is highly parameterizable and allows experiments with combinations of different features and methods. Currently selected features involve the position of the segment inside the input stream, the segment length as well as linguistic properties (inclusion and importance of centroids, terms, named entities, facts). The scoring formula is the following:

$$SCORE\ (s) = w_p P_p + w_c P_c + w_T P_T + w_n P_N + w_f P_f$$
$$given\ that\ length(S) \geq P_l$$

P_l is a cut-off filter for segments shorter or equal than a predefined threshold (currently 2).

P_p is the positional score, favouring segments closer to the beginning of the document.

P_c is the centroid score, i.e. inclusion of central words for the document cluster.

P_t is the score for spotted terms

P_n is the score for spotted named entities

P_f is the score for spotted facts. (Facts are defined as the *most significant* events that characterize the data of a specific domain)

and $w_{\{p,c,t,n,f\}}$ are the adjusted feature weights

Next, in order to reduce redundancy, segments are re-scored according to their cosine similarity to already selected ones. If the similarity exceeds a selected threshold of 70% then the segment is dropped from summary.

After the scores refinement has been finished, the top-ranked segments, in their original order, are selected to form the extract (their number determined by a compression factor currently set to 10%).

4.1.2. State of the art

Recent work on textual-based summarization has focused on extracts rather than abstracts, reflecting the difficulty in tackling the problems introduced by the complexities of language (anaphora, polysemy, world knowledge, etc.). Corpus-based systems follow up classical approaches combining the calculation of corpus statistics in a learning framework. (Kupiec et al., 1995) developed a Bayesian classifier, viewing the extraction problem as statistical classification one, while (Barzilay and Elhadad 1997) computed lexical chains, the strongest of which pinpoint to significant sentences. A centroid-based summarization of multiple documents is presented in (Radev et al., 2004), and (Witbrock and Mittal 1999) used statistical models to choose important terminology and their syntactic context in order to produce headline summaries.

More sophisticated NLP techniques try to identify key passages based on the analysis either of word relatedness or of discourse structure. (Salton et al. 1997) try to identify salient passages by calculating the degree of lexical connectedness (e.g. common terms) between a candidate passage and the rest of the text. Other research rewards passages that include topic words, i.e. words that correlate well with the topic of interest to the user or with the general theme of the source text (Strzalkowski et al. 1999).

Alternatively, a summarizer may reward passages that occupy important positions in the discourse structure of the text. (Marcu 1997) derives the rhetorical structure of texts using discourse usages of cue words. (Teufel and Moens 2002) show how particular types of rhetorical relations in scientific journal articles can be reliably identified through the use of classification.

4.2. Visual-based Summarization

4.2.1. Architecture-methodology

Our visual summarization component takes as input a set of *key frames*, i.e. representative frames for the shots of the input video. A shot cut detection algorithm (Osian et al., 2004) is applied which detects shot cuts by analyzing motion-compensated image differences and

characterises each shot (of variable length) with a single key frame.

Taking into consideration that a summary should be able to adapt to users' preferences and situation, a flexible solution is proposed in the form of a *hierarchical summarization tree*. Depending on the desired output, an arbitrary number of representative images can then be extracted at runtime. The hierarchical summarization tree is constructed in an offline phase, as follows. We repeatedly group visually similar key frames into clusters, choosing each time for each cluster as prototype the image with the highest *relevance value* (see below). This results in a hierarchical tree structure, which can be sliced at a certain height, yielding the desired number of clusters, each with its corresponding prototype image.

To compute the visual distance between two images, we use the L2-norm of the distance between the colour histograms of the images, with colours measured in RGB-space and using histograms of 256×3 bins. The visual distance between two clusters is defined as the visual distance between their prototypes. The matrix containing the visual distances between all key frame pairs of one example video fragment is shown in figure 5. Hierarchical clustering of key frames is done based on this distance matrix. The resulting hierarchical summarization tree is shown in figure 6.

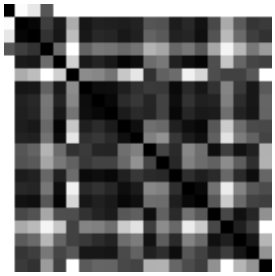

Figure 5: Matrix with visual distances between key shots

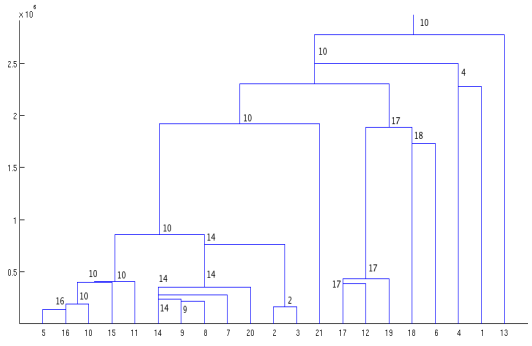

Figure 6: Hierarchical Summarization Tree

When combining two clusters, a new prototype must be chosen, among the prototypes of the merging clusters. We do this selection on the basis of a *relevance value*. It is clear that in a summary only the most relevant key frames must be present. Our relevance value is a combination of two factors: the *time duration* and a *salience measure.*

We assume that, in a general video sequence, longer shots are more relevant and hence their corresponding key frames should be more likely to appear in the final summarization. Indeed, a key frame from a long shot explains a larger part of the video. For instance, during a speech in the European parliament the speaker is shown most of the time, in a few long shots, while the audience is shown only briefly in a few intermittent shots. The time duration of a shot is thus a good measure for the relevance of its key frame. When combining two or more shots into a cluster, time durations are added.

A second component of the relevance value is the *salience*. This measure takes into account that in some cases shots that are shown for a long time in the video segment may be relatively unimportant nevertheless. Examples are shots showing the anchor person of a news broadcast, or the moderator of a parliament session. We can detect this type of shots because they also appear a lot in the video material outside the story segment under scrutiny, for instance during the rest of the news broadcast. The salience measure S_i adds relevance to shots which are repeated frequently inside the story segment, but do not appear often outside it:

$$S_i = \frac{avg_j\, dist(k_i, k_j)}{avg_l\, dist(k_i, k_l)}$$

where k_j is a key frame inside the same story part as k_i, and k_l outside that story part. Histogram distance is denoted as *dist*. This is similar to the *tf-idf* mechanism (text frequency – inverted document frequency) often exploited in text document analysis.

When combining two key frames into one cluster, the relevance value of the selected prototype image, computed as the product of the salience and the time duration $R_i=S_i d_i$, is transferred to the higher level of the tree:

$$R_{n+1} = \sum R_n$$

Figure 6 shows the resulting clustering tree, with at each branch the number of the chosen prototype. Now, this tree can be cut at a chosen height to yield the desired number of representative images for the story segment.

4.2.2. State of the art

Visual-based summarization typically boils down to the selection of the most relevant shots or key frames, either using some highlighting

mechanism or based on clustering. The former approach (Smith and Kanade, 1997; Lienhart et al., 1997; Ma et al., 2002) relies on heuristics about what humans typically consider relevant video fragments, e.g. scenes with a lot of contrast, scenes with close-ups of faces, scenes with a lot of motion activity, etc. The latter approach, on the other hand, studies the (dis)similarity between frames, e.g. using singular value decomposition (Gong and Liu, 2000), non-negative matrix factorization (Cooper and Foote, 2002), or clustering techniques (e.g., Mundur et al., 2005). In REVEAL THIS, we follow the latter clustering-based approach, yet using a heuristics-based weighting scheme.

Visualization of video summaries is typically either in the form of a new, shorter video (so-called video skim or video trailer (Smith and Kanade, 1997; Gong and Liu, 2000) or in the form of a set of representative key frames (Shipman et al., 2003; Wactlar, 2000; Mundur et al., 2005). A video skim is a shortened video that maintains as much semantic content within the desired time constraint as possible. This visualization method is very intuitive and informative, yet difficult to extract and not suited for interaction with the user. Representative key frames, on the other hand, can be spatially organized to reflect the structure or content of the video, and can be made clickable such that a user can zoom in on a specific part of the video to drill deeper. However, the spatio-temporal properties and audio content are lost.

5. Conclusion and future work

In this paper, we have addressed the issue of creating cross-media summaries by synthesizing salient parts of the media elements that comprise the original audiovisual content. We have proposed a template-based technique that takes into account both the characteristics of the media and the users' profiles. With this strategy in mind, two domains (broadcast TV news and European parliament sessions) and two types of users (generic, specialized) have been accounted for. The resulting layout templates, encoded as SMIL presentations, serve both as preview generic summaries and as entry points for further exploration of the digital information content.

With regard to future work we plan to:

(1) Expand our work to the travel documentaries domain.

(2) Integrate translation capabilities in order to provide summaries in the user's preferred language regardless of the language of the original media.

(3) Evaluate our approach by involving user focus groups in the retrieval task of the project.

6. Acknowledgements

Work described in this paper was fully supported by the research project "Retrieval of Video and Language for The Home user in an Information Society" (REVEAL THIS), FP6-IST-511689, funded in the framework of the specific research and technological development programme "Integrating and strengthening the ERA" as well as the Fund for Scientific Research Flanders (FWO) and the Flemish Institute for the Advancement of Science in Industry (IWT).

7. References

Agnihotri L., Devara K., McGee T., and Dimitrova N. (2001). Summarization of Video Programs Based on Closed Captioning*, SPIE Conference on Storage and Retrieval in Media Databases, San Jose, CA, January 2001,* (pp. 599--607).

Aner A., and Kender J. R. (2002). Video Summaries through Mosaic-Based Shot and Scene Clustering. *In Proceedings of the European Conference on Computer Vision.*

Barzilay R., and Noemie E. (1997). Using lexical chains for text summarization. *In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization,* (pp. 10--17). Madrid.

Benitez AB., Chang S-F. (2002), Multimedia Knowledge Integration, Summarization and Evaluation*, In Proceedings of the Third International Workshop on Multimedia Data Mining MDM/KDD 2002* (pp. 39--50).

Christel M.G. et. al. (2002). Evolving Video Skims into Useful Multimedia Abstractions. *In ACM CHI '98,(pp. 171--78), Los Angeles, CA.*

Cooper M. and Foote J (2002). Summarizing video using non-negative similarity matrix factorization. *IEEE Workshop on Multimedia Signal Processing.*

Gong Y. and Liu X. (2000). Video summarization using singular value decomposition. *In Proceedings IEEE Conference on Computer Vision and Pattern Recognition, volume 2,* (pp 174--180).

Kim JG., Chang H., Kang K., Kim M., Kim J., Kim H. (2003). Summarization of News Video and its description for content-based access. *In Interscience Wiley Periodicals* Vol 13 (pp. 267--274).

Kupiec J., Pedersen J., and Francine Chen (1995). A rainable document summarizer*, In SIGIR95 and in Advances in Automatic Text Summarization,* Mani and Maybury, eds., 1999.

Ma YF., Lu L., Zhang HJ., and Li MJ. (2002). A user attention model for video summarization. *ACM Multimedia,* (pp 533--542).

Marcu Daniel (1997). The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. *Ph.D. thesis, University of Toronto, Toronto.*

Mundur P., Rao Y., and Yesha Y. (2005). Keyframe-based video summarization using delaunay clustering*, submitted to the International Journal of Digital Library.*

Osian M., Van Gool L. (2004). Video shot characterization, *Machine Vision and Applications Journal (15), No. 3,* July 2004, (pp 172--177).

Pastra K., Piperidis S. (2006). REVEAL THIS Deliverable 2.3: Collection of training & test data.

Papageorgiou H., Prokopidis P., Protopapas A. and Carayannis G. (2005). Multimedia Indexing and Retrieval Using Natural Language, Speech and Image Processing Methods *In Multimedia Content and the Semantic Web: Methods, Standards and Tools* Wiley (pp 279--297).

Radev D., Hovy E. and McKeown K. (2002). Introduction to the Special Issue on Summarization. *In Computational Linguistics 28(4),* (pp 399--408). MIT Press.

Radev D., Allison T., Blair-Goldensohn S., Blitzer J., Celebi A., Dimitrov S., Drabek E., Hakim A., Lam W., Liu D., Otterbacher J., Qi H., Saggion H., Teufel S., Topper M., Winkel A., Zhang Z. (2004). MEAD - a platform for multidocument multilingual text summarization *In Proceedings of LREC 2004*, Lisbon.

Salton, G., Singhal A., Mitra M., and Buckley C. (1997). Automatic text structuring and summarization. *Information Processing & Management, 33(2)* (pp. 193--207).

Shipman F., Girgensohn A., and Wilcox L. (2003). Generation of interactive multi-level video summaries. *In ACM Multimedia, 2003.*

Smith M.A. and Kanade T. (1997). Video skimming and characterization through the combination of image and language understanding techniques. *In Conference on Computer Vision and Pattern Recognition.*

Strzalkowski, T., Gees S., Wang J., and Bowden W. (1999). A robust practical text summarizer. *In I. Mani and M. T. Maybury, editors, Advances in Automatic Text Summarization. (*pp. 137--154). MIT Press, Cambridge.

Teufel S. and Moens M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *In Computational Linguistics 28(4),* (pp. 409--445).

Wactlar H. (2000). Informedia - search and summarization in the video medium. *In Proceedings of Imagina 2000.*

Witbrock M. and Mittal V. (1999). Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries. *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley,* (pp. 315--316).

Uchihashi S., Foote F., Girgensohn A. and Boreczky J. (1999). Video Manga: Generating semantically meaningful video summaries*, ACM Multimedia 1999,* (pp. 383--392).

Zimmerman, J., Marmaropoulos G., van Heerden, C. (2001). Interface design of Video Scout: A Selection, Recording, and Segmentation System for TVs. *In HCI '01: Proceedings of Ninth International Conference on Human Computer Interaction, Lawrence Erlbaum Associates* (pp. 227--281).

From Media Crossing to Media Mining

Franciska de Jong[*†]

[*]University of Twente, Human Media Interaction group
Enschede, The Netherlands
{fdejong}@ewi.utwente.nl
[†]TNO ICT, Delft, The Netherlands

Abstract

This paper reviews how the concept of Media Crossing has contributed to the advancement of the application domain of information access and explores directions for a future research agenda. These will include themes that could help to broaden the scope and to incorporate the concept of medium-crossing in a more general approach that not only uses combinations of medium-specific processing, but that also exploits more abstract medium-independent representations, partly based on the foundational work on statistical language models for information retrieval. Three examples of successful applications of media crossing will be presented, with a focus on the aspects that could be considered a first step towards a generalized form of media mining.

1. Introduction

According to numerous policy makers, IT gurus and advocates of IT initiatives, there is something to be gained by the possibility for users of all kinds to watch or listen audiovisual content anytime, anywhere and via all kinds of platforms. The kind of information referred to by the various kinds of visionaries, can be characterized as heterogeneous in topic, in format, in language, and collection structure. A wide range of reactions on this vision can be observed. From the side of the sceptics one of the major questions is whether the concept of nomadic consumption of arbitrary types of content really match realistic user needs? But there is also the view that the target is right, but that the vision expressed above is not ambitious enough and that the important trend should not focus primarily on the use case of advanced content consumption, but on the more abstract modeling and representation of information, independent of use cases. This dichotomy could be interpreted as two compatible views, corresponding to two different agendas. On the one hand that of software vendors and researchers looking for opportunities to demonstrate the added value of their work, and on the other hand the long term agenda of research communities?

1.1. From Media Crossing ....

With the growing interest in cross-media functionality (for applications such as indexing, browsing, generation, etc.), the insight in the inherent limitations grows as well. Attempts to build cross-media indexing environments have been around now for more than 10 years. Some of these attempts even have yielded unquestionable successes, either commercially or researchwise.

Any serious review of the added value of tools and techniques that have been prosposed and its potential for future advancement of the fiels of information processing will make clear that in the long run we can not be satisfied with a framework that is characterized by the combination of multiple medium-specific processing. Instead, a future should imply a perspective and a vision on how to build a parameterized representation space that could serve both as a foundational framework for formal information models as well as a conceptual basis for application development.

1.2. ... to Media Mining

This paper will describe and compare three different experimental systems that can be considered successful in their contribution to the research agenda. That is, looking backwards. Another way to characterize them is to call them

relatively simple instantiations of media-crossing applications. In what follows below it will be shown how the three experimental systems each call for a more ambitious and multidisciplinary approach that could help define the next-generation of content consumption tools. Though conceptually and formally these tools can clearly be associated with well-establsihed and familiar concepts such as multimedia retrieval or cross-media indexing, a more careful investigation can reveal that they each have taken one or more small steps into a direction that is far more ambitious. For this reason they could be considered to as examples of 'media mining' *avant la lettre.*

In the next section the concept of media mining will briefly be introduced an even widened. The range of research themes that can be linked to it, will be addressed via the description of three cases in the section 3-5. The concluding section will summarize the main findings.

2. Media Mining

Very soon after the introduction of the notion of data-mining in the nineties, it became clear that *'knowledge discovery'*, a term often used for data mining techniques, was not just applicable to the digging up of more or less hidden data patterns in traditional databases. Via text mining, audio mining en media mining, the concept has been claimed to be applicable to all sorts of digital data. Even the term *reality mining* has been introduced and rapidly taken up: Google produced 3000 hits already in Spring 2005, and over 13000 in April 2006). There seems no limit to the applicability of the concept of reality mining: it is hard to come up with something that is NOT covered by the word 'reality'. But even without trying to tackle this philosophical challenge it is easy to see that there is a magnitude of data almost above imagination and that the diversity of types of 'beings' that can be captured with mining techniques is enormous. In principle knowledge about reality and its inhabitants do not follow, let alone obey the borders of media formats and modalities. As a consequence developing tools that support the crossing of format borders in exploring digital archives can be no more than a very first step towards fully exploiting the treasures out there.

Still, and luckily, tools for simply crossing media in search environments do exist. Already for several few years there is even a search task within the context of the TRECVID that stimulates the exploitation of speech transcripts for the retrieval of video. (Cf. (Smeaton et al., 2003).) Relativizing the advancedness of such tools obliges one even more to assess their value. Of course it would be unjust to deny that they can be extremely useful, and particularly in very specific uses cases with very specific user tasks and data sets, crossing media may be the only way to go. But even if one accepts the always-everything-anywhere-mantra, research agenda's should take a wider perspective and develop frameworks that can accommodate more ambitious functionalities for the tackling of the problem of indexing multifaceted collections. Actual user needs for this domain may be hard to predict, but there is clearly an interest from e.g., content syndication parties, portal owners and content providers.

In addition to more and better medium-specific analysis tools, there is the need for analysis models that deliver features that can be integrated in a medium-independent representation and for search models that can abstract away from media-specific features. Ad hoc merging of ranked lists based on word occurrence statistics and image features can be effective, but the real goal should be transformation and integration of representations into one medium-independent representation. The attempt to use conceptual structures as a representation that is independent of language and modality is one of the most salient features of what has more recently become known as *semantic web*, and it will be crucial to take up lessons learned from that framework, including work on the bridging of the *semantic gap*, or more in particular content-based image analysis. And thirdly the metaphor of *translation* could help to clarify the difference in ambition between media crossing and media mining.

As announced, the next sections we will review three experimental approaches each illustrating the media crossing paradigm in a different way.

- Content reduction
- Content merging
- Content enhancement

3. The Content Reduction Case

In various domains, professional information analysts have to deal with large amounts of information which is refreshed on a daily basis and disseminated via various media types: traditional newspapers, news wires and magazines, internet sites and also television broadcasts via air or cable. Analysis and monitoring of these open news sources, which in some cases are coupled to non-public sources of information, is often crucial for efficient and effective workflow. Various mining tools can support the task of news analysts. In this section the crucial role of content reduction as prerequisite for mining will be discussed.

3.1. Parameterized abstraction and summarization

Multimedia news browsing differs from multimedia retrieval in several respects. A major difference is that browsers are supposed to support information search by offering the user not just access to data, but also one or more perspectives on the available data. (Examples of different perspectives are e.g., chronology and geography.) The more flexible a browser, the more different perspectives. etc. In other words: browsers offer a wider range of access functionalities in an integrated way. Indexing can be the basic functionality, but in addition clustering, classification, extraction of headlines and proper names, and summarization can be exploited to build. If a disclosure system integrates news content from heterogenous sources and in multiple formats, e.g., text, audio and video, a salient feature of the browsing functionality could be that that the content can be accessed at various levels of abstraction. For this purpose a variety of content reduction tools can be applied.

Examples of automatic content reduction techniques are abstraction tools such as classification, redundancy detection (via topic-based clustering), summarization, or the generation of a network representation. There is no absolute criterion for the adequacy of these techniques. Whether or not a classification or a summary of a document is useful may vary per user, per user task, per location, etc. To set the system parameters that eventually determine the output, tools that generate abstractions should collect frequency and co-occurrence data of content features, weigh them against background models, and combine them with information about the user and the context of use.

3.2. Abstraction *versus* reduction

Effective content abstraction is a key feature for improved efficiency of the information analysis task. In this context the notion 'abstraction' refers both to conceptual structure, as well as to (reduced) content size. Both forms may pay a role in the automatic enrichment of content via a multifaceted metadata structure.

Various useful levels of abstraction can be distinguished, as different analysis tasks may impose different requirements on the level of conciseness, and even different perspectives on the content can correspond to different metadata requirements. For example, a proper name index on a cluster gives another perspective than a list of topic labels generated by thesaurus-based classification. Metadata types such as keywords and headlines help the user to select potentially interesting clusters for further inspection. This more detailed inspection step can subsequently involve looking at the titles of the individual news items and reading a multi-document extract. Though content abstraction implies content reduction, the reverse only holds if the reduced representations (e.g., summaries, headlines) are representative from one or more perspectives. This is independent of whether the abstraction techniques yield reduced representations in running text, such as extractive or abstractive summaries, or extracted headlines, or structured objects, such as networks of list of proper names or named entities, topic labels for clusters, list extracted key words, etc.

3.3. Lessons learned from Novalist

At TNO a news browser for heterogenous media archives has been developed which is called Novalist. It aims to facilitate the work of information analysts in the following way: (i) related news stories are clustered to create dossiers, sometimes also called 'threads', (ii) dossiers resulting from clustering are analysed and annotated with several types of metadata, and (iii) a browsing screen provides multiple views on the dossiers and their metadata.

One of the reasons why it offers an interesting case for the perspective of this paper is the content base for which the browser functionality can be demonstrated. The corpus disclosed by the demonstrator system consists of a collection of news items published by a number of major Dutch newspapers and magazines, web crawls, a video corpus of several news magazines and a video archive with all 2001 broadcasts of *NOS Journaal*, the daily news show of the Dutch public TV station. The autocue files for the video archive function as collateral text, i.e., text that is not the primary target of search, but that supports the disclosure of video via the time links to media fragments. The entire collection consists of some 160,000 individual news items from 21 different sources.

Another crucial aspect is the technique known as *document clustering* applied in combination with *topic detection* (also known as topic discovery). The system has to deal with dynamic information, about which no full prior knowledge is available. There is no fixed number of target topics and events types. The system must both discover new events as the incoming stories are processed, and associate incoming stories with the event-based story clusters already created. Clustering is done incrementally: for a new incoming story, the system has to decide instantaneously to which topic cluster the story belongs. Since the clustering algorithms are unsupervised, no training data is needed.

Via document clustering, structure is generated in news streams, while the annotations can be applied as filters: search for relevant items need not to apply on analyzed data but can be limited to relevant subsets of the collection. Novalist supports the fast identification of relevant dossiers during browsing. Dossiers are visualized in a compact overview window with links to a time axis. Additional functionality could consist of the automatic generation of links to related sources, both internal and external.

The screen dump of the end-user application in Figure 1 illustrates the browser functionality. For a detailed explanation of the concept of topic detection and the similarity concept applied in the language modeling approach that is underlying Novalist, and for an overview of the performance evaluation of some components, cf. (Spitters and Kraaij, 2002), (Jong and Kraaij, 2005).

Novalist demonstrates that multiple document abstractions effectively mediate different levels of granularity. The analysis can be performed independently of end-user queries. Due to the emphasis on content preprocessing it can support an entire chain of users: content portals that select subsets of news according to filters to serve their users, professional information analysts that link the portal content to their own repositories, and nomadic news consumers. The media crossing proper is limited to linking audiovisual data via their textual transcriptions to the items in the text-based clusters, and though the source crossing that comes in via the combinations of a wide range of open source titles could be viewed as a distinguishing feature as well. But it is the clustering that brings in the mining perspective, strengthened by the fact that the clustering could easily be extended to numerical data, click patterns to set profiles, etc.

Similar dossier generation applications, with topic clustering as basis and content reduction as additional functionality, could be applied in other domains than news, and/or for other combinations of media. In addition to text from newspapers and autocue files (=teleprompter) files, transcripts of broadcast audio generated with automatic speech recognition (ASR) could be taken into account, Assuming that the material can be properly segmented, such sources could be linked to the related topical clusters. Cf. also section 5.

4. The Content Merging Case

In the spectrum of attempts to exploit textual resources for the disclosure of media archives most attention has gone into the role of speech transcripts and their added value on the retrieval of news video broadcasts. Less attention has been given to the possibility to apply information extraction techniques for video archives. The project MUMIS which was completed in 2003 can be considered to fill this gap. In addition to investigating the possibility for the generation of speech transcripts for sports programmes, it payed attention to the possibility to exploit the redundancy in the target document collection. The system components developed provide an analysis for news, commentaries, structured tables from reports, covering international football

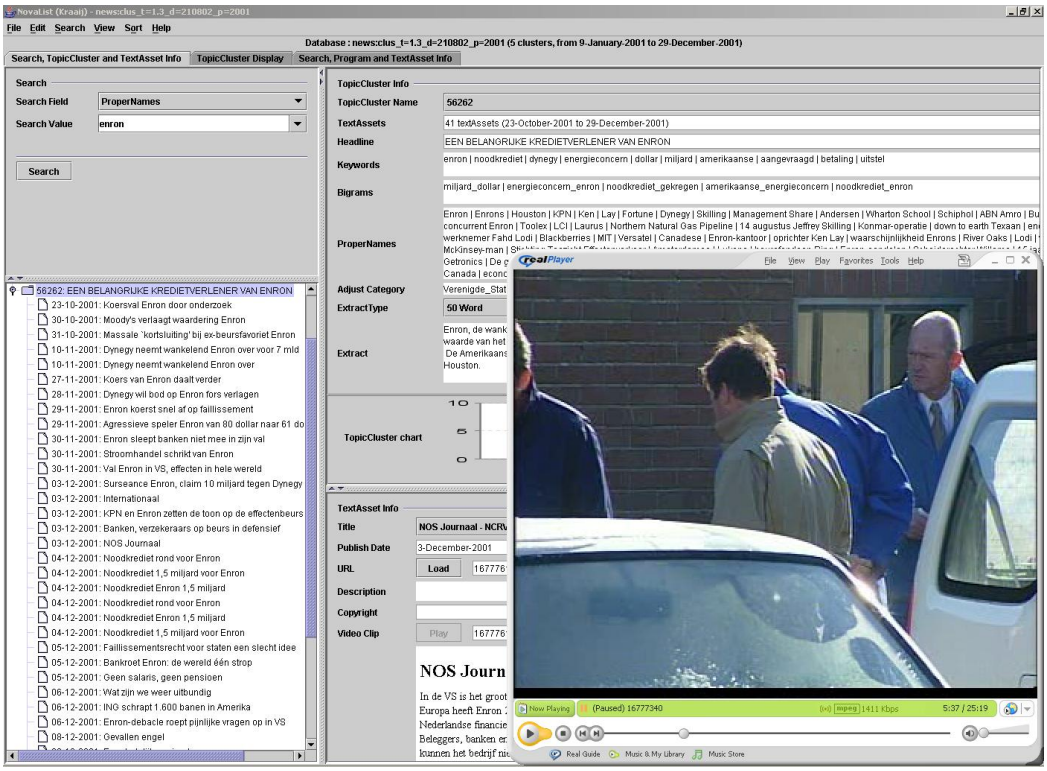

Figure 1: Novalist: browsing multimedia dossiers through associated metadata; query term: 'Enron'

games in multiple languages and multiple modalities, and the resultant data are merged to function as a searchable conceptual knowledge base of all content with links to the timecodes of the corresponding media fragments.

4.1. Multi-source Information Extraction

The combination of research issues central to MUMIS were given in by the characteristic features of the archive studied: football commentaries for an international tournament. What are the crucial features for this type of archive: it consists of video recordings accompanied by several textual sources that cover the same event, but do not necessarily give identical or overlapping information about that event. Cf. Table 1 for examples. Rather the documents should be called parallel, or even less weakly: comparable. As a consequence of the availability of such a combination of sources, the relations between the extraction results can be analysed in order to generate one single merged representation. Errors originating from one of the texts can be removed based on information from the other texts, redundancies can be taken out, and furthermore the merged partial knowledge from separate sources provides a more complete and coherent annotation of the material to be disclosed. The MUMIS disclosure approach can be termed *multi-source information extraction*. As IE modules , such as GATE (cf. (Saggion et al., 2002)) and SCHUG (cf. (Declerck, 2002)), have been applied developed for three languages, it is also a case of *multilingual* IE.

Formal text
England 1 - 0 Germany Shearer (52) Bookings Beckham (42)...
Ticker
41 mins: Beckham is shown a yellow card for retaliating on Ulf Kirsten seconds after he is denied a free-kick.
40' Hoekschop Engeland met David Beckham. Slecht getrapt. Meteen maakt Beckham daarna een fout en krijgt een gele kaart.
Match
David Beckham - a muted force in attack - was shown a yellow card for a late challenge on Kirsten...
Transcription
...it's gonna be a card here for David Beckham it is yellow mmm well again his was the name in the post match headlines...
David Beckham hielt die Sohle noch druber schauen Sie mit dem Hinterteil auch harter Einsatz gegen Kirsten und Collina zeigt ihm Gelb eine der Unarten leider von David Beckham
Beckham met*x Kirsten dat is nou weer dom wat die Beckham doet ja zal ie dat dan nooit leren Kirsten overdrijft nu hoor maar Kirsten gaat 't duel in geeft een zet en dan reageert Beckham op deze manier in ieder geval krijgt ie dan weer geel

Table 1: Different accounts of the same event in different languages

4.2. Improved retrieval via merging

In MUMIS the goal of merging is to yield improved metadata based on information from all documents available from the various sources. As is to be expected, complete recognition of events in natural language sentences is extremely difficult. Often, events will be only partially recognised. The result of merging is one description for all events of a single match that in terms of completeness and correctness has been optimized. A merged annotation is supposed to offer better retrieval results for the multimedia content. The example below, taken from actual results on the Euro 2000 match Netherlands vs. Yugoslavia, gives a rough indication of how merging results in a more complete picture of what happened in the 30-31st minute of the match.

> The IE component recognizes in document A a description of an action of the type SAVE, performed in the 31st minute. In addition, it recognizes the names of two instances of the concept PLAYER: Van der Sar (the Dutch goalkeeper) and Mihajlovic (a Yugoslavian player), but the IE system can not figure out which of these two performed the save.
> In document B IE component recognizes an event of the type FREE-KICK in the 30th minute, and the names of the same two players. It fails to detect which player took the free-kick.
> The fact that the same two players are involved, plus the small difference between the time-stamps, strongly suggests that both descriptions are about the same event. The merger component matches the partial descriptions from A and B, and concludes that it was Mihajlovic who took the free-kick which was followed by a save by Van der Sar.

Figure 2: MUMIS: Example of event merging (informal)

The merging procedure exploits the fact that all available information sources make reference to a time line for the soccer match. This timeline can either be explicit, but sometimes remains implicit. The examples indicates that merging is a combination of three subtasks: time-alignment, unification, and re-ordering. Figure 3 shows the first step for a set of two documents.

As reported in e.g., (Kuper and et al, 2003) experiments have indicated that merging seems to improve retrieval performance.

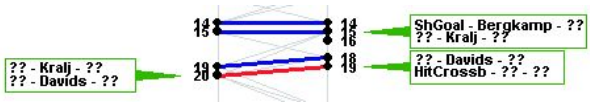

Figure 3: MUMIS: Time-alignment. Vertical lines denote documents, numbers are time stamps; thin lines: possible bindings; thick lines: strongest bindings.

4.3. Lessons learned

MUMIS is clearly a case of information access via media crossing: via time-coded text links into video fragments can be provided. But from the mining perspective it is the merging aspect that is more salient. MUMIS showed that when multiple sources of information are available at the same time, it is likely that the quality and/or reliability will diverge. In combination with IE technology, domain models that link up to templates sets of an interesting size, and reasoning techniques it becomes possible to single out, or even generate optimal representations. Clearly something can be won by having available more, and ideally certified sources. In the MUMIS case the enhancement of content representations via merging exploited primarily text, but integration with all kinds of gazeteers, and repositories of numerical data sources could have additional impact. Especially the scalability of this type of content processing is an issue for further research.

5. The Content Enhancement Case

Most tools that support the searching and browsing of media content in some way or other deploy the concept of *matching*. A representation of the search query is matched onto a representation of the information that is available. In case the formats for query and content representation differ there is a mismatch to be solved. Speech is a carrier of language and therefor a candidate format for content oriented search. But for simple text-based querying of spoken word archives there is an initial mismatch. Only if the either the query or the content has been converted to the format of the other can a matching algorithm be applied. As explained in e.g., (Goldmann and et al, 2005), in the case of spoken document search there are several ways to create matchable representations. Most common nowadays are ap-

proaches that seek to preprocess the audio signal and to apply automatic speech recognition (ASR) to produce a textual transcript. The transcripts can be the basis for a time-coded index that can support the search for audio fragment. This technique has been widely applied for various languages, and in absence of generally applicable tools for video analysis, speech has become the number one entry to large volumes of video content. A clear and widely reported example of media crossing.

Less widespread is work on the exploitation of textual content to complement the speech transcripts. The first role of text is of course to feed the language models needed to build large vocabulary speech recognition. But in addition there are other roles. One of them is that parallel or comparable texts can help to decrease or even eliminate the word error rate of ASR systems. If a manually produced transcript is available, e.g., the minutes of parliamentary sessions, or subtitles for broadcast data, the two parallel texts could be aligned. The timecodes of the ASR transcripts could then be fed into the manually produced text, which in turn could be used for user feedback during search. Also the particular the out-of-vocabulary rate could be decreased: if a (non-perfect) ASR transcript is used a the basis for a search of related text, and the terms referring to named entities in the most similar texts are fed into the language models, a second run of the ASR could yield improved recognition results.

Finally there is of course also the possibility to use an audio fragment as a query for textual documents. An obvious application domain for this option is, again, news. But it works also in other domains than news, e.g., oral history archives, meeting or lecture recordings, digital story telling, etc. In combination with e.g., manually generated minutes, historical studies, policy plans, etc., ASR can provide a welcome or even required additional perspective on the recorded A/V content. Another option to consider would be link generation to geographical data (maps) or other kinds of repositories with a non-linear structure.

Initial experiments with the exploitation of ASR generated transcripts for the search of related text in the cultural heritage domain has been reported in e.g., (Morang et al., 2005).

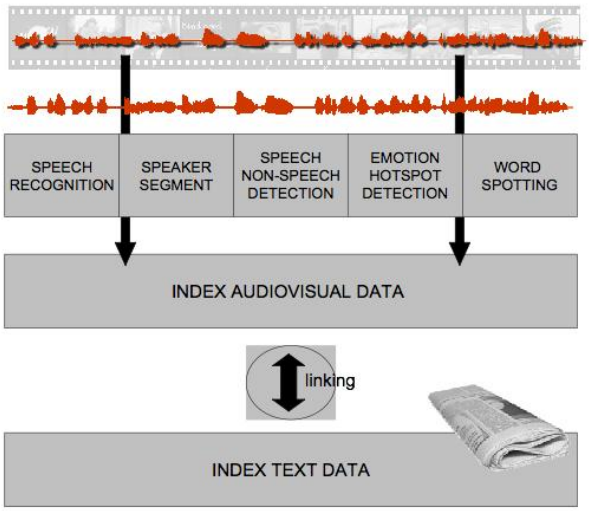

Figure 4: Linking audio to text

6. Conclusions and recommendations

The primary scope of the systems described in this paper are aiming at information access. Pattern detection, which is at the heart of all kinds of mining, is in principle applicable in a much wider range of domains, so the mining concept has a broader scope. Even if we exclude purely numerical data patterns, it could cover diverge topics such as interaction patterns, advertisement campaigns, opinions, moods on certain topics, synonyms, etc. The majority of these examples will serve the needs of professional users first of all (including computational linguists). But as noted above, one of the crucial differences between media mining and media crossing, is the relative independence of use cases for mining. Only after mining tools have been applied the usability of the patterns become evident. In the domain of NLP this holds for example for web-based Question Answering, a search application heavily uses precompiled lists of facts mined from the web.

For development of a research agenda addressing issues that can help building next-generation of media access tools that do not just cross modalities and formats, but that can generate medium-neutral, or in some cases normalized representations, I see a number important sources of inspiration. First there is the analysis of content in term of ontologies, a theme that is getting wide attention under the label of 'semantic web

technology', and that will undoubtedly help to advance the field of media crossing. Second, there is the field of machine translation. Some decades ago it introduced the concepts of interlingual and language-independent meaning representation. The former could be taken as the counterpart of media-crossing approaches that take textual representations as interlingua. The later is close to the generalized interpretation that could help to establish the more general framework and that in this paper has been linked to the concept of media mining. In particular if the number of modalities and formats to be covered increases, the need for medium-independent intermediate representation will increase as well.

Interestingly enough a basis for generalizing approaches to the processing of multimodal information could come from the language (*sic*) modeling approach to Information Retrieval. Cf. (Ponte and Croft, 1998), (Hiemstra, 1998). The language modeling approach to retrieval is based on the philosophy that the language in a relevant document follows the same distribution as that in the query. This same philosophy can also be applied to content-based image and video retrieval, where the only difference lies in the definition of language. Content-based image retrieval systems are usually based on a vector-space model (Smeulders et al., 2000). Collection images are represented as vectors in a high-dimensional feature space, and similarity between images is estimated by a distance metric defined on this space. A drawback of this model is that it is far from obvious how to combine similarity in one representation (e.g., color histograms) with that of another one (e.g., texture); especially when a combination is concerned of different modalities, such as video shots represented by their visual, audio, and speech content. Recently, several attempts have been made to investigate whether discrete bag-of-words models (as used in full-text retrieval) can be also developed and effectively implemented for visual content and the so-called 'visual words' it consists of. Cf. e.g., (Squire et al., 2000), (Jin and Hauptmann., 2002), (Vries and Westerveld, 2004).

With the availability of more abstract models, media formats can not only be crossed, but eventually really integrated, and the content they store can be explored in a genuinely general way. The approaches mentioned rely on heavy processing power, but with the likely advances in grid-processing there is no reason to doubt that the required capacity will become available. While seeking on the one hand collaboration with communities involved in foundational research on abstract content models, language engineers should on the other hand continue to carefully design the media-crossing applications, and to apply all the heuristics they can get hold of to improve their tools and to demonstrate the added value of language processing for the disclosure of information.

Acknowledgements

This work was partly supported by the Dutch bsik-programmes MultimediaN (www.multimedian.nl) and BioRange (www.nbic.nl), and the EU projects AMI (IST-FP6-506811), MESH (IST-FP6-027685), and MediaCampaign (IST-PF6-027413).

References

Declerck, T., 2002. A set of tools for integrtaing linguistic and non-linguistic information. In *Proceedings of SAAKM 2002, ECAI-2002*. Lyon.

Goldmann, J. and S. Renals et al, 2005. Accessing the spoken word. *International Journal on Digital Libraries*.

Hiemstra, D., 1998. A linguistically motivated probabilistic model of information retrieval. In *Proceedings of ECDL 1998*.

Jin, R. and A.G. Hauptmann., 2002. Using a probabilistic source model for comparing images. In *Proceedings International Conference on Image Processing (ICIP02)*. Rochester, NY.

Jong, F.M.G. de and W. Kraaij, 2005. Content reduction for cross-media browsing. In H. Saggion and J.-L. Minel (eds.), *RANLP workshop 'Crossing Barriers in Text Summarization Reserach*. Borovets, Bulgaria.

Kuper, J. and H. Saggion et al, 2003. Intelligent multimedia indexing and retrieval through multi-source information extraction and merging. In *18th International Joint Conference of Artificial Intelligence (IJCAI)*. Acapulco, Mexico.

Morang, J., F.M.G. de Jong, R.J.F. Ordelman, and A.J. van Hessen, 2005. Infolink: analysis of dutch broadcast news and cross-media browsing. In *Proceedings of CBMI 2005*. Amsterdam.

Ponte, J. and W. Croft, 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval (SIGIR'98)*.

Saggion, H., H. Cunningham, D. Maynard, K. Bontcheva, O. Hamza, C. Ursu, and Y. Wilks, 2002. Extracting information for automatic indexing of multimedia material. In *Proceedings of LREC 2002*.

Smeaton, A.F, W. Kraaij, and P. Over, 2003. Trecvid - an overview. In *Proceedings of TRECVID 2003*. USA: NIST.

Smeulders, A.W.M., M. Worring, S. Santini, and R. Jain A. Gupta, 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

Spitters, M. and W. Kraaij, 2002. Unsupervised clustering in multilingual news streams. *Proceedings of the LREC 2002 workshop: Event Modelling for Multilingual Document Linking*:42–46.

Squire, D. McG., W. Muller, H. Muller, and T. Pun, 2000. Content-based query of image databases: inspirations from text retrieval. In *Pattern Recognition Letters*, volume 21.

Vries, A. de and T. Westerveld, 2004. A comparison of continuous vs. discrete image models for probabilistic image and video retrieval. In *Proceedings International Conference on Image Processing (ICIP'04)*.