Automated Annotation of Synchronized Multimedia Presentations

Harald Sack and Jörg Waitelonis

Institut für Informatik
Friedrich-Schiller-Universität Jena
D-07743 Jena, Germany
email: {sack|joerg}@minet.uni-jena.de

Abstract. Semantic annotation of multimedia data for improving search engine performance is an important issue. We have focused on the automated annotation of video recordings of university lectures. Most times these lectures are supported by a desktop presentation provided by the lecturer. This presentation in general contains an outline of the given lecture and thus, enables automated annotation of the corresponding video recording. We have implemented a tool for automated MPEG-7 annotation of video recorded lectures, which enables efficient content based search within a video database. Given a particular search term our search engine prototype is able to deliver only the relevant sequences of the matching video recordings.

1 Introduction

Searching and information retrieval of multimedia data in the World Wide Web (WWW) is an important problem. Today, most WWW search engines are based on keyword search, i. e. they rely on metadata that describe a document – so called *descriptors* – generated with the help of information retrieval techniques. The automated generation of descriptors for text documents already is a complex task, because linguistic ambiguities have to be identified and semantic content has to be qualified and properly extracted. Considering multimedia data, as e. g. graphics, audio, or video data the task of generating descriptors in an automated way becomes even more difficult. To describe the semantic content of a single picture one requires profound conceptual knowledge and experience. On the other hand, the description of audio and video data content – as far as there is spoken language involved – can be considered as a mere transcription of spoken language into a text document.

The multimedia content of the WWW is rapidly growing and major search engines, as e. g. Google provide keyword based multimedia search possibilities [1, 2]. Digital video broadcasting and video archives of television networks provide a large amount of growing multimedia resources available on the WWW. Amongst others the WWW offers also lectures and scientific presentations recorded on video, e.g. the audio/video search engine Singingfish [4] delivers almost 7.000 video clips for the search term "lecture". The possibility of synchronizing the video image of the lecturer with the live presentation given during the lecture opens up new possibilities of automatically generating descriptors for multimedia data [16].

Our work is focused on the search in video recorded lectures. These lectures combine at least two different streams of data: a video recording of the lecturer and most times also a recording of the lecturer's presentation given on his computer's desktop. Additionally, the presentation itself is offline available as a file. We have synchronized the lecturer's video recording, the live desktop presentation, and the offline presentation file. Thus, we are able to determine which video sequence matches a slide and consequently, we annotate the video sequence with the text

contained in the corresponding slide by means of an MPEG-7 encoding. Based on the assumption that the lecturer's talk closely corresponds to the content of his presented slides, we have the possibility to annotate the video sequences with its content in an automated way. We use the generated MPEG-7 file as a basis for content based search. By formulating search terms (descriptors) the user is able to search our database of video recorded lectures by means of its contents. The result of a search query will be given in terms of all video sequences that are annotated with the according search term. Thus, the user has not only the possibility to determine, which lectures in general deal with a given subject, but he is also in the position to directly examine short video sequences that contain the topic he is looking for.

Our objective is to annotate an entire database of recorded lectures dealing with a great variety of topics. Automated semantic annotation that provides content based search in that particular area will enable the user to create lectures on demand. According to a given topic the user might select the most interesting video sequences and thus, he is able to create a personalized lecture fulfilling his own personal requirements.

The paper is structured in the following way. Section 2 explains how to create and how to deal with synchronized multimedia presentations of video recorded lectures. In Section 3 the way of converting a synchronized multimedia presentation into an annotated MPEG-7 file is presented, while Section 4 continues with covering the content based search on MPEG-7 files. Section 5 shows how to cope with different file formats using alternative presentation software, and section 6 concludes the paper.

2 Multimedia Search and Synchronized Multimedia Presentations

While the video recording of a lecture provides content remarks, comments, annotations, etc. in spoken language, the lecturer's presentation in general supplies an outline of the most important topics discussed during the lecture. The presentation – if staged on a computer – simply can be transcribed into semantically annotated text. In difference to the spoken language of the video recording the presentation also provides additional structural information, e. g. headings, text attributes, enumerations, or list elements. The only prerequisite to use these additional data is the synchronization of the video recording and the single slides of the given presentation. Then, the textual content of the presentation is used to generate semantic annotations for the recorded video that is encoded with the MPEG-7 standard [10].

State-of-the-art multimedia analysis systems are restricting themselves by resorting mostly to visual descriptions on a very low level of abstraction, e. g. describing the most *dominant color* of a picture. On the other hand there are approaches that focus on the abstract content dimension of multimedia documents and their correspondent semantic annotation, as e. g. *person* or *landscape* (see [13, 9] for a more detailed discussion). Furthermore, there has been much effort in scene detection and automated segmentation of videos [17, 26, 24]. Both approaches are not rather helpful for our purpose. The scenery provided in a video recording of a lecture is rather limited. Most times there is only some part of the lecturer visible within an almost static environment for the entire lecture. Most times a recorded lecture provides a single scene without any cut or even any camera movement. Thus, the content of the entire recorded video could be described by *person in classroom*, providing no further details about the actual content of the lecture.

To get access to the lecture's content the audio layer with the recorded voice of the lecturer can be analyzed. Based on speech recognition these approaches [11, 27, 28, 21] generate a time stamp for words and clusters of words, such that search

engines are able to find the exact position of a particular information inside a video recording. Here, the difficulty lies not only in the insufficient accuracy of the speech recognition software[14], but also in the ambiguity of spoken language and in addition also in the difficulty of considering which information is important and which is not. Alternatively, the content based annotation of the recorded lecture can be provided manually (e.g. [22]), which is not efficient in large scale (see also [7]).

Our approach is focused on combining the recorded video together with a synchronized presentation given by the lecturer (see fig. 1). The presentation has to be supplied by means of a computer, i. e. it's textual content should be accessible and each slide of a presentation has to be marked with a time stamp during the live presentation. Thus, we are able to decide which slide has to be associated with a recorded video sequence. Usually, each slide also provides some textual content. After removing stop words that provide no further information and after word stemming [20] each video sequence will be annotated with a couple of keyword roots describing its contents. In difference to speech recognition the slides of the synchro-

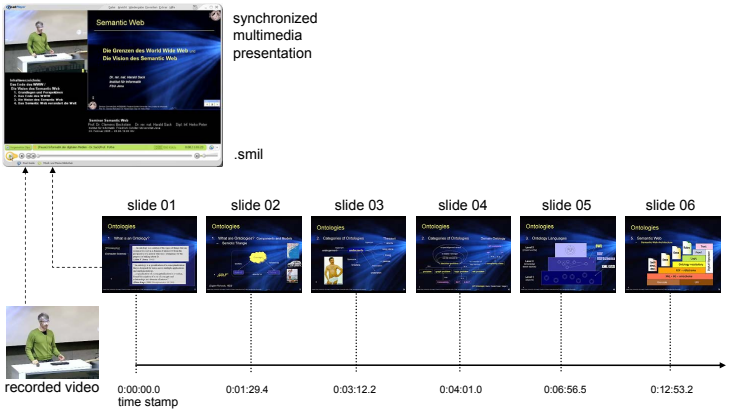

Fig. 1. Synchronizing Video and Presentation of a Recorded Lecture

nized presentation contain only text that is (*or at least should be considered to be*) significant for the associated lecture. Additionally, slides provide some rudimental semantics as e. g. headings or list entries that can be used to decide about the importance of several keywords related to a video sequence. To extract important concepts from spoken language is a much more complex task than extracting important concepts from a slide that already contains structured text representing an excerpt of the talk. Thus, our approach is also capable to filter important concepts from a given video sequence.

3 Automated Annotation of Synchronized Multimedia Data

The synchronization information combining the video recording and the slides of the presentation is a fundamental requirement. It can be achieved with a logfile that is maintained during the presentation on the lecturer's computer. Assumed the lecturer uses Microsoft PowerPoint (PPT) for his presentation, the implementation of this application can either be realized as an *AddIn* for PowerPoint itself or as an

external application, which monitors the presenter's actions. Both possibilities use the *Microsoft PowerPoint Object Library* [5] to handle PowerPoint events. By the time a *SlideShowNextSlide* PPT event occurs the logfile is extended by a new entry including the time stamp of the event, the slide number, and the path/file name of the affected PPT presentation. Thus, a synchronization between any given slide and a corresponding video sequence will be possible.

PPT presentations can be exported into several different file formats, e. g. XML, MHT (MHTML), or JPG. We chose the MHT file format [19] to extract textual information from the PPT presentation. The MHT file is composed of the single slides encoded in Hypertext Markup Language (HTML), embedded images, and the presentation's outline also encoded in HTML. To extract the descriptors for a slide the following processing has to be performed: First, all HTML tags are removed except tags that determine the document structure (e. g. headings, lists, paragraphs, etc.). Stop words that contain no further information are removed. To the resultant keywords the *porter word stemming* algorithm [20] is applied. Thus, a search engine working on the slides uses only keyword roots as descriptors providing a better recall of search results.

Subsequently, the slide headlines are extracted from the presentation's outline. For each PPT presentation we create a list of tuples providing slide number, descriptors, headline, and slide text.

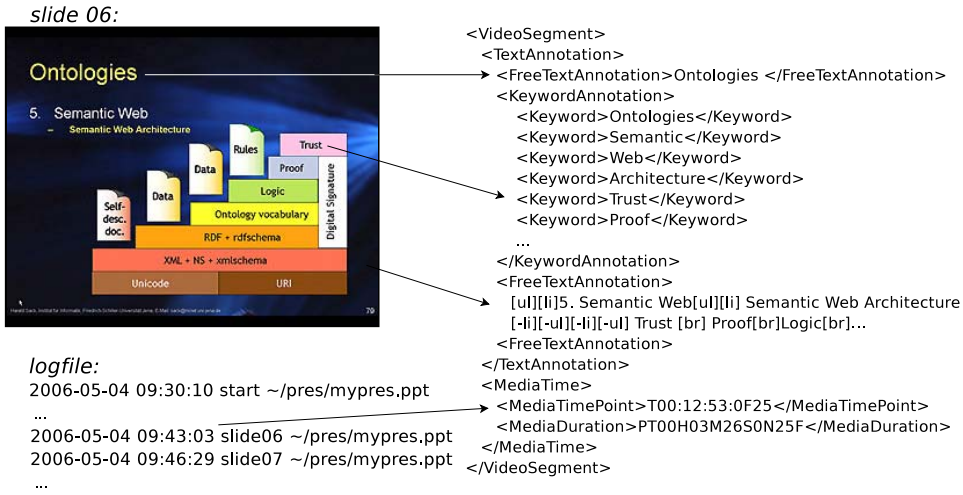

Fig. 2. Annotation of a Single Video Segment

Now, recorded video, logfile, and descriptor information have to be synchronized and encoded with the MPEG-7 standard. The time stamps within the logfile are used to partition the video recording into single segments. Thus, each video segment can be mapped to one particular slide of the presentation. MPEG-7 allows each video segment to be annotated with textual information [18, 25]. We utilized the MPEG-7 free text annotation (`<FreeTextAnnotation>`) and keyword annotation (`<KeywordAnnotation>`). Fig. 2 illustrates an example annotation of a video segment.

Each video segment annotation supplies text annotation together with time annotation to enable a temporal decomposition of the recorded video. The text annotations comprise a `<FreeTextAnnotation>` as headline, the `<KeywordAnnotation>`

providing descriptors, and a `<FreeTextAnnotation>` with the structured slide content. The structured slide content is only used to enable well readable search results and does not affect the search process itself. HTML tags enclosed in the structured slide content have to be masked to avoid interference with the XML encoding of the MPEG-7 file. The time annotations `<MediaTimePoint>` and `<MediaDuration>` are calculated from the time stamps provided in the logfile.

Additional media information (e. g. URI, creator, applications, etc.) and media usage information is stored using MPEG-7 content management elements (`<DescriptionMetadata>`, `<MediaInformation>`) in the annotations header. Further metadata about the lecturer (e. g. name, title, organization, institute, etc.) and the lecture itself (e. g. topic, date, place, duration, etc.) are also encoded in the MPEG-7 file (cf. [8], see fig. 3).

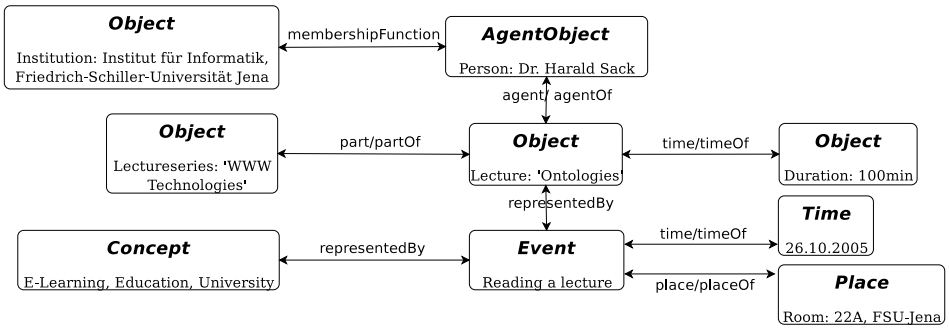

Fig. 3. MPEG-7 Encoded Semantic Description of a Video Recorded Lecture

There are several possibilities to maintain the MPEG-7 annotation. On the one hand, the annotation can be created directly via the PowerPoint AddIn. On the other hand, a separate application can be used to provide the annotation. We chose a web server providing a web interface for uploading the logfile and the presentation. The web server creates the annotation and as well provides the search engine that is described in the following section.

4 Searching MPEG7

The MPEG-7 encoding is based on XML (Extensible Markup Language). Therefore, searching MPEG-7 requires to parse a XML document, which follows the document object model (DOM). We provided all information required for content based search within the MPEG-7 document. Thus, no further input is required.

At first, we have to find out in which video segment a user given search term occurs. For this task the `<Keyword>` nodes of the MPEG-7 document have to be examined. The information related to the corresponding video segment can be acquired from the parent nodes of the resultant `<Keyword>` nodes that provide time stamps and video segment durations. Next, adjoined video segments have to be identified and merged into new contiguous video sequences. Segment i and j are adjoined, if $time(i) + duration(i) = time(j)$. All the resulting segments contain the user requested search term.

A different challenge is the creation of an entire knowledge base of recorded lectures and the provision of an efficient search facility. DOM parsing of the MPEG-7 files dramatically slows down search performance. Another issue when dealing with

an entire knowledge base will be the ranking of the search results. To speed up search performance an index over the descriptors in all MPEG-7 files has to be created. To avoid access to the MPEG-7 documents during the search process the index must contain all information provided by the MPEG-7 document.

The search is implemented in two steps: In the first step it is determined in which MPEG-7 document – i. e. in which video recording – the search term occurs. Secondly, for each resultant MPEG-7 document the particular video segments w.r.t. the search term together with the time stamp information has to be acquired. Thus, we have a two-level index architecture. In the first level the requested MPEG-7 documents have to be determined and in the second level for every MPEG-7 file the time stamp information related to the single video segments has to be determined. With a growing number of video recordings, one should consider the use of a database system.

As a second issue - how to rank the achieved search results - for each search term an ordering of the resultant video sequences has to be determined. This ordering should come up with the searchers preferences. The most fundamental approach is based on simple keyword frequency. The more often a descriptor occurs for a given video segment, the higher has to be its rank and it has to be put in front of lower ranked results. For further ranking refinement additional parameters can be defined:

- The more adjacent video segments are merged into a contiguous video sequence, the higher has to be the sequence's rank.
- The longer a video sequence takes related to the duration of the entire video recording, the higher has to be its rank.

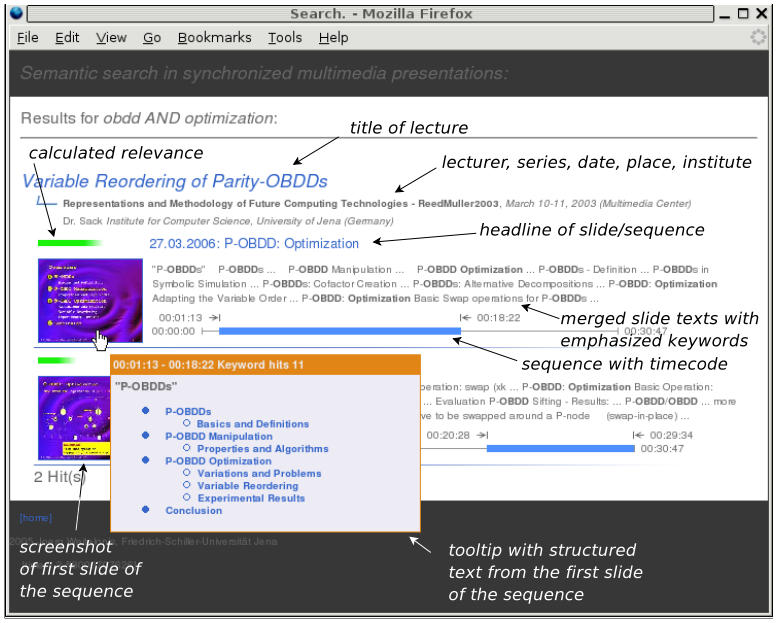

Fig. 4. Screen Shot of the Results Overview

We have implemented a prototype search tool [3] working on a small video database of about 25 recorded lectures most of them given in german. The implementation of our *search interface* consists of three components: search mask, results overview, and video segment replay. The search mask itself consists of a simple text

field. The search results are generated and visualized as shown in fig. 4. The results overview page offers many additional information elements that assist the user to decide, if a result is more or less relevant. At first, a list of all matching video segments is displayed. For each video segment the following information is available:

- title of the lecture,
- headline of the first slide in the video sequence,
- merged slide texts of all video segments,
- a screenshot of the first slide of the video segment,
- structured slide text of the first slide of the video segment available as a tooltip, and
- position and duration of the video segment shown in a timeline.

To refine the search the user can combine more than one keyword with Boolean operators 'AND' and 'OR'.

Choosing a video segment directs the user to the *view interface*, which implements the RealPlayer browser plugin for replaying the video recording. Depending of the video segment's start time the RealPlayer starts playing the recorded lecture from this position. Thus, a simple navigation within the multimedia database for pervasive learning is possible. To include new lectures into the knowledge base we have implemented a *web administration interface*. There, the administrator can upload the MHT file of the presentation, the logfile and the screenshots of the slides in one single step.

5 Alternative Presentation Sources

We used Microsoft PowerPoint based synchronized multimedia presentations as a starting point for automated annotation of recorded video lectures. For the synchronization a logfile is maintained during the presentation providing time stamps, which relate the content of a single slide to its corresponding video sequence. Given a PowerPoint based logfile plugin we are able to record the presenter's actions, as e.g. the transition from one slide to another, or also the page number of the currently displayed slide. As already described in section 3 the PowerPoint presentation is exported into Multipart HTML (MHT) to enable a matching of the corresponding slide content to a given time stamp.

But, different presentation software, as e. g. Adobe's Portable Document Format (PDF) [15], which is a rather common and platform independent document format for document exchange and document presentation, necessitates an upgrade of our processing. Again, the key problem is to match keywords with video sequences. Therefore, we have to synchronize the presentation and the recorded video. But, this time, depending of the applied presentation software, we are not necessarily able to directly determine the page number of the currently displayed slide.

Again, an external application running on the presenter's computer records user actions into a logfile. To match slide numbers with a recorded video sequence, the logfile contains time stamps of user actions together with scaled down screen shots of the currently presented slide. This offers the possibility to use any presentation software, even without being able to directly access the page number of the currently presented slide.

The screen shots maintained within the logfile can serve as input to an out-of-the-box optical character recognition (OCR) software. Thus, the text contained in a slide for a given time stamp can directly be determined. But, the weak accuracy of the OCR applied to the scaled down screen shot prevents its usage for direct index generation. Nevertheless, the textual content of the displayed slide is also directly available from the offline presentation. The accuracy of the OCR is sufficient

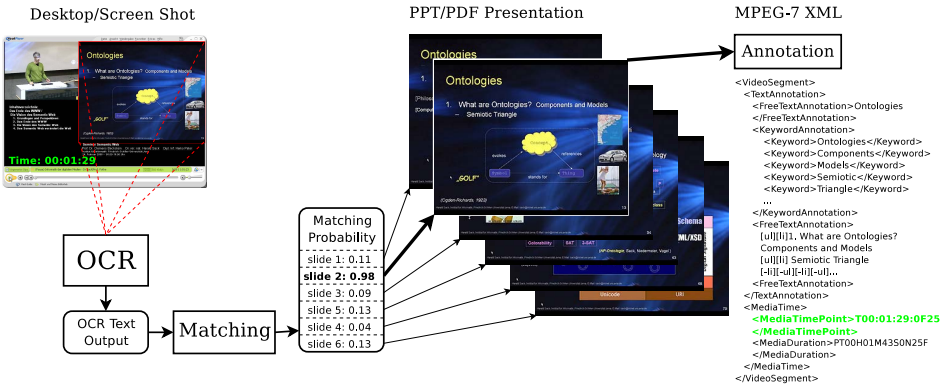

Fig. 5. OCR Supported Annotation

to match the text identified by the OCR with the original text from the offline presentation with a high probability and thus, we are able to determine which slides to match with a given video sequence (see fig. 5).

It is even possible to match slides and recorded video even without maintaining a logfile on the presenter's computer. But, this is only possible in the case of a recorded synchronized presentation, as e. g. recording the presentation desktop also as a video stream. If we have given two synchronized video streams, the presenter video stream and the presentation video stream, no logfile recording is required. In a post processing step the presentation video will be analyzed: Screen shots of the presentation are collected while writing time stamps and corresponding screen shots into a separate logfile. The OCR software determines the textual content of the screen shots, identifies consecutive screen shots of the same slide to create new time stamps for each slide change. Finally, the slide's content is processed as already described to determine, which slides to match with a particular video sequence.

6 Conclusions

We have shown how to implement content based search in a multimedia environment. We've focused on lectures represented as synchronized multimedia data streams. Recorded video lectures are annotated with the content of the slides presented by the lecturer. Thus, the content of a given slide is synchronized with the corresponding recorded video sequence. This annotation is encoded with the MPEG-7 standard, which serves as the basis of our search engine implementation. For reasons of performance we designed an index data structure, which enables fast access to particular video sequences related to a given search term.

While providing an automatic semantic annotation for recorded video lectures based on Microsoft PowerPoint presentations, we have also shown how to process alternative presentation formats as e. g. Adobe PDF, or how to process a recorded desktop video.

Based on our implementation a knowledge base can be set up, providing content based access to recorded lectures. By selecting the search results for a given search term the user is able to design an entire lecture according to his personal requirements and needs. Thus, we open up the way to personalized lectures on demand, supplying just the information that is needed independent from the user's location at any time.

References

1. Google image search, http://www.google.com/imghp/.
2. Google video search, http://video.google.com/.
3. Semantic search in synchronized multimedia presentations, prototype at http://www.minet.uni-jena.de/~joerg/cgi-bin/semsearch/search.cgi.
4. Singingfish – audio/video search engine, http://search.singingfish.com/.
5. How to handle PowerPoint 2002 events by using Visual Basic .NET 2002. Technical report, Microsoft Help and Support Article ID: 308330, http://support.microsoft.com/default.aspx?scid=KB;EN-US;q308330\&, 23 August 2005.
6. Thomas Sikora B. S. Manjunath, Philippe Salembier, editor. *Introduction to MPEG-7*. Chichester etc.: John Wiley & Sons, 2002.
7. David Bargeron, Anoop Gupta, Jonathan Grudin, and Elizabeth Sanocki. Annotations for streaming video on the web: System design and usage studies. In *Proceedings of the Eighth International World-Wide Web Conference*, 1999.
8. Ana B. Benitez, Jose M. Martinez, Hawley Rising, and Philippe Salembier. Description of a Single Mulitmedia Document. In B. S. Manjuta, Philippe Salembier, and Thomas Sikora, editors, *Introduction to MPEG-7*, chapter 8, pages 111–138. John Wiley & Sons, Chichester, 2002.
9. Stephan Bloehdorn, Kosmas Petridis, Carsten Saathoff, Nikos Simou, Vassilis Tzouvaras, Yannis S. Avrithis, Siegfried Handschuh, Ioannis Kompatsiaris, Steffen Staab, and Michael G. Strintzis. Semantic annotation of images and videos for multimedia analysis. In *ESWC*, pages 592–607, 2005.
10. S. F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):688–695, June 2001.
11. Jonathan Foote. An overview of audio information retrieval. *Multimedia Syst*, 7(1):2–10, 1999.
12. J. Geurts, J. van Ossenbrugen, and L. Hardman. Requirements for practical multimedia annotation. In *Multimedia and the Semantic Web, 2nd European Semantic Web Conference*, 2005.
13. Alexander G. Hauptmann and Michael G. Christel. Successful approaches in the TREC video retrieval evaluations. In *ACM Multimedia*, pages 668–675, 2004.
14. W. Hürst, T. Kreuzer, and M. Wiesenhütter. A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web. In *ICWI*, pages 135–143. IADIS, 2002.
15. Adobe Systems Incorporated. PDF Reference - Adobe Portable Document Format, 5th edition, http://partners.adobe.com/public/developer/en/pdf/PDFReference16.pdf, 2006.
16. Lauer, T., Mller, R., and Trahasch, S. Web technologies and standards for the delivery of recorded presentations. In *Proceedings of the IEEE International Conference on Information Technology Based Higher Education and Training (ITHET)*, pages 104–110, 2004.
17. Zhu Liu and Qian Huang. Detecting news reporting using audio/visual information. In *ICIP (3)*, pages 324–328, 1999.
18. National Institute of Standards and Technology. NIST MPEG-7 Validation Service and MPEG-7 XML-schema specifications, http://m7itb.nist.gov/M7Validation.html.
19. Jacob Palme, Alex Hopmann, and Nick Shelness. MIME encapsulation of aggregate documents, such as HTML (MHTML). Internet proposed standard RFC 2557, March 1999.
20. M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.
21. Stephan Repp and Christoph Meinel. Semantic indexing for recorded educational lecture videos. In *4th Annual IEEE Int. Conference on Pervasive Computing and Communications Workshops (PERCOMW'06)*, 2006.
22. J. R. Smith and B. Lugeon. A visual annotation tool for multimedia content description. In *Proc. SPIE Photonics East, Internet Multimedia Management Systems*, 2000.
23. R. Troncy. Integrating structure and semantics into audio-visual documents. In Dieter Fensel, Katia P. Sycara, and John Mylopoulos, editors, *The Semantic Web - ISWC*

2003, Second International Semantic Web Conference, Sanibel Island, FL, USA, October 20-23, 2003, Proceedings, volume 2870 of *Lecture Notes in Computer Science*, pages 566–581. Springer, 2003.
24. Ba Tu Truong, Svetha Venkatesh, and Chitra Dorai. Scene extraction in motion pictures. *IEEE Trans. Circuits Syst. Video Techn*, 13(1):5–15, 2003.
25. Toby Walker, Jörg Heuer, and Jose M. Martinez. Basic Elements. In B. S. Manjuta, Philippe Salembier, and Thomas Sikora, editors, *Introduction to MPEG-7*, chapter 7, pages 95–109. John Wiley & Sons, Chichester, 2002.
26. Jihua Wang and Tat-Seng Chua. A framework for video scene boundary detection. In *ACM Multimedia*, pages 243–246, 2002.
27. Michael Witbrock and Alex Hauptmann. Speech recognition and information retrieval:, January 27 2004.
28. Dongru Z. and Yingying Z. Video browsing and retrieval based on multimodal integration. In *Proceedings of the IEEE/WIC International Conference on Web Intelligence,Halifax, Canada*, 2003.