Multimedia Search Without Visual Analysis: The Value of Linguistic and Contextual Information

Franciska M. G. de Jong, Thijs Westerveld, and Arjen P. de Vries

Abstract—This paper addresses the focus of this special issue by analyzing the potential contribution of linguistic content and other nonimage aspects to the processing of audiovisual data. It summarizes the various ways in which linguistic content analysis contributes to enhancing the semantic annotation of multimedia content, and, as a consequence, to improving the effectiveness of conceptual media access tools. A number of techniques are presented, including the time-alignment of textual resources, audio and speech processing, content reduction and reasoning tools, and the exploitation of surface features.

Index Terms—Context, language processing, multimedia search, semantic metadata, speech and audio analysis, surface features.

I. INTRODUCTION

TO tackle the challenge handed in by ever increasing creation and storage of multimedia content on the one hand, and the promising steps made in multimedia analysis on the other hand, it is tempting to put the focus on the advancement of image analysis and its integration with recently gained insights from relevant domains such as knowledge extraction and semantic web technology. It would, however, be a definite mistake to ignore contextual content and, in particular, the available insights in the successful role that can be played by the modality for which matured analysis frameworks exist: natural language.

As is widely acknowledged, the exploitation of linguistic content in multimedia archives can boost the accessibility of multimedia archives enormously. Already in 1995, Brown *et al.* [1] demonstrated the use of subtitling information for retrieval of broadcast news videos, and in the context of TRECVID the annual video retrieval benchmark event that is organised by the National Institute of Standards and Technology (NIST),[1] a

common feature of the best performing video retrieval systems has been for several years the exploitation of speech transcripts [2]–[4]. Of course the added value of linguistic data is limited to video data containing textual and/or spoken content, or to video content with links to related textual documents, e.g., subtitles, generated transcripts etc. But when available, the exploitation of context and linguistic content can play a crucial role in the generation and exploitation of rich metadata, and therefore in bridging the semantic gap between media features and user needs.

This paper presents an overview of technologies that support media retrieval in a way that is complementary to visual analysis. The paper aims to emphasize that technology based on linguistic and contextual sources can bring more than basic keyword search in collateral text and we showcase a number of techniques aimed at deriving additional metadata from these resources. We illustrate how linguistic, knowledge-based, and visual resources can be combined to detect high-level concepts, and how contextual information can improve retrieval results obtained via visual analysis. Furthermore, it is discussed that the linguistic and contextual techniques may be used to create a corpus with rich semantics. Such a corpus can be used as training material for the development of visual analysis tools based on semantic concept detection.

As said, this work does not detail the analysis options for all modalities in equal manner. Rather, it focuses on the full potential of metadata automatically derived from nonvisual information in a multimedia retrieval setting. But it also underlines the importance of combining approaches originating from fields as disparate as image processing, knowledge engineering, information retrieval, semantic analysis and artificial intelligence, and it shows how the presented techniques and resources can be used as a stepping stone into fully developed and integrated multimedia access technology. By taking this perspective it aims at both broadening and deepening the coverage of this special issue.

The remainder of this paper is organized as follows. After a brief overview of relevant developments in the field of conceptual search for media archives with content in multiple media types and modalities in Section II, Section III explores some methods that deal with the exploitation of linguistic content in, or attached to multimedia databases. The next section continues with methods for using and enhancing speech recognition transcripts. Next, Section IV discusses methods of content reduction. We discuss methods for abstracting from document representations to cluster representations and methods for merging different views on the same document. Section V discusses how characteristics at the surface of documents (like production date

F. de Jong is with the University of Twente, 7500 AE Enchede, The Netherlands, and with The Netherlands Organisation for Applied Scientific Research (TNO), NL-2600 GB Delft, The Netherlands.

T. Westerveld is with Center for Mathematics and Computer Science (CWI), NL-1098 SJ Amsterdam, The Netherlands.

A. de Vries is with Center for Mathematics and Computer Science (CWI), Amsterdam NL-1098 SJ Amsterdam, The Netherlands, and also with the Technical University of Delft, NL-2600 GA Delft, The Netherlands (e-mail: arjen@acm.org).

[1]NIST, Washington, CA, http://www.nist.gov/

or size) can be used to improve retrieval results. Section VI addresses the implication of integrating multiple annotations, and the paper ends with our main conclusions and a look at the multimedia retrieval research agenda for the near future.

II. CONCEPTUAL SEARCH AND MULTIMEDIA ANNOTATION

The semantic gap between user needs and content features is as old as the concept of archiving itself. The traditional approach towards the creation of indexes is to rely on manual annotation with controlled vocabulary *textual* index terms. Shared vocabulary between annotators and users is a powerful instrument that circumvents gap bridging, but it requires a time and labor consuming effort. With the emergence of digital archiving this approach is still widely in use, and for many archiving institutes the creation of manually generated metadata is and will be an important part of the daily work.

The automation of metadata generation is often considered as something that can enhance the existing process rather than replace it. The eventual metadata available in innovative systems is therefore likely to be a combination of highly reliable and conceptually rich annotations, and (semi)automatically generated metadata. The numerous semantic web initiatives will contribute to the establishment and convergence of common terminology and formalisms for the capture and integration of higher level semantic layers in text, image, audio and speech. Innovations for media annotation are to be expected from various directions and fields.

Early image retrieval work in the nineties was based on matching via low-level features (see [5] for an overview. More recently, attempts have been taken up to combine low-level feature analysis with knowledge-based approaches, and to exploit the outcome for image and video retrieval via higher level content annotations. A good example is the automatic detection of image concepts, as widely studied in the context of TRECVID's high-level feature detection task [4] and the LSCOM project [6]. Other examples aiming at the support of conceptual search include the many recent approaches to automatically annotate images based on learned relations between (low-level) visual features and textual terms [7]–[9]. Among the most promising alternatives is the approach taken by [10] to link the annotations generated automatically via visual concept detectors to concepts from linguistically motivated semantic networks such as WordNet (cf. [11]), which allows the use of textual queries as a starting point for matching. The promising results on TRECVID data illustrate the advancements made for query-by-concept in comparison to the more "traditional" query-by-example or query-by-keyword methods, and how the integration of knowledge engineering, semantic analysis and image processing is maturing.

The observed added value of linguistic resources in building conceptual search tools for digital media archives stems in part from the fact that natural language expressions are by nature closer to the level of concepts than low-level image features. In addition, texts and transcripts can be processed with state-of-the-art methods for natural language processing that inherently address the semantic layers and that facilitate more rich and natural information access. These include named-entity detection, information extraction, automatic topic classification,

translation, summarization, novelty detection. For an overview, cf. [12]. Also, ontology-driven and concept-based variants of text-analysis have become widely available [13]–[15]. This all can be seen as reinforcing the large amounts of work on conceptual search in the IR field, partly originating from before the Semantic Web era, bringing in more densely populated concept spaces.

III. EXPLOITING COLLATERAL TEXT

Depending on the resources available within an organization that administers a media collection, the amount of detail of the metadata and their characteristics may vary. Large national audiovisual institutions annotate at least (descriptive metadata): titles, dates and short content descriptions. However, many multimedia archiving institutes do not have the resources to apply even some basic form of archiving. To still allow the conceptual querying of video content, collateral textual resources with content related to the A/V items can be exploited. They can be either available because they play a role in the production or broadcast process, or they can be generated via speech recognition.

A. Speech-Based Indexing

A well known example of a collateral textual resource is subtitling information for the hearing-impaired (e.g., CEEFAX pages 888 in the U.K.) that is available for the majority of contemporary broadcast items, and in any case for news programs. Subtitles contain a nearly complete transcription of the words spoken in video items and can easily be linked to the video by using the time-stamps that come with the subtitles. An early proof of concept for the exploitation of subtitles for indexing was delivered in the nineties by the IST project Pop-Eye [16]. Textual sources that can play a similar role are teleprompter files (also referred to as auto-cues): the texts read from screen by an anchor person. In all these examples, the time stamps in these sources are crucial for the creation of a textual index into video.

Recent years have shown that large vocabulary speech recognition can successfully be deployed for creating multimedia annotations allowing the conceptual querying of video content (e.g., [17]). In particular, this holds for the broadcast news domain, where the collection of training data for creating a speech recognition system is relatively easy. For the broadcast news domain, speech transcripts approximate the quality of manual transcripts, at least for a few languages, including American English. It should be noted, however, that in domains other than broadcast news, and for most nonEnglish languages, a similar recognition performance is usually harder to obtain [18]. Complicating factors include the lack of domain-specific training data, large variability in audio quality, speech characteristics and topics being addressed.

B. Text Alignment

Relatively limitedly referenced is the exploitation potential for external textual content to complement speech transcripts. In automatic speech recognition (ASR) development, the first role of text is of course to feed the language models that determine which constructions and vocabulary are covered by the
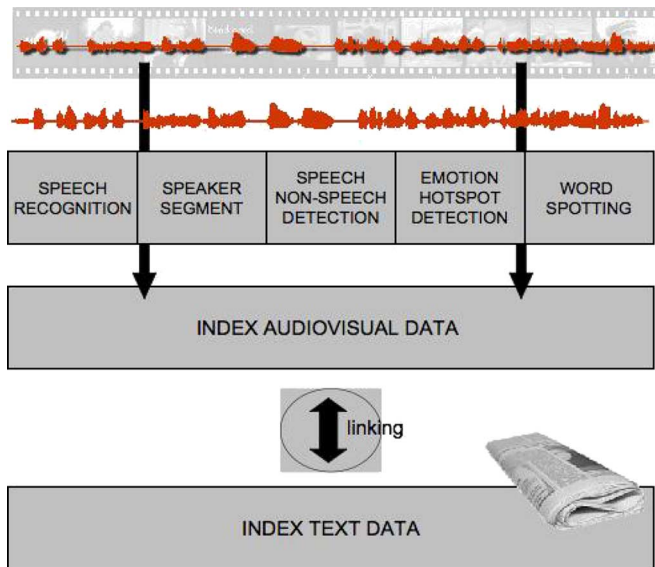
Fig. 1. Linking audio to text.

recognition engine. But when offline language model generation has been completed, whether or not in an iterative process, textual content has still a role to play, both during and after the speech recognition process. With relatively simple alignment techniques, parallel or comparable texts can help to reduce the word error rate of ASR systems and thereby increase the retrieval performance.

Reversely, ASR transcripts can help to enhance manually generated transcriptions. In manually generated annotations or transcriptions that describe a media fragment, the available time-labels are not always fully reliable and outside the news broadcast domain they will often be absent. Examples of such text sources are minutes of meetings or written versions of lectures and speeches. In such cases, the text files can be synchronized with imperfect speech transcripts [19]. This technique for transcript enhancement can be applied independent of a search action, and is generally known as time-alignment.

C. Cross-Collection Browsing

To optimize indexing of audio and/or video, ideally one could just synchronize audiovisual material with content that approximates the speech part. One way to take this simple form of media-crossing a step further is to exploit *any* collateral textual resource, or even better: any kind of textual resource that is accessible, including open source titles and proprietary data (e.g., trusted web pages and newspaper articles). In the context of meetings for example, usually an agenda, documents on agenda topics and CVs of meeting participants can be obtained and linked to the media repository. Based on a text fragment one could search for transcripts with similar or related content, and use the transcripts to jump to the corresponding media fragments (Fig. 1).

In the context of TRECVID, a first step is taken by the National University of Singapore [20]. They use a comparable corpus of collected newspaper articles and Google news to expand their queries. An initial query is fired against this comparable corpus to identify persons, locations and points in

time related to the event searched. These entities are then used to match against the entities identified in the TRECVID video corpus.

Crossing media boundaries is not limited to searching audio with text. There is also the reverse option to take 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, e.g., oral history archives, meeting or lecture recordings, digital story telling, etc.

The examples illustrate how cross-media search—textual querying of individual multimedia documents can be turned into something more ambitious: the mining of a truly multimedia and even distributed database. This concept gains potential impact if it also accommodates a wide range of metadata. In Section IV, an example of such a multifaceted media browser will be described.

IV. CONTENT REDUCTION

Information extraction (IE) techniques for both image and textual data are highly fit to be coupled to summarization tools for the generation of more concise content presentation [21]. This section illustrates how on top of IE, topic classification and reasoning can be applied to generate and merge multimedia annotations for document clusters. We consider this the ultimate proof for the claim that the described analysis yield annotations that are inherently semantic. Two cases of content reduction will be described. One is based on abstraction, the other on merging multiple event descriptions into a single representation.

A. Content Reduction Via Abstraction

Effective content abstraction is a key feature for improved efficiency of the information analysis task [22]. In the context of this paper, the notion "abstraction" refers both to conceptual structure, as well as to (reduced) content size. Both forms may play 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, may 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 multidocument extract.

An example is the Novalist news browser for heterogeneous media archives developed by TNO [23], [24]. It aims to facilitate the work of information analysts in the following ways: 1) related news stories are clustered to create dossiers, sometimes also called "threads;" 2) dossiers resulting from clustering are analyzed and annotated with several types of metadata at different levels of abstraction; and 3) a browsing screen provides multiple views on the dossiers and their metadata.

The corpus used in a case study consisted of a collection of news items published by a number of major Dutch newspapers
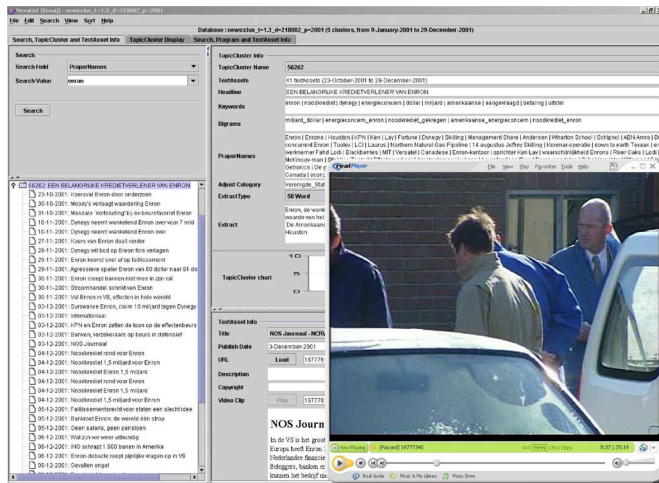
Fig. 2. Novalist: browsing multimedia dossiers through associated metadata; query term: "Enron."

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.

Fig. 3. MUMIS: example of event merging (informal).

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 auto-cue files for the video archive function as collateral text. The entire collection consists of some 160 000 individual news items from 21 different sources.

Via document clustering, structure is generated in news streams, while the annotations can be applied as filters on the clusters: search for relevant items 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 Fig. 2 illustrates the browser functionality.

B. Content Reduction Via Merging

The MUMIS system described in [25] provides a different type of content reduction. Here, a number of components provide an analysis for news, commentaries, structured tables from reports, covering international football 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.

The archive consists of video recordings of the football matches of an international tournament, plus commentaries. A typical feature for this type of archive is that the video content is accompanied by several textual sources that cover the same event, but do not necessarily give identical or overlapping information about that event. Via IE tool sets such as GATE (cf. [26]), each commentary is analyzed and for each cluster of commentaries (i.e., the entire set of commentaries for one match) the resulting data are compared and merged into a single 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 example in Fig. 3, taken from actual results on the Euro 2000 match Netherlands versus Yugoslavia, gives a rough indication of how merging results in improved metadata: a more complete formal description of what happened in the 30-31st minute of the match. The link between the timecodes of the various sources is kept to ensure that the corresponding media fragments can be played.

The merging procedure exploits the fact that all available information sources make reference to a time line for the football match. This timeline can be explicit, but sometimes remains implicit. As the examples indicate, merging is a combination of three subtasks: time-alignment, unification, and re-ordering. Experiments have indicated that merging can indeed improve retrieval performance.

V. EXPLOITING SURFACE FEATURES

Apart from textual data, another nonvisual information source can be exploited: surface features. Surface features are properties at the surface of (multimedia) documents; they do not describe content. Examples are the length of a document, references to the document's location, and the production date, but also speech features e.g., speaker age, speaker gender. Although these features do not directly relate to the document's coverage, they can be valuable additional sources of information in a retrieval or recognition setting. Surface feature evidence can be combined with traditional content-based or text-based retrieval scores, to improve retrieval results. In text retrieval for example, the length of a document is often used as an indicator of relevance (longer documents are more likely to be relevant). Similarly, on the web, the number of hyperlinks pointing to a document is an indicator of the importance of a document [27], [28]. Also, in the design of video *browsing* interfaces, the importance of surface features, like the temporal structure of video, is well-known, see for example [29]. In video search systems however, surface features are mostly ignored. A recent analysis of the correlation between these types of features and the relevance of video shots in the context of TRECVID shows that in fact they can be quite useful [30].

Fig. 4. *Floods*. Relevant images are directly related to a news event (five consecutive shots are shown).

Fig. 5. *Umbrellas*. Relevant images cluster with news events (five consecutive shots are shown).
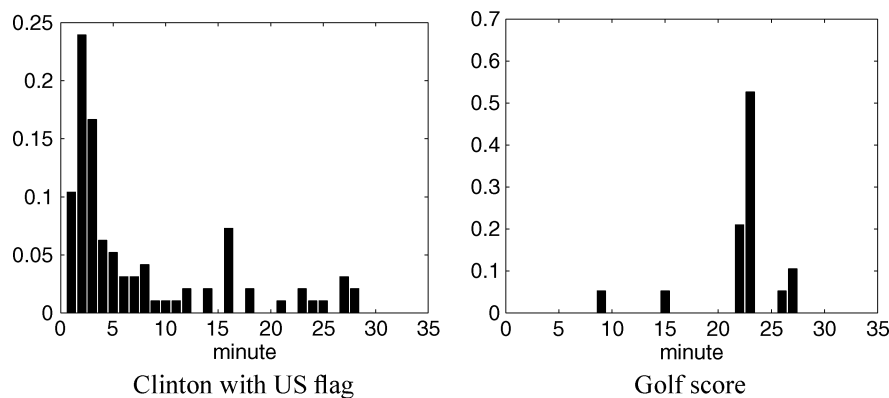

Clinton with US flag

Golf score

Fig. 6. Distribution over minutes of shots relevant to *Clinton in front of the US flag* (left) and *Golf ball going into the hole* (right).

A. Video Surface Features

A first observation is that relevant shots tend to cluster: when a shot is relevant for a given topic, it is likely that its neighboring shots are relevant as well. An explanation for this is that news broadcasts are organized in stories. When a query is directly related to a news event, it is obvious that all, or at least many, shots from the story are relevant (see Fig. 4). Interestingly, the same appears to hold for queries asking for visual items not directly related to news events. When one of the shots shows a relevant item, it is likely to re-appear in other shots of the same story. For example, a news story that happens to be shot on a rainy day, is probably a good source of information when one is searching for shots with umbrellas (see Fig. 5). Analysis of the local distribution of relevant shots in the TRECVID 2004 corpus showed that almost half of the relevant shots has a relevant neighbor.

Another source of information is the metadata associated with videos (e.g., the date of broadcast or the name of the broadcaster) or individual shots (e.g., the duration of the shot or its time within the video). Although these features do not tell us anything about the content of the data, they can help to locate relevant information. For example in a news collection, the distribution of shots related to a particular event are not distributed evenly over the year. e.g., floods tend to occur in the rainy season and for shots regarding Clinton's impeachment trials one should focus on material broadcast in early 1999. Similarly, different broadcasters cater for a different public and hence they show different kinds of material. At a finer granularity, one could exploit the regular patterns of for example news broadcasts, they usually open with major news events (war, politics, or disasters) while sports is usually shown at the end of the news show. See, for example, the different distributions over time for shots showing *Clinton in front of the US flag* and shots showing *a golf ball going into the hole* in Fig. 6.

B. Contextual Speech Features

Another type of surface feature can come from the audio channel. There is more information in speech than words alone. Speaker characteristics can be extracted from the speech (speaker's voice, word usage, syntax) as well, and may serve as an additional level of information. It will not always be useful as a source for primary indexing or cross-linking, but speaker turns and speaker identification are useful features for search and they can add structure for browsing. Speaker features help to create context for content. Therefore, annotations of this type are especially beneficial to support professional information analysts exploring cultural heritage collections. Historians for example, may be interested both in the exact words that were spoken, but also in the speaker's profile. The latter may partly be reconstructed using speaker characteristics such as accent, age, gender, speaking behavior and even emotion and cognitive state. Vice versa, contextual information of this kind may improve speech recognition results.

Research has also been directed to extract features from multiple voices, for example emotional features in order to detect so called "hot spots" in collections. Typical examples of such "hot spot" detections are the cheering of a crowd in a sports game, or laughter in the context of meetings (cf. [31]).

C. Use Cases for Search With Surface Features

For some types of queries, we can expect a user to guide the system in exploiting the surface features. For example, a user familiar with news broadcasts may tell a retrieval system to focus on shots that start around the 20th minute of the broadcast when searching for sports related items. For other types of queries (like the umbrella example), this is perhaps less obvious. An alternative is to obtain their characteristics through relevance feedback. Once a user has pointed out some relevant shots, a system may analyze the surface features of these shots to find patterns. Based on the results of the analysis the search space could be reduced to data with specific surface feature characteristics, or shots with these characteristics could be pushed towards the top of the results list.

VI. TOWARDS INTEGRATION OF MULTIPLE ANNOTATIONS

The above sections illustrate that the added value of nonvisual metadata for multimedia retrieval seems beyond doubt, whereas contextual metadata seem to be effective in particular in combination with other annotations. Smart integration of different annotation layers is likely to increase retrieval effectiveness in many scenarios, as the following examples may illustrate.

In [32], it was shown that there is a near-linear relationship between ASR performance and retrieval accuracy but that the IR performance degradation slope is relatively gentle. When recognition performance remains within certain boundaries (an ASR performance of 50% word-error-rate is typically regarded as a lower bound for successful retrieval) the damage in terms of retrieval performance may be acceptable, especially when no other means (metadata) are available for searching. More recently in 2005 the Cross-Language Evaluation Forum (CLEF) has started a speech retrieval track on spoken interviews [33]. With a state-of-the-art ASR system, transcriptions of 589 hours of spontaneous conversational speech were created with a recognition error between 20% and 30%. These figures are comparable to the error rates obtained at the TREC-SDR track for broadcast news of 1999. The recent CLEF IR performance figures for conversational speech seem lower, although it was reported to be sufficiently accurate to be useful for some purposes. It was also shown that speech recognition transcripts of around 30% WER could not beat IR performance for available systems that used nonspeech metadata as primary source for searching. Using both speech transcripts and manually generated metadata gave the best retrieval results.

For audiovisual oral history collections, queries of the following kind are conceivable: "give me fragments with male, native Dutch speakers talking about...without expressing any emotions." All queried aspects pertain to a different conceptual layer. Such multiaspect queries raise questions on the efficient storage of multiple annotation layers, and the integration of retrieval scores for them, the selection of appropriate segments boundaries, etc. These problems are studied by us within

the MultimediaN project. Hereto, we integrate XML database technology with information retrieval [34]. Ongoing and future research will have to make clear whether this can facilitate the combination of multiple probabilistic models for more heterogenous archives, and whether scalability issues can be solved.

VII. SUMMARY AND CONCLUSION

This paper makes a case for a continued exploration of the potential role of nonvisual content and contextual metadata for the purpose of multimedia analysis and access.

Several relevant approach to multimedia retrieval based on nonvisual features ave been presented. The state-of-the-art in simple keyword search in ASR transcripts has been revisited. Note that this is sometimes seen as "low hanging fruit" [35], but still not widely in use [36]. Also more challenging approaches have been described, among which the use of parallel and comparable corpora for ASR performance improvement, techniques for cross-media document clustering, cluster annotation, exploitation of contextual features for narrowing the search space or improving retrieval results, and frameworks for querying multilayered annotations.

We have illustrated that even though image analysis technologies have recently shown promising progress, there is a lot of added value in text and speech, and it would be a waste of resources not to use the relatively mature tools and techniques for processing language. Though this statement has been made already more than ten years ago, it is even more valid, now that the convergence of knowledge engineering, semantics and multimedia analysis is bringing in massive support for the creation of richly annotated corpora.

We would like to argue in addition that the development of next-generation access tools for heterogenous archives requires a research agenda that is not just focusing on crossing modalities and media, but also takes up the generation of medium-neutral or normalized representations. There is a potential for exploiting annotation types across media types and at various levels of abstraction: image features combined with speech transcripts, or image and speech features combined with manually generated metadata, etc. The development of abstract models for the representation of both content and query seems an obvious next step. From this perspective the massive conceptual annotation initiatives that we see nowadays are just an initial, but highly welcome step towards the creation of a an infrastructure for training resources that will be required to develop more generalized multimedia search methodology.

REFERENCES

[1] M. G. Brown and J. Foote *et al.*, "Automatic content-based retrieval of broadcast news," in *Proc. Third ACM Int. Conf. Multimedia*, Nov. 1995, pp. 35–43.

[2] A. Smeaton, W. Kraaij, and P. Over, "TRECVID—An overview," in *Proc. TRECVID 2003*, 2003 [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs.org.html

[3] W. Kraaij, A. Smeaton, and P. Over, "TRECVID—An overview," in *Proc. TRECVID 2004*, 2004 [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs.org.html

[4] P. Over, T. Ianeva, W. Kraaij, and A. Smeaton, "TRECVID—aN overview," in *Proc. TRECVID 2005*, 2005 [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs.org.html

[5] A. Smeulders, M. Worring, S. Santini, and R. J. A. Gupta, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[6] LSCOM lexicon definitions and annotations version 1.0, DTO Challenge workshop on large scale concept ontology for multimedia Columbia University, Tech. Rep. 217-2006-3, Mar. 2006.

[7] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. 26th Annu. Int. ACM SIGIR Conf.*, Toronto, ON, Canada, 2003, pp. 119–126.

[8] K. Barnard and P. Duygulu *et al.*, "Matching words and pictures," *J. Mach. Learning Res.*, vol. 3, pp. 1107–1135, 2003.

[9] G. Carneiro and N. Vasconcelos, "A database centric view of semantic image annotation and retrieval," in *Proc. 28th Annu. Int. ACM SIGIR Conf.*, New York, 2005, pp. 559–566.

[10] L. Hollink and M. Worring, "Building a visual ontology for video retrieval," in *Proc. 13th Annu ACM Int. Conf. Multimedia*, New York, NY, 2005, pp. 479–482.

[11] C. Fellbaum, Ed., *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[12] T. Strzalkowski, *Natural Language Information Retrieval*. Norwell, MA: Kluwer, 1999.

[13] P. Buitelaar, M. Sintek, and M. Kiesel, "Feature representation for cross-lingual, cross-media semantic web applications," in *Proc. ISCW Workshop Knowledge Markup Semantic Annotation (SemAnnot2005)*, 2005, pp. 502–513.

[14] P. Castells and F. Perdrix *et al.*, "Neptuno: Semantic web technologies for a digital newspaper archive," in *Proc. ESWS*, 2004, pp. 445–458.

[15] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval," *J. Web Semant.*, vol. 2, no. 1, pp. 49–79, 2004.

[16] F. de Jong, J.-L. Gauvain, J. den Hartog, and K. Netter, "Olive: speech based video retrieval," in *Proc. CBMI'99*, Toulouse, Oct. 1999, pp. 75–80.

[17] J.-L. Gauvain, L. Lamel, and G. Adda, "Transcribing broadcast news for audio and video indexing," *Commun. ACM*, vol. 43, no. 2, pp. 64–70, 2000.

[18] R. Ordelman, "Dutch speech recognition in multimedia information retrieval," Ph.D. dissertations, University of Twente, Enschede, The Netherlands, Oct. 2003.

[19] F. de Jong, R. Ordelman, and M. Huijbregts, "Automated speech and audio analysis for semantic access to multimedia," in *Proc. 1st Int. Conf. Semant. Dig. Media Technol. (SAMT 2006)*, Athens, Greece, 2006, pp. 226–240.

[20] T.-S. Chua, S.-Y. Neo, H.-K. Goh, M. Zhao, Y. Xiao, and G. Wang, "TRECVID 2005 by NUS PRIS," in *Proc. TREC Video Retrieval*, 2005 [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs.org.html

[21] E. André, , R. Mitkov, Ed., "Natural language in multimedia/multimodal systems," in *Handbook of Computational Linguistics*. Oxford, U.K.: Oxford Univ. Press, 2003, pp. 650–669.

[22] W. Kraaij and W. Post, "Task based evaluation of exploratory search systems," in *Proc. SIGR Workshop, Eval. Exploratory Search Syst.*, 2006, pp. 24–28.

[23] M. Spitters and W. Kraaij, "Unsupervised clustering in multilingual news streams," in *Proc. LREC 2002 Workshop: Event Modelling Multilingual Doc. Linking*, 2002, pp. 42–46.

[24] F. de Jong and W. Kraaij, "Content reduction for cross-media browsing," in *Proc. RANLP Workshop Crossing Barriers in Text Summariz. Res.*, H. Saggion and J.-L. Minel, Eds., Borovets, Bulgaria, 2005, pp. 64–69.

[25] J. Kuper and H. Saggion *et al.*, "Intelligent multimedia indexing and retrieval through multisource information extraction and merging," in *Proc. 18th Int. Joint Conf. Artif. Intell. (IJCAI)*, Acapulco, Mexico, Feb. 2003, pp. 409–414.

[26] H. Saggion, H. Cunningham, D. Maynard, K. Bontcheva, O. Hamza, C. Ursu, and Y. Wilks, "Extracting information for automatic indexing of multimedia material," in *Proc. LREC 2002*, May 2002, pp. 29–31.

[27] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, no. 1–7, pp. 107–117, 1998.

[28] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[29] H. Lee and A. F. Smeaton, "Designing the user interface for the Físchlár digital video library," *J. Dig. Inf.*, vol. 2, no. 5, 2002 [Online]. Available: http://jodi.tamu.edu/Articles/v02/i04/Lee

[30] T. Westerveld, A. P. de Vries, and G. Ramírez, "Surface features in video retrieval," in *Proc. 3rd Int. Workshop Adaptive Multimedia Retrieval, AMR'05*, Jul. 2005, pp. 180–190.

[31] K. P. Truong and D. A. v. Leeuwen, "Automatic detection of laughter," in *Proc. InterSpeech*, Sep. 2005, pp. 485–488.

[32] J. Garofolo, C. Auzanne, and E. Voorhees, "The TREC SDR track: A success story," in *Proc. 8th Text Retrieval Conf.*, 2000, pp. 107–129.

[33] R. W. White, D. W. Oard, G. J. F. Jones, D. Soergel, and X. Huang, "Overview of the CLEF-2005 cross-language speech retrieval track," in *Proc CLEF 2005* , pp. 744–759.

[34] H. E. Blok and V. Mihajlovic *et al.*, "The TIJAH XML information retrieval system," in *Proc. 29th ACM Conf. SIGIR*, 2006, pp. 725–726.

[35] A. Hauptmann, "Lessons for the future from a decade of informedia video analysis research," in *Proc. Int. Conf. Image Video Retrieval (CIVR05)*, Singapore, 2005, pp. 1–10.

[36] A. Jaimes and M. Christel *et al.*, "Multimedia information retrieval: What is it, and why isn't anyone using it?," in *Proc. 7th Int. Conf. ACM SIGMM*, 2005, pp. 3–8.

Franciska M. G. de Jong is a Full Professor of language technology at the University of Twente, Enschede, The Netherlands, since 1992. She is also with The Netherlands Organisation for Applied Scientific Research (TNO), Delft, The Netherlands

She has a background in theoretical linguistics and started to work on language technology in 1985 at Philips Research where she worked on machine translation. Currently, her main research interest is in the field of multimedia indexing, semantic access, cross-language retrieval, and the disclosure of spoken word archives, and she coordinates part of the research of the Human Media Interaction group. She is frequently involved in international program committees, expert groups and review panels, and has initiated a number of EU-projects. In 2001–2003, she was a member of the EU/NSF "spoken word archives" working group. She is a Project Leader of the MultimediaN-project on semantic multimedia access (2004–2008). She chairs the steering committee of the Dutch IMIX project on multimodal information extraction and since 2004.

Dr. de Jong is a member of the board of the Dutch Research Council for the Humanities.

Thijs Westerveld received the Ph.D. degree in computer science from the Human Media Interaction Group, University of Twente, Enschede, The Netherlands, in 2004, for his work on the use of generative probabilistic models for multimedia retrieval.

He is a Researcher in the Database Architectures and Information Access group at the Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands. He has worked on numerous national an European projects in the areas of information retrieval and multimedia retrieval, and published in international journals and conferences in the field. He has served as a Program Committee Member for several international information retrieval conferences and he has been a reviewer for international journals in the areas of information retrieval and pattern recognition. His current interests include the combination and integration of information sources for information retrieval in the textual or multimedia domain, language models for IR, and the evaluation of (multimedia) retrieval systems. Since 2006, he coordinates the multimedia track of the initiative for the evaluation of the initiative for the evaluation of XML retrieval (INEX).

Dr. Westerveld received the Best Paper Award in the International Conference on Image and Video Retrieval (CIVR) in 2004.

Arjen P. de Vries received the Ph.D. degree in computer science from the University of Twente, Enschede, The Netherlands, in 1999, for his work on the integration of content management in database systems.

He is especially interested in the design of database systems that support search in multimedia digital libraries. He has worked on a variety of research topics, including (multimedia) information retrieval, database architecture, query processing, retrieval system evaluation, and ambient intelligence. He works as a postdoctoral researcher at the Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands. He is also an Associate Professor in the area of multimedia data management at the Technical University of Delft, Delft, The Netherlands.