Streaming Media Architectures, Techniques, and Applications:
Recent Advances

Ce Zhu
Nanyang Technological University, Singapore

Yuenan Li
Tianjin University, China

Xiamu Niu
Harbin Institute of Technology, China

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE

Hershey • New York

Director of Editorial Content: Kristin Klinger
Director of Book Publications: Julia Mosemann
Acquisitions Editor: Lindsay Johnston
Development Editor: Michael Killian
Publishing Assistant: Milan Vracarich Jr.
Typesetter: Michael Brehm
Production Editor: Jamie Snavely
Cover Design: Lisa Tosheff

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Streaming media architectures, techniques and applications : recent advances / Ce Zhu, Yuenan Li, and Xiamu Niu, editors.
p. cm.
Includes bibliographical references and index.
Summary: "This book spans a number of interdependent and emerging topics in streaming media, offering a comprehensive collection of topics including media coding, wireless/mobile video, P2P media streaming, and applications of streaming media"--Provided by publisher.
ISBN 978-1-61692-831-5 (hardcover) -- ISBN 978-1-61692-833-9 (ebook) 1. Streaming technology (Telecommunications) I. Zhu, Ce, 1969- II. Li, Yuenan, 1981- III. Niu, Xiamu, 1961-
TK5105.386.S3746 2011
006.7--dc22
2010016311

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 17
Perspectives of the Application of Video Streaming to Education

Marco Ronchetti
Università Degli Studi di Trento, Italy

ABSTRACT

The field of e-learning has been a precursor in using the video streaming over the Internet. Both the synchronous and the asynchronous options have been explored over the last decade, with the asynchronous one becoming the dominant paradigm in recent years. Pedagogical research lecture reported evidence that video-streaming is an effective way of teaching, provided certain conditions are met. Technological research has attempted to investigate various ways to better produce or deploy video lectures: video segmentation, summarization, multimodal extraction of text and metadata, semantic search and gesture analysis are among the research areas that were involved. The present paper reviews the main technological research achievements and trends, and suggests directions in which we may be seeing the streaming of lectures to venture in near future.

INTRODUCTION

The field of e-learning has been a precursor in using the video streaming over the Internet. To our knowledge, the first proposal of an architecture for recording and distributing lectures in the form of video streaming over the Internet dates back to 1995 (Tobagi). Tobagi also implemented and demonstrated a first prototype: however apparently the system was never brought into routine production. The first systematic application of video streaming to teaching followed three years later (Hayes 1998). At that time, a VHS based system for delivering lectures to a geographically remote place (from USA to France) was substituted first with an audio stream with synchronized power point images, and shortly thereafter it evolved into a video transmission that included both the teacher and the slides with a technique called chroma key[1]. In recent years several custom systems were developed, some were commercialized, some were put in the public domain and others

DOI: 10.4018/978-1-61692-831-5.ch017

were used locally as prototypes. A review of the desirable features for such systems can be found in (Ronchetti 2008). After a short time in which pioneers opened the way showing a possible but uncertain future, big players are coming onto the scene today, such as in the case of the Massachusetts Institute of Technology on-line video collection[2] and the one of University of California at Berkeley who, according to U.S. Government news[3] of January 2008, was the first University with a plan to offer full courses on You Tube.

Research has shown that, from a pedagogical point of view, the video streaming of lectures is an effective practice (for a review see Ronchetti 2009). Moreover its production costs are already quite low. Systems like Lode[4] and OpenEya[5] are available at no cost since they are either open-source or free, and their hardware requirements are rather basic for today's computers. Hence, although there are still some problems, like the unwillingness of some teachers of being recorded, we believe that there are little doubts that the use of video streaming applied to education is here to stay, and will continue to expand. Already today, the list of websites dedicated to offer on-line video-lectures is impressive (see e.g. the partial catalog[6] compiled by University of Wisconsin at Milwaukee).

Although some of the early applications of video streaming to teaching were focusing on synchronous usage, in recent years most of the cases are concerned with asynchronous consumption. These two modalities present very different implications: in synchronous lectures it is desirable to allow remote users to interact with the speaker/teacher, while for the asynchronous ones the focus is shifted to other issues such as intelligent information extraction, ability to search, interconnect, navigate and annotate lectures. This second area has been very active, and has elicited several research streams.

In this chapter we shall review the research directions and results that matured over the last decade. Our coverage will reflect what has been happening in the last years, in which not much emphasis has been paid to the synchronous aspects, and many efforts have gone towards the view of a flexible, searchable collections of multimedia material built around video streaming, available on demand, in which the content allows random access and the user can easily find information about the content and locate interesting spots: a perspective that goes under the name of next generation digital library.

Synchronous Videolectures

Synchronous video streaming breaks spatial constraints and allows users to participate in real time to remote events. Students' physical presence in the teacher's location is not required any more. This solution allows for distributed classes (e.g. two classrooms on the same campus to accommodate a very large audience) and/or distributed individuals (e.g. students following a lecture in real time from their home). However, by simply using traditional video streaming solutions, an important obstacle arises: the lack of interaction. How can students in the remote classroom, or in their home, request teacher attention for asking a clarification? Ron Baecker (2003) has addressed this issue when designing the e-Presence system. e-Presence is a web-casting tool that was originally designed with the scientific seminar model in mind. It is based on unidirectional synchronous video streaming, with synchronized slides that accompany the speaker's video. The interaction problem is attacked by integrating a textual chat into the system, and introducing the role of the "mediator": a person that is located in the physical place where the event takes place, and monitors the chat (Schick et al. 2005). S/he acts as a proxy for the remote user: when a question comes into the chat, s/he calls the attention of the speaker and poses the question. A small problem came from the fact that the streaming of the "synchronous" flow was always 10-15 seconds late (because of the delay induced by the real-time compression

that is performed on the flight). As an evolution of the textual chat model, Baecker et al. (2006, 2007) also investigated the possibility of introducing a VOIP channel for allowing direct interaction between student and teacher, and/or among students.

Bidirectional video streaming among multiple points is actually fairly common in many videoconferencing systems. Obviously, we cannot review here the whole vast world of videoconferencing, as this chapter focuses on specific applications on teaching and learning. We shall hence limit the discussion to the use of videoconference in the didactic framework. As we shall see, however, even though marrying videoconference and distance education seems to be a quite natural idea, it never really caught up in a significant way. In 2004 Raymond et al. noticed that, in spite of the diffusion of videoconferencing tools, it was difficult to find examples of successful application to the educational field. This is in striking contrast with the vast success that asynchronous video lecture systems started having in those years. Five years later, the situation has not changed much. Raymond et al. argued that the problem is connected with the lack of friendliness on multi-user videoconference systems. Even today, systems like Skype are widely used for point-to-point interaction, but their support for multipoint conferences is not as simple and intuitive as for the point-to-point case. Even systems that were designed with videoconference in mind are typically limited to a relatively small number of participants. In contrast, a lecture is a traditional academic setting involves a large number of actors (of the order of one hundred) with strongly asymmetric responsibilities, so that the flow of information is mostly unidirectional. Typical videoconference tool are aimed at much smaller communities, with more symmetric roles. They have been used in teaching for various activities, such as tutoring, or allowing faculty members to participate in a thesis defense at a remote institution, but their impact and diffusion is orders of magnitude less than their asynchronous counterpart.

In part this is due to a relatively complex technical set-up. Certainly the scenario of a synchronous, interactive tele-lecture suffers from the extra managerial burden imposed to the teacher. In an asynchronous scenario the teacher can almost forget that s/he's being recorded, and can run "business as usual", while in a mixed setting (in which s/he is teaching to a traditional classroom and at the same time s/he is involved in a videoconference) there is a cognitive overload, unless a mediator is used, as in the above-quoted Baecker's approach.

Our feeling is that an implicit cost/benefit analysis has severely limited the use of synchronous video streaming as opposed to the asynchronous version: the users' perception is that the extra effort required to apply synchronous video streaming to teaching largely outperforms the obtainable benefits.

Synchronous video streaming still has interesting uses in educational setting in more restricted and specific application areas. It is used in point-to-point versions for involving a remote expert in a traditional lecture: in such case the teacher becomes the natural mediator between the class and the expert. It can be used for inclusion in class of a remote child in hospitals (see e.g. the TelecomItalia Smart Inclusion project[7]), or for inclusion of a few children living on remote places (e.g. in the "Isole in rete" project[8] in which children living on a small island are virtually included in a larger remote classroom). In all these cases a simple point-to-point paradigm is used, and additional tools, such as e.g. interactive whiteboards, augment the interactivity palette.

Asynchronous Applications

In the previous section we concluded that asynchronous video streaming of lectures has been perceived by teachers, institutions and students as a higher value, simpler activity than the synchronous one. We should not overlook an obvious but essential advantage of the asynchronous

mode: while synchronous streaming only allows breaking spatial constraints, the asynchronous one also breaks temporal constraints, and hence it supports e.g. students who have full time jobs, or who need catching up on missed classes. Moreover, asynchronous mode allows for more freedom in using the material: like in books, it is not necessary to take the entire content, or to consume it sequentially. In fact, the analysis of usage patterns has shown that students tend not to watch an entire video-lecture, but rather to jump to video-fragments that are interesting for them for checking their notes, or re-hearing an explanation. Evidence in this sense was provided by the work by Zupancic & Horz (2002), and reinforced by the report by Ronchetti (2003). Moreover, Soong et al. (2006) reported that students accessed mostly those parts of lectures, which they did not understand, implicitly confirming that they watch fragments rather than full lectures. Zhang et al. (2006) stated that students using video-lectures with the possibility of random access performed better than those in other settings, and showed better learner satisfaction. The possibility to quickly navigate the lectures is therefore essential (and this is probably the main aspect that makes digital recording deeply different from more traditional VHS-based videos). Watching an entire video is in fact a time consuming experience, and we need ways to quickly identify and access the information of our interest. Moreover, the focus is shifted from a single lecture to the whole collection of available material: the notion of a digital library becomes central, and the "library" is not any more (digital) text only, but a vast collection of multimedia. Video streaming becomes one component in the new, broader scenario.

This brought research to focus on a set of topics that include the ability to:

- automatically summarize video-lectures;
- segment videos into semantically homogeneous chunks;
- generate indexes that enable effective search;
- mine audio and video to extract explicit information that can be used for post-processing the videos and/or be passed to the users;
- extract metadata;
- produce personal annotations (which might be textual or multimedia-based), and share them with peers.

In the rest of this chapter we shall discuss these issues, because even though they are not directly related to video streaming, their impact on video streaming usage is fundamental. In fact the success of these research streams will enable a new generation of on-demand video streaming, in which random access to information tokens in a large collection of videos, and within the videos themselves, will be made possible.

Summarizing Videolectures

When we try to find some information from traditional sources (e.g. books) we have tools that help us. For instance, we search for books on a given topic through an OPAC (On-line Public Access Catalogue[9]) service and we identify a set of candidates. To actually understand if one of the candidate books is relevant to us, we have a wealth of techniques to gather information. On the back or inner cover of the book, we can read a short summary and get some information on the author. We can look at the index and read the titles of the chapters. We can read the introduction (of the book or of a specific chapter). We also browse the book, and skim the text. In a few minutes we are able to get a pretty good understanding of the utility of the book for us, and we are able to do a random access to the part that is most relevant for us.

When it comes to multimedia, we do not have (yet) anything similar. Digital libraries start to offer services for searching a multimedia resource,

but then it is difficult to gain perspectives of a document without watching the video (or listening an audio recording) in its entirety. The research area called "video abstracting" deals with this issue. The review article by Truong and Venkatesh (2007) describes this area in general, without specific reference to the case of video-lectures. Unfortunately many of the techniques described there are not very useful in the case of video-lectures, especially when these are recording of events that took place in a classroom. Both key-frame sampling and video skimming are not very useful when the images are rather homogeneous, even though they can be helpful e.g. in detecting slide transitions. Most e-lecture systems however already deal in a special way with slides (keeping track of the time at which slide change occur, and maintaining snapshots of each individual slide).

He et al. (1999) proposed an ad hoc technique for summarizing video-lectures. It is interesting to follow their line of thinking, since they envisioned several ways top mine information although they followed only a subset of them. We'll follow their paper, but for sake of completeness we'll also integrate some additional information not present in their work. Their top taxonomy comprises:

- Video channel;
- Audio channel;
- Speaker action;
- End users' actions.

The former can in principle be used for several means, such as detecting slide transitions (if such information does not come from other sources), identifying sections in which the speaker writes on the blackboard, and identifying the use of multimedia within the lecture (such as when the speaker gives a live demonstration through a computer simulation or shows a video in class). Also, speaker gesture can be extracted and analyzed. In their paper they apply none of these techniques because the video consisted only of a "talking head", and hence not much could be inferred from video analysis.

Audio could be used by attempting to extract meaning from the spoken works. While several other authors followed this line of research (as we shall discuss later), He et al. (1999) focused on the audio channel examining pitch, pause, intonation and other prosody information. For instance, it is known form previous research that the introduction of a new topic often corresponds with an increased pitch range. Pauses were used to detect the beginning of phrases. In such way it was possible to avoid including in the generated summary segments that start in the middle of a phrase. In fact, He's et al. report that users found segments starting in the middle of a phrase to be very annoying.

Speaker actions could be deducted from the video or captured through other means. For instance, many video-lecture capturing software applications offer the possibility to record both the time at which slide changes occur, and the slide itself. As we mentioned, gesture analysis could be a valuable source of information. Also facial expressions could be helpful.

End user actions are a very valuable source of information. If videos are watched on-line, the server can record users' activity – such as jumps to different parts of the lecture, logs of which parts of lectures were watched the most etc. He et al. actually deployed such information. Some systems however rely on local watching (after downloading the whole video). Sometimes this is done because watching on-line suffers from network congestion, so that users find it more convenient to obtain (in batch) a local copy and then use it on their own machine. In such case it is much more difficult to gather information about behavioral patterns.

He et al also defined the desirable attributes of a good summary:

- Each segment should be concise;
- The set of selected segments should cover all key points;
- Earlier segments should establish the right context for the following ones;
- The flow in the summary should be natural and fluid, so as to provide overall coherence.

They proposed three summarization algorithms: the first one was one only on slide transitions points, the second one used only the pitch information while the third one used all the available information (slide transition points, pitch and user-access patterns). They did not find any significant difference among the three approaches, and concluded that the simplest one is therefore preferable.

Yokoi and Fujiyoshi (2007) proposed a technique that, although is not a summarization technique, reduces the time needed to watch the video of a lecture. Their idea is that there are portion of lectures that are not significant, and hence they cut or compress them. They identified content-free segments (those characterized by pauses and silence) and cut them out. Also, they spotted the chalkboard writing segments and apply a fast-forwarding increasing the video speed by 3 times during these segments. The identification of chalkboard writing phases was based on image analysis. The final result was that the processed lecture is 20% to 30% shorter than the original one. The whole process was automatic and the final result was comparable with what can be obtained by manual editing of the videos. At the time of their report they were not able to produce a lecture index, but they mentioned it as a future activity. Of course, since most of their compression comes from chalkboard writing identification, this approach is best suited to traditional lectures (not the ones mostly based on electronic presentations).

Extracting Text from the Audio Track

The idea of deploying Automated Speech Recognition (ASR) techniques to the audio tracks of video-lectures to generate a transcript is very natural. The transcripts can then be used in a variety of ways. Wald (2005) suggested that they could be used to create captions (e.g. for deaf learners), and to assist those who, for cognitive, physical or sensory reasons, find note taking difficult. Also the possibility to allow searching multimedia material by using the transcripts is mentioned in his work.

However, using an ASR to extract text from a video-lecture in not trivial, as a good quality of the sound is not always guaranteed. Poor acoustic conditions, differences in speakers' style and accent, and the use of generic vocabularies are the main obstacles. In ideal conditions (i.e. anechoic room, slow speaking rate, and limited vocabulary) a previously trained state-of-the-art ASR system can achieve a Word Error Rate (WER) of less than 3%. In general conditions however the error ranges from 20% to about 45%.

The performance of ASR systems can be improved by creating ad-hoc acoustic model training the system on the speaker's voice. Typically, this requires the teacher to run a training session in which s/he reads a predefined text, so that the system can adjust itself comparing its prediction and the known (exact) expected result. Although such operation takes only a relatively short time (such as half a hour) it might be considered annoying by the teachers, with the result of increasing their unwillingness to use lecture recording systems. Many teachers are in fact nervous about the idea of being recorded, as such operation exposes their performance outside the classroom, and possible mistakes or imprecise wording cannot be hidden behind a "you did misunderstand what I said". Adding an additional nuisance is certainly not needed. Hence a "speaker independent" ASR (i.e. one that does not need specific training) is a better choice, even if performances are lower.

Another way to improve the ASR performance is to have a specific language model. The likelihood of the used words is not flat across all domains: by changing domain certain words become more common and other more rare. Some words are domain-specific, and sometimes – especially in languages other than English – words in a different language are used. Hence, by knowing in advance the domain and the corresponding word distribution, one can greatly improve the results of the ASR system. Textbooks relative to the specific domain of the lecture can be used as a baseline to create the language model, by calculating word frequencies and correlations between words. Often such sources can be found in electronic form, which makes the process of building an ad-hoc language model relatively easy. However, often lecture language resembles conversational language (Glass et al. 2007): the effect is that, in spite of the coincidence of the semantic domain, textbook material has been found to be a rather poor predictor of the spoken language (Park et al. 2007) and was not helpful in reducing the WER. In contrast, it was found to improve the results in terms of retrieval performance (i.e. when the transcripts are used to respond to user queries to identify a relevant portion of the associated video). Park et al. consider this artifact to be caused by the spontaneous nature of the language used in class, as opposed to the more formal one employed by books. Although we could not find decisive evidence in literature, one could deduce that probably, by using textbooks as sources of the vocabulary, the word error rate is decreased on the most relevant words (i.e. the domain specific ones) and increased on more generic and common ones. Hence although the overall WER does not improve, terms that are most likely to be used in queries are better identified.

Munteanu et al. (2007) demonstrated that the language model can be improved also by taking into account the set of slides that often accompany the video-lecture. Their result seems to be in agreement with the above hypothesis, since slides are likely to contain almost only domain specific words.

Choudary et al. (2007) used textbook indexes, as they are manually created by experts, contain no trivial words and hence are very effective in representing the instructional videos.

Including the user in the loop can improve the results. Munteanu et al. (2006) suggested using a wiki-like system to allow users to manually intervene and correct errors in transcripts. Problems connected with conflicts, spam and history of the text can be solved in the traditional wiki way that has been popularized by Wikipedia. Munteanu's work shows an implementation of such technology, but as far as we know the results of their pilot study aimed at measuring the effectiveness of such approach were never published.

Lecture Segmentation

Once the text transcripts are available, further mining becomes possible. A useful task is to segment a lecture into smaller chunks. In fact a lecture typically includes several topics, but finding the boundaries among different subjects is not easy. Generic video segmentation in many cases relies on image analysis, but as we already discussed, in the case of video-lectures this not a very useful option. Rarely the video carries semantic meaning, and its usefulness relies mostly on psychological reasons: learners show a better concentration in front of a video than on audio + slide version (Glowalla 2004), and viewing the speaker gives a sense of familiarity that helps getting emotionally more involved. A limited extraction of information from the video has however been attemped by some authors. For instance, Liao and Syu (2008) identify three classes of scenes: *Teacher-Blackboard* (when the teacher is writing something on the blackboard or is explaining something), *Teacher-Student* (when the teacher is talking with students) and *Students* (when the scene shows the audience). Such information is reinforced by a classification of audio features

(background noise, uniqueness of the speaking voice). This knowledge is then used to provide a first level of segmentation of the lecture.

In most cases however detecting lecture segments relies on the availability of audio transcripts. Text segmentation is an active research area. For a short review of the general principle and techniques used, see section 2.1 in Lin et al. (2004). Here we shall focus on the specific applications of text segmentation to lecture transcripts.

Yamamoto et al. (2003) associated ASR transcripts with the textbook used in the lecture. They considered a window (called "Analysis section") in the ASR text and compared its content with the textbook content. They used term frequency–inverse document frequency (TF-IDF[10]) measure to associate the window content with the books sections represented by a vector-space model. They moved the analysis section along the speech transcript and obtain a map of the speech on the textbook sections. Some post-processing allowed cleaning the results by removing the noise generated by analysis sections that are wrongly associated and fall in the middle of otherwise homogeneous sections. Evaluation of their results reports a correct chapter association of 98%. Smaller grain (book) section association turned out to be correct in 89% of the cases.

This approach can be traced back to Hallyday and Hasan (1976) lexical cohesion theory, according to which text segments with a similar vocabulary are likely to belong to the same coherent topic segment.

Lin et al (2004) used a similar technique. They used a text window of fixed length (120 words) and slided the window through the text moving by 20 words at the time. They calculated the similarity between adjacent windows by taking into account seven language features (noun phrases, verb classes, word stems, topic words, combined features, pronouns, and cue phrases) and for each of them calculating a derivation of TF-IDF, which they named TF*ISF, here ISF stands for "Inversed Segment Frequency". The TF*ISF values is used to detect boundaries of sections. The best results were obtained for values obtained from noun phrases. This approach has the advantage of being universal, since it does not need a domain reference as in Yamamoto's case. However, Yamamoto's approach has the advantage that the reference provides a baseline for extracting semantic clues and understanding the section topic, while Lin's approach only identifies the section without providing any semantic indication.

Repp and Meinel (2008) poposed a hybrid approach that uses slides, transcripts and raw audio to extract various indicators: pauses in the audio track (the longest silences being used as segment boundaries), slides transition markers (they were assumed as boundaries in the transcripts), sliding window (with the same parameters used by Lin), clusters of adjacent similar words in the transcript, similarity between the text contained in sliding windows and adjacent slides, correlation between relevant keywords extracted from the slides and the text in the sliding window. Their results show that imperfect ASR transcripts severely harm the effectiveness of their approach, and that reasonable results can be obtained based on slide transition markers and on pauses. Their results are partially bases on the work reported in another paper (Repp et al. 2007) where they devise an algorithm to find slide transition times based on slide content and transcript in the case that the acquisition software did not already provide slide transition markers.

Search, Semantic Indexing and Multimodal Access

The ability to perform indexing and search on a video stream is another important feature that can be added when once the text transcripts are available. Search involves mainly two kind of queries:

1. given a collection of lectures, identify which lecture is dealing with a given search target;
2. given a lecture, identify the time locations where the search target is present.

Although the two queries can be combined, it is better to keep them conceptually separated. The first one can rely on additional metadata that may be present (e.g. if lectures are stored in a learning management system), while the second one depends on the availability of a temporally annotated transcription. Many ASR provide temporal annotation of the transcribed words, i.e. for every identified word they provide the time at with it was uttered. In such case, it is easy to create an inverted index of all the words contained in the speech. Using the index to search a given word, one can retrieve the portion of phrase that contains the searched word, and the time at which the word occurred during the speech.

Zhang & Nunamaker (2004) manually segmented lectures into semantically homogeneous fragments, provided metadata for each fragment and then allowed user to express queries in natural language. In response to the queries, the system used the metadata to identify the relevant segments and proposed them to the user. Such approach, that addresses the first type of query, strongly relies on human intervention (both for the segmentation and for the generation of metadata), and hence it is difficult and costly to use it on a large scale.

Yoshida et al. (2003) created a system that requires teachers to define a set of keywords for every lecture. These keywords are then matched again a text transcribed by an ASR, and the user is shown a trackbar where the location of any chosen keyword is highlighted.

Fuji et al. (2006) built a system, which searches a lecture video for specific segments in response to a text query. Their results showed that by using specific acoustic and language models (i.e. adapting speech recognition to the lecturer and the topic of the target lecture), the recognition accuracy was increased and consequently the retrieval accuracy was comparable with that obtained by human transcription. This result is partially in contrast with Hürst's finding (Hürst 2005): an investigation of the impact of the ASR errors turned out not to be dramatic, and showed that the imperfect transcripts of recorded lectures are anyway useful for further standard indexing processes.

Hürst and Deutschmann (2006) and Fogarolli et al. (2007) implemented systems that allow searching for arbitrary words in a video-lecture. In both cases the search was multimodal, as it also allowed searching in the slides accompanying the lectures. An excellent review of multimodal video indexing has been published by Snoek and Worring (2005).

Akiba et al. (2009) used a collection of (spoken) lecture documents and evaluate effectiveness of retrieval of searched targets. They concluded that correct retrieval from lectures is much more difficult that for broadcast news. They do not offer an explanation for this fact, but it could be argued that while news span over vastly different topics, a lecture generally has a precise semantic focus – hence the set of terms used throughout a lecture is probably less heterogeneous than in news and this causes more difficulties.

An interesting alternative approach to search has been reported by Iwatsuki et al. (2007). They used a (patented) technique called Fast-Talk Phonetic-Based Searching designed to build search databases from phonemes. The process hence does not need to go through the step of extracting text by using an ASR, and does not search keywords. The authors claim speaker independence and recognition rates of 98%. The main drawback of this method is that adding a semantic layer is impossible, since the concept of "word" does not play any role in the system.

Indexing and searching is an interesting option, but an important step forward would be being able to provide a semantic layer. Such a layer could be useful for both the type of queries we mentioned, and would allow automatic generation of metadata. Although most research on extracting semantics is performed in the framework of the Semantic Web vision (Berners-Lee et al. 2001) where ontologies

and logic reasoning play a important role, there are alternative approaches that attempt to extract semantic information such as the latent semantic indexing technique (Deerwester et al. 1990) and explicit semantic analysis (Gabrilovich & Markovitch 2007, Jambunathan et al. 2008), or that use lighter forms of ontologies (Giunchiglia et al. 2006). There have been several papers and projects on approaching e-learning under the umbrella of the Semantic Web, but to our knowledge most of them were not dealing with the specific theme of video-lectures. Repp, Linckels and Meinel (2008) built a system based on the use of ontologies, description logics and natural language processing. The system generated automatically semantic annotation and used it to provide a Query/Answer system. Repp and Meinel (2006) also showed that a smart semantic indexing can be done even with partially incorrect transcripts.

One of the main problems when working in the Semantic Web perspective is the need to use ontologies: getting a good ontology for a generic application domain is not (at least, yet) a trivial task. Hence approaches to semantics with a vision alternative to the Semantic Web are interesting, and a few papers concerned with extraction of semantic information from video-lectures using alternative approaches have been published. For instance, Choudary et al. (2007) have dealt with semantic retrieval from instructional videos. They dealt with the lack on an ontology by using textbook indexes to define the semantic concept space, represented each video in such space, and performed semantic retrieval. Fogarolli & Ronchetti (2008b) used Wikipedia as a reference for additional information, trying to extract semantics from the ASR text by relating it with Wikipedia's content. They have combined the terms extracted from the corpus (a set of lectures) with lexicographic relationships from Wikipedia. Wikipedia has been used as an alternative to ontologies and as a basis for cross-language references (Fogarolli et al. 2008a). They also used relations extracted from Wikipedia pages to graphically represent the main concepts and their relations within one video lecture (Fogarolli, Seppi & Ronchetti 2009).

Gesture Analysis

Early work on gesture analysis applied to video-lectures was done by Ju et al. (1998). They analyzed the speaker movements and defined three temporal gesture models. Their aim was to be able to identify the portion of a projected slide that the speaker wants to attract students' attention to. The first gesture is characterized by the hand entering over the projected slide, pausing for at least 1/3 of a second and then exiting. The second gesture is more complex: a hand enters, pauses, then moves and pauses an arbitrary number of times before exiting. The last is a waving gestures in which the hand enters and never comes to rest, but rather moves continuously within a small spatial neighbourhood: in such case they to determine the location of a waving gesture by selecting the centre by the pointing positions.

Wang, Ngo and Pong focused in 2004 on three basic gestures: circling (draw a circle around something), lining (draw a line along something) and pointing (point to somewhere for emphasis). Gesture detection was then used to automatic editing the videos, performing close-up of a particular slide region. In 2006 they (Wang et al. 2006) introduced a feature to predict the completion of these gestures: prediction is relevant for real time processing. They also included a study of the relations among gestures, ASR text and slide content. These correlations were used to improve accuracy and responsiveness of gesture detection. In a following work (Wang et al. 2007) automatic video editing was improved by including an analysis of poses, gestures and texts in lectures. They defined a Finite State Machine that, based on the information obtained by this analysis, can generate a simulated camera motion from the existing video.

It is interesting to mention that recent development of work done on real-time gesture detection brought Pong's group (Wang et al. 2008) to develop a system that can be used during the lecture (and not "*a-posteriori*" on the video-lecture) that can simulate an interactive whiteboard: the lecturer can draw on the slides simply by performing gestures in front of the projected slide.

Videolectures Annotation

The idea of allowing users to annotate generic web documents and possibly share these annotations with others has been around for quite a long time (see e.g. the Annotea project (Kahan et al. 2001). Somehow related ideas include tools like Google Notebook[11] and bookmark sharing systems like Delicious[12]. The original ideas about annotation were more ambitious, in that they aimed at being able to attach annotation not just to a page, but to a particular position in that page (hence allowing also multiple annotation of the same page). Similarly, annotation of streaming video has been a subject able to spawn a whole research branch that was also applied to video-lectures (see e.g. Bargeron et al 1999, Correia and Cabral 2005). For example, anchoring discussions to specific (video-lecture) resources rather than collecting them in bulletin boards or forums sounds like a reasonable thing to do, since it would provide a context for the discussion. There have been proposals in this sense (Abowd et al 1998, Haga 2002, Lauer et al. 2005). Another idea, a precursor of the Web 2.0 fashion, was to have students taking lecture notes, attaching them to the video or to a slide-cast and sharing them (e.g. Truong 1999, Kam et al 2005). Curiously, although the needed technology seems to be ripe and these approaches seem to be interesting and useful, none of these initiatives appears to have yet gathered much success. It is possible that the new emphasis given by the Web 2.0 will revitalize this area, allowing the creation and growth of social communities around multimedia resources in the learning domain.

OTHER TOPICS

We already discussed how gesture analysis could be used to post-process videos, e.g. zooming in the region of interest. Other researchers have attempted to implement a "virtual cameraman". In fact, capturing visual details – such as following the teacher to better capture his/her expressions and body language, or zooming on the blackboard when needed is one of the advantages offered by a (costly) human operator. In principle, a virtual cameraman (either in the acquisition phase or during post-processing) could achieve the same goal at a fraction of the cost. We shall not review here what can be done during the acquisition phase, and we'll rather focus on the operations that can be performed as post-processing of an existing video. A possible approach is to use just one camera with panoramic video capturing, and then to extract the portion of image of interest (Sun et al. 2005). A somehow similar idea is implemented in the EYA system (Canessa et al. 2008). They used a wide-angle photo camera to record high resolution pictures every 10 seconds. At present they leave to the user the possibility to focus on the details of interest: the browser shows the video and a large thumbnail of the current picture. When the user moves the mouse over the thumbnail, a high-resolution subset of the image is shown where other systems put the slide. In this way, the user can focus on the detail s/he wants (be it the blackboard, the projected screen or other). They are presently working to enrich the system by automatically extracting some feature, like detecting the slide transitions that occur during the lecture.

The problem of automatically detect slide transitions has been faced by many authors during the last decade, since slide transitions carry a semantic meaning and can help segmenting a lecture. A first system was proposed by Mukhopadhyay (1999), but it required a special synchronization tone to be emitted during the lecture recording. Later approaches were based on computer vision-based,

statistical techniques: reviews can be found in the papers by Gigonzac et al (2007) and De Lucia et al. (2008).

Chen & Heng (2003) used the ASR transcripts to match in the slide to identify transitions. Other techniques include detecting slide changes through an http proxy – but this approach only works for HTML-based presentations (Tanaka et al. 2004).

Since in many lectures the blackboard still plays an important role, it is important to be able to effectively capture what happens there. Again, this operation can be done in the recording phase (e.g. using an interactive whiteboard[13]), or an ad-hoc digital desk (see e.g. Joukov & Chiueh 2003) Early work (since 1996 till 2001) on the capture of several experiences in the classroom, including whiteboard traces, was performed in the framework of the Classroom2000/eClass project (see eClass 2001). The other possibility is to act during post-processing of the video. The E-chalk project (Friedland & Rojas 2006) extracts handwriting from a traditional blackboard via image analysis.

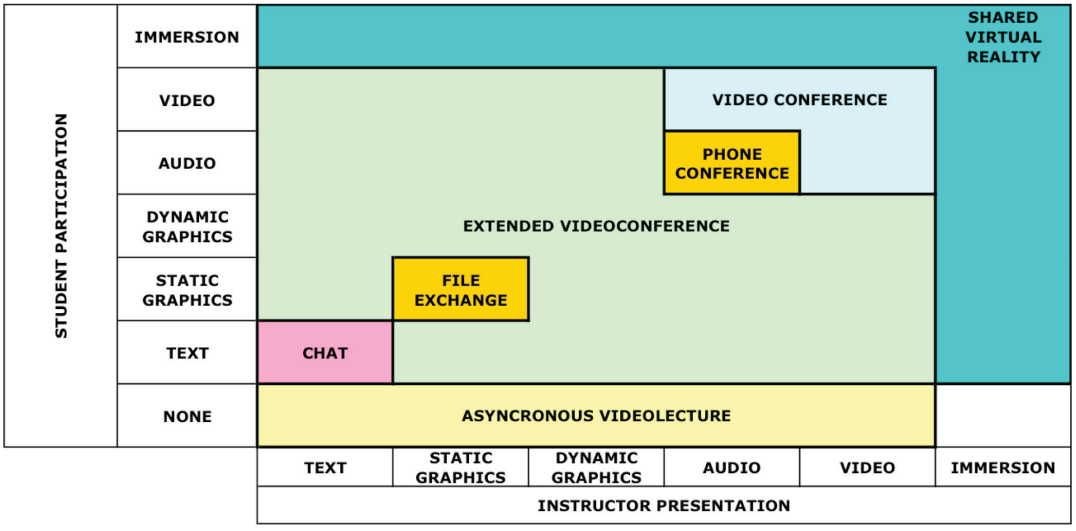
Finally, we mention that a careful and comprehensive study of a user interface that would offer the possibility of getting the most out of recorded video-lectures is, to our knowledge, still missing. Steps in this direction have been taken by Mertens et al. (2004, 2006). Such a study should take into consideration also mobile and ubiquitous devices, and analyze also their pedagogical effectiveness. There is in fact little doubt that new generation devices like the Apple iPhone, that has a reasonably high-resolution screen and that already delivers you-tube videos can open new frontiers also in the field of video-lectures, and in fact Apple recently started the iTunes U[14] initiative dedicated to the diffusion of video-lectures coded ad-hoc for the iPhone. Other work that pays attention to mobile devices has been reported by Friedland & Rojas (2006), who used mobile phones and iPods to show the videos of their E-chalk videos.

CONCLUSION

Given the growing needs for continuous education requested by today's society, video streaming applied to education is become more and more popular, and we expect this trend to continue.

To summarize the state of art, it is interesting to refer to the tele-education space as defined by Pullen (2000) and shown in Figure 1. Up to now,

Figure 1. The Tele-education space



most efforts (both on research and deployment) have been concentrated on a tiny portion of the whole space in the bottom part, where the interaction level between teacher and learner is null. Attempts to deploy the part that seems to be the most interesting one (the extended videoconference/video-lecture region) are either limited to point-to-point interactions, or failed to become a standard model. The main challenges seem to be mainly on finding a convincing, natural and efficient user interface model, and probably on modifying and evolving the teaching paradigm, that is still too much teacher-centered. At present, extensions that will allow the participants to live an immersive learning experience are only at the beginning. In the meantime, we can enjoy the growing wealth of asynchronous video lectures that can support institutional and continuous education. We expect that within a few years they will be enriched by harvesting the research lines discussed here, which will ultimately allow a more efficient use of our time while learning.

REFERENCES

Abowd, G. D., Atkeson, C. G., Brotherton, J., Enqvist, T., Gulley, P., & LeMon, J. (1998) Investigating the capture, integration and access problem of ubiquitous computing in an educational setting, *Proc. of the SIGCHI Conf. on Human factors in computing systems*, (pp. 440-447)

Akiba T., Aikawa K., Ithoh Y., Kawahara T. & Nanjo H. (2009) Construction of a Test Collection for Spoken Document Retrieval from Lecture Audio Data. *Journal of Information Processing (17)* 82-94

Baecker, R. M. (2003). A Principled Design for Scalable Internet Visual Communications with Rich Media, Interactivity, and Structured Archives. *Proceedings of CASCON*, *2003*, 83–96.

Baecker R. M., Baran M., Birnholts J., Laszlo J., Rankin K., Schick R. and Wolff P. (2006) Enhancing interactivity in webcasts with VoIP. *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*

Baecker R.M., Birnholtz J., Causey R., Laughton S.,Kelly Rankin K., Mak C.,Weir A., Wolff P. (2007) Webcasting Made Interactive: Integrating Real-Time Videoconferencing in Distributed Learning Spaces, *Human Interface and the Management of Information. Interacting in Information Environments, LNCS (4558)* 269-278

Bargeron, G. A. Grudin J, & Sanocki E. (1999) Annotations for streaming video on the web. *CHI '99 Extended Abstracts on Human factors in computing systems* pp. 278-279

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, (May): 2001.

Canessa, E., Fonda, C., & Zennaro, M. (2008) Academic Webcasting using the Automated EyA Recording System *INTED-International Technology, Education and Development*, Valencia/Spain, March 2008.

Chen, Y., & Heng, W. J. (2003) Automatic synchronization of speech transcript and slides in presentation. *Proc. of the Int. Symp. on Circuits and Systems, ISCAS '03.* vol. 2 pp. II-568-571

Choudary, C., Liu, T., & Huang, C. (2007) Semantic Retrieval of Instructional Videos. *Ninth IEEE International Symposium on Multimedia Workshops* (2007)

Correia, N., & Cabral, D. (2005). *VideoStore: A system to store, annotate and share video based content. Recent Research Developments in Learning Technologies 2005* (pp. 1299–1303). FORMATEX.

De Lucia, A., Francese, R., Passero, I., & Tortora, G. (2008). Migrating legacy video lectures to multimedia learning objects. *Software, Practice & Experience*, (38): 1499–1530. doi:10.1002/spe.877

Deerwester, S., Dumais, S. T., Furnas, G. W., & Landauer, T. K. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science American Society for Information Science*, *41*(6), 391–497. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

eClass (2001) for various references on the eClass project, see http://www.cc.gatech.edu/fce/eclass/pubs/index.html Retrieved Sept. 16, 2009

Fogarolli, A., Riccardi, G., & Ronchetti, M. (2007) Searching information in a collection of video-lectures. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications ED-MEDIA 2007*, p. 1450-1459

Fogarolli, A., & Ronchetti, M. (2008a). Intelligent Mining and Indexing of Multi-Language e-Learning Material. In Virvou, M., Howlett, R. J., & Jain, L. C. (Eds.), *New Directions in Intelligent Interactive Multimedia George* (pp. 395–404). Springer. doi:10.1007/978-3-540-68127-4_41

Fogarolli A., Ronchetti M., (2008b) Extracting Semantics from Multimedia Content. *Special Issue of Scalable Computing: Practice and Experience (9)* 1895-1767

Fogarolli, A., Seppi, G., & Ronchetti, M. (2009) RDF Graph Representation for Digital Content Visualization Summrization and Navigation. *International Conferences on Digital Libraries and the Semantic Web (ICSD2009)* pp 165-177

Friedland, G., & Rojas, R. (2006). Human-centered Webcasting of Interactive-Whiteboard *Proc. of the Eight IEEE Int. Symposium on Multimedia*

Fujii, A., Itou, K., & Ishikawa, T. (2006). LODEM: A system for on-demand video lectures. *Speech Communication*, (48): 516–531. doi:10.1016/j.specom.2005.08.006

Gabrilovich, E. and S. Markovitch S. (2007) Computing Semantic Relatedness is using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp.1606--1611

Gigonzac, G., Pitie, F., & Kokaram, A. (2007) Electronic slide matching and enhancement of a lecture video. *4th European Conference on Visual Media Production, IETCVMP 2007*, pp. 1 – 7

Giunchiglia, F., Marchese, M., & Zaihrayeu, I. (2006) Encoding classifications into lightweight ontologies. *Proceedings of ESWC'06*

Glass, J., Hazen, T. J., Cyphers, S., & Malioutov, I. (2007). Recent progress in the MIT spoken lecture processing project. Proc. *Interspeech*, *2007*, 2553–2556.

Glowalla U. (2004) Utility and Usability von E-Learning am Beispiel von Lecture-on-demand Anwendungen. *Entwerfen und Gestalten*, 2004 (in German)

Haga. (2002) Combining video and bulletin board systems in distance education systems. *The Internet and Higher Education (5)*, 119-129

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London, UK: Longman.

Hayes, M. H. (1998) Some approaches to Internet distance learning with streaming media, *Second IEEE Workshop on Multimedia Signal Processing* pp. 514-519

He, L., Sanocki, E., Gupta, A., & Grudin, J. (1999). Auto-Summarization of audio-video presentations. *Proceedings of the ACM Multimedia Conference (ACMMM)* pp.489–498.

Hürst, W. (2005) *Multimediale Informationssuche in Vortrags- und Vorlesungsaufzeichnungen*. Doctoral dissertation Universitaet Freiburg, Fakultaet fuer Angewandte Wissenschaften. (in German)

Hürst, W., & Deutschmann, N. (2006) Searching in recorded lectures. *Proc. of World Conf. on Educational Multimedia, Hypermedia and Telecommunications ED-MEDIA 2006* pp. 2859-2866

Iwatsuki, M., Takeuchi, N., Kobayashi, H., & Yana, K. (2007). Automatic Digital Content Generation System for Real-Time Distance Lectures. *International Journal of Distance Education Technologies*, (5): 7–18.

Jambunathan, A., & Ronchetti, M. (2008). Exploiting the collective intelligence contained in Wikipedia to automatically describe the content of a document. In Ronchetti, M. (Ed.), *The Semantic Web: a view on data integration, reasoning, human factors, collective intelligence and technology adoption* (pp. 209–216). Bangkok, Thailand: AIT e-Press.

Joukov, J., & Chiueh, T. (2003). Lectern II: a multimedia lecture capturing and editing system. *Proc. of Int. Conf. on Multimedia and Expo, 2003. ICME '03*. vol. 2 pp. II - 681-684

Ju S.X., Black M., Minneman S. & Kimber D. (1998) Summarization of videotaped presentations: Automatic analysis of motion and gesture. *IEEE Transactions on Circuits and Systems for Video (8)* no. 5, 686-696

Kahan, J., & Koivunen, M.-R. Prud'Hommeaux, E. and Swick R.R., (2001) Annotea: An Open RDF Infrastructure for Shared Web Annotations, *Proc. of the WWW10 International Conference*

Kam, M., Wang, J., Iles, A., Tse, E., Chiu, J., Glaser, D., et al. (2005) Livenotes: a system for cooperative and augmented note-taking in lectures. *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems* pp.531-540

Lauer, T., Trahasch, S., & Zupancic, B. (2005) Anchored Discussions of Multimedia Lecture Recordings. *Proceedings 35th Annual Conference - Frontiers in Education - FIE'05* pp. 12-17

Liao, Y.-C., & Syu, M.-H. (2008) An Actor-Based Video Segmentation System Using Visual and Audio Information in E-Learnin". *Eighth International Conference on Intelligent Systems Design and Applications. ISDA '08.* vol. 3 pp. 575 - 580

Lin, M., Nunamaker, J., Chau, M., & Chen, H. (2004) Segmentation of lecture videos based on text: a method combining multiple linguistic features. *Proc. of the 37th Annual Hawaii Int. Conf. on System Sciences, 2004*. pp. 1-9

Mertens, R., Ketterl, M., & Vornberger, O. (2006) Interactive Content Overviews for Lecture Recordings. *Eighth IEEE International Symposium on Multimedia* pp. 933-937

Mertens, R., Schneider, H., Muller, O., & Vornberger, O. (2004) Hypermedia navigation concepts for lecture recordings. *E-Learn: World Conference on E-Learning in Corporate*, pp. 2480–2847

Mukhopadhyay, S., & Smith, B. (1999) Passive capture and structuring of lectures. *MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia* Part 1 pp.477-477

Munteanu, C., Penn, G., & Baecker, R. (2007) Web-Based Language Modelling for Automatic Lecture Transcription *Proceedings of the Tenth ISCA European Conference on Speech Communication and Technology – EuroSpeech / Eighth International INTERSPEECH Conference*, pp. 2353–2356

Munteanu, C., Zhang, Y., Baecker, R., & Penn, G. (2006) Wiki-like editing for imperfect computer generated webcast transcripts, *Proc. Demo track of ACM Conf. on Computer Supported Cooperative Work – CSCW, pp. 83–84*

Park, A., Hazen, T. J., & Glass, J. R. (2005) Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. *IEEE International Conference on Acoustics*

Pullen, J. M. (2000) The Internet-based lecture: converging teaching and technology. ITiCSE '00: *Proceedings of the 5th annual SIGCSE/SIGCUE ITiCSEconference on Innovation and technology in computer science education* pp. 101-104

Raymond, D., Kanenishi, K., Matsuura, K., & Yano, Y. (2004). IP Videoconferencing in Distance Education: Ideas for a Successful Integration. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2004* pp. 4179-4184

Repp, S., Linckels, S., & Meinel, C. (2008) Question answering from lecture videos based on an automatic semantic annotation. *Proc. of the 13th annual conf. on Innovation and technology in computer science education ITiCSE '08* pp.17-21

Repp, S., & Meinel, C. (2006) Semantic indexing for recorded educational lecture videos. *Proceedings of the 4th IEEE Conference on Pervasive Computing and Communications Workshops (PerCom)*, pp 240–245

Repp, S., & Meinel, C. (2008) Segmentation of Lecture Videos Based on Spontaneous Speech Recognition. *Tenth IEEE Int. Symp. on Multimedia ISM 2008* pp. 692 – 697

Repp, S., Waitelonis, J., Sack, H., & Meinel, C. (2007) Segmentation and annotation of audiovisual recordings based on automated speech recognition. *Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)* pp. 620–629

Ronchetti, M. (2003) Has the time come for using video-based lectures over the Internet? A Test-case report *CATE - Web Based Education Conference 2003*

Ronchetti, M. (2008) Requirements for videolectures: which system is the best for you? *World Conference on Educational Multimedia, Hypermedia and Telecommunications (EDMEDIA) 2008*. pp. 2192-2199.

Ronchetti, M. (2010in press). The impact of Internet-carried video-lectures on education. In Magoulas, G. (Ed.), *E-Infrastructures and Technologies for Lifelong Learning*. IGI Global.

Schick, R., Baecker, R. M., & Scheffel-Dunand, D. (2005). Bimodal Text and Speech Conversation During On-line Lectures, *Proceedings of ED-MEDIA 2005*

Snoek, C., & Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, (25): 5–35. doi:10.1023/B:MTAP.0000046380.27575.a5

Soong, S. K. A., Chan, L. K., & Cheers, C. (2006) Impact of video recorded lectures among students. *Proceedings of the 23rd annual ascilite conf.: Who's learning? Whose technology?* Pp.789-792

Sun X., Foote L, Kimber D. & Manjunath B.S., (2005) Region of interest extraction and virtual camera control based on panoramic video capturing. *IEEE Transactions on Multimedia, (7)* n.5 981 - 990

Tanaka, Y., & Itamiya, T. Hagino. T., & Chiyokura, H. (2004) HTTP-proxy-assisted automatic video indexing for e-learning. *International Symposium on Applications and the Internet Workshops. SAINT 2004*. pp. 502 - 507

Tobagi, F. (1995) Distance learning with digital video. *Multimedia IEEE (2)* n.1 90 - 93

Truong, B. T. and Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl. (3)*, n.1, 1-37

Truong, K., & Abowd, G. (1999) StuPad: integrating student notes with class lectures. *CHI '99 extended abstracts on Human factors in computing systems* pp. 208 – 209

Wald, M. (2005) 'SpeechText': Enhancing Learning and Teaching by Using Automatic Speech Recognition to Create Accessible, Synchronized Multimedia *World Conf. on Educational Multimedia, Hypermedia and Telecommunications EDMEDIA-2005*

Wang, F., Ngo, C.-W., & Pong, T.-C. (2004) Gesture tracking and recognition for lecture video editing. *Proceedings of the 17th International Conference on Pattern Recognition ICPR 2004. vol. 3* pp. 934 - 937

Wang, F., Ngo, C.-W., & Pong, T.-C. (2006) Prediction-Based Gesture Detection in Lecture Videos by Combining Visual, Speech and Electronic Slides. *IEEE International Conference on Multimedia and Expo, 2006* pp. 653 - 656

Wang F., Ngo C-W., Pong T-C. (2007) Lecture Video Enhancement and Editing by Integrating Posture, Gesture, and Text. *IEEE Transactions on Multimedia (9)* n.2. 397–409

Wang F., Ngo C-W., Pong T-C. (2008) "Simulating a Smartboard by Real-Time Gesture Detection in Lecture Videos". *IEEE Transactions on Multimedia (10)* n.5 926 - 935

Yamamoto, N., Ogata, J., & Ariki, Y. (2003) Topic segmentation and retrieval system for lecture videos based on spontaneous speech recognition, *European Conference on Speech Communication and Technology.* pp. 961–964

Yokoi and Fujiyoshi. (2006) Generating a Time Shrunk Lecture Video by Event Detection. *2006 IEEE International Conference on Multimedia and Expo*, pp. 641 – 644

Yoshida, T., Tada, K., & Hangai, S. (2003). A keyword accessible lecture video player and its evaluation. *Proceedings of the International Conference on Information Technology: Research and Education*, *ITRE2003*, 610–614.

Zhang D. & Nunamaker, J.F.Jr. (2004) A natural language approach to content-based video indexing and retrieval for interactive e-learning. *IEEE Transactions on Multimedia (6)* n.3 450 - 458

Zhang, D., Zhu, L., Briggs, L. O., & Nunamaker, J. F. Jr. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information & Management*, (43): 15–27. doi:10.1016/j.im.2005.01.004

Zupancic, B., & Horz, H. 2002. Lecture recording and its use in a traditional university course. *Proc. of the 7th Annual Conf. on Innovation and Technology in Computer Science Education ITiCSE '02*. pp.24-28

KEY TERM AND DEFINITIONS

Automated Speech Recognition: A computational technique that converts spoken words to text.

Continuous Education: An all-encompassing term within a broad spectrum of post-secondary learning activities and programs.

Digital Library: A library in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers.

Distance Education: A field of education that focuses on the pedagogy and andragogy, technology, and instructional systems design that aim to deliver education to students who are not physically "on site".

E-Learning: A term that encompasses all forms of Technology-Enhanced Learning, i.e. support of any pedagogical approach that utilizes technology.

Gesture: A form of non-verbal communication in which visible bodily actions communicate conventionalized particular messages, either in place of speech or together and in parallel with spoken words.

Multimodality, Multimodal (Interaction): A form of man-machine interaction using multiple modes of input/output.

Semantic Web: An evolving development of the World Wide Web in which the meaning (semantics) of information and services on the web is defined, making it possible for the web to "understand" and satisfy the requests of people and machines to use the web content.

Text Segmentation: The identification of lexical units in writing systems.

Video Abstracting: A research area that deals with gaining perspectives of a video document without watching it entirely.

Video Conference: A set of interactive telecommunication technologies which allow two or more locations to interact via two-way video and audio transmissions simultaneously.

Video-Lecture: lecture recorded in a video and delivered through a variety of media.

Video Segmentation: The identification of boundaries among regions that differ for content or aspect in a video.

ENDNOTES

1 http://it.wikipedia.org/wiki/Chroma_key
2 http://watch.mit.edu/
3 http://www.america.gov/st/educ-english/2008/January/200801221815081CJsamohT0.1036035.html
4 http://latemar.science.unitn.it/LODE
5 http://www.openeya.org
6 http://www4.uwm.edu/libraries/media/streaming.cfm
7 http://www.telecomitalia.it/cgi-bin/tiportale/TIPortale/ep/contentView.do?channelId=-9793&LANG=IT&contentId=33796&programId=9596&programPage=%2Fep%2FTImedia%2FTICSList.jsp%3Ffonte%3DTelecom%2BItalia&tabId=6&pageTypeId=-8663&contentType=EDITORIAL
8 http://www.scuola-digitale.it/isoleinrete/content/index.php?action=read_pag1&id_cnt=6679
9 http://en.wikipedia.org/wiki/Online_public_access_catalog
10 http://en.wikipedia.org/wiki/Tf%E2%80%93idf
11 http://www.google.com/notebook/
12 http://delicious.com/
13 http://en.wikipedia.org/wiki/Interactive_whiteboard
14 http://www.apple.com/education/mobile-learning/