Internet Multimedia Search and Mining

Xian-Sheng Hua, Marcel Worring, and Tat-Seng Chua

August 28, 2010

Contents

List of Figures

List of Tables

Chapter 1. Cross-modality Indexing, Browsing and Search of Distance Learning Media on the Web

Alon Efrat
The University of Arizona
Arnon Amir
IBM Almaden Research Center
Kobus Barnard
The University of Arizona
Quanfu Fan
IBM T.J.Watson Research Center

1.1 Cross Modality Indexing and the Slides Channel

Cross modality indexing is the indexing one modality of a multimodal stream using another modality, where the association between the modalities is represented by some mapping. A simple example is the use of speech recognition in video search. The indexing and search are performed on the audio channel while the "document" is a visual moving picture. In this example, the straightforward mapping between the two is temporal and, given the separated two channels, is determined by the time offset which aligns the two. The mapping may be more involved in the case of closed captions and speech. In this case, a complete sentence appears on the screen for a certain duration which may or may not cover the duration of the corresponding speech. In the case of live closed captions there is a delay between the voice and the captions. This requires an appropriate temporal model for mapping, or crossing between the modalities.

Snoek & Worring [23] gave the following definition to the term multimodality:

> **Definition 1:** *Multimodality* The capacity of an author of the video document to express a predefined semantic idea, by combining a layout with a special content, using at least two information channels.

Where the three main information channels, or modalities that Snoek & Worring propose are:

- Visual modality: contains the mise-en-scene, i.e., everything, either naturally or artificially created, that can be seen in the video document;
- Auditory modality: contains the speech, music,and environmental sounds that can be heard in the video document;
- Textual modality: contains textual resources that describe the content of the video.

We discuss multimodal content on the web in the context of indexing, browsing and search of distance learning media, and refer the reader to [23] for a survey of the other applications of cross-modality. We focus mostly on recorded media, although many of the techniques discussed below are also applied to live streaming of lectures. A special attention is given to electronic *slides* used in the lecture, and accessible to the system. Though they can be considered as either part of the visual channel (as they are seen in the video) or the textual channel (as they commonly contain text that describes the lecture content), we propose to acknowledge their unique nature, and consider them to be an independent channel, which we call the *"slides channel"*. The slides channel emerges once video segments are linked to the slides that were used in the presentation. Its is determined by the temporal correspondence — which slides (if any) appears in which frame of the video, and by the spatial correspondence — where in the frame the slide appears, and under which transformation. Once the temporal and spatial correspondences are found, text extracted from the slides can be used as semantic handles into the video content. The special characteristics of the slides channel are detailed below.

In Section 1.2 we motivate the methods developed here specifically for the important field of educational video. We discuss systems that do not use a slides channel and contrast them with the advantages of systems that time-align and link slides to video segments to improve video search and browsing. We also discuss manual, semi-automated, and automated methods to achieve this time alignment. In Section 1.3 we characterize the slides channel that is accessible once the alignment is available. Then in Section 1.4 we elaborate on how the multiple modalities enhance authoring of distance learning material by linking together multiple sources of knowledge, generating a dynamic presentation that a user can conveniently navigate. Finally, in Section 1.5 the capabilities that the slide channel enables for linking, retrieving, and accessing videos over the internet.

1.2 Indexing Distance Learning Materials

We open by discussing recent developments and the significance of distance learning video. According to a 2006 survey conducted by the Sloan Consortium [7], more than 96% of the largest educational institutions in the United States have on-line offerings, and nearly 3.2 million college students took at least one on-line course in the Fall 2005 term.

This trend suggests a future in which a much larger segment of people across the world will have access to distance learning opportunities. Distance learning is used by most large universities and corporations. In many cases, the content is created by video capture of lectures or presentations, and distribution is conducted by video streaming over the internet. The video quality, however, varies greatly. A high quality production with multiple cameras is costly and labor intensive. Low-

cost video production, on the other hand, suffers in quality as a result of common problems such as inadequate illumination, significant color distortion, and images in video that lack clarity. Furthermore, reduced video quality affects readability of the text slides captured in the video, impairing learning ability of the viewer. Finally, the video stream is likely to lose the vibrant, interactive, and insightful context that exists in the lecture room.

Since the usage of the slides channel is central to this chapter, we discuss separately work that does not use this channel, and others that do use it.

Indexing distance learning media without using the slides channel

Systems for analyzing, indexing, searching and browsing educational videos have been studied extensively. Several of these systems deal with presentation slides. One of the earliest web-based browsing tools for educational videos is the Berkeley Lecture Browser [21], which provides access to streaming video and manually-synchronized slides. The Cornell Lecture Browser [20] is a similar tool, but uses automatic synchronization. The eClass system (originally known as Classroom 2000 [2]) uses an augmented "smart classroom" to capture white boards, slides, video and audio and create a virtual learning environment.

The CueVideo [3] system provides tools to quickly convert videotaped lectures or conference presentations into searchable on-line video proceedings, as well as automatic slide matching for topical indexing [24]. Other systems use information captured from whiteboards during lectures. However applying OCR to images of whiteboards is rather challenging. Instead, tracing the motion of the whiteboard marker can capture the temporal aspect of handwriting. See the recent book [16].

Web-based remote education systems, also referred as *e-learning* or *distance learning* systems, are significantly affecting the way people learn, teach and share knowledge. E-learning offers alternatives for those who cannot go to class and provides other advantages such as low cost, unlimited access of learning materials, and self-paced learning. Studies have reported that no significant differences are observed between the learning effectiveness of students learning over distance and students learning in classrooms [22].

Creating the slides channel: Matching slides to video frames

The specific task of synchronizing slides with video was first addressed by manually editing time stamps (e.g., the BMRC Lecture Browser [21]). Subsequently, the Classroom 2000 Project [2] introduced hardware to record time stamps during the presentation. Their ClassPad provided easy browsing and annotation of slides in the classroom for both teachers and students. Today, several commercial e-learning systems can deliver synchronized multimedia presentations over the Internet to end users where slides are displayed side by side with the video and the instructional

content can be accessed directly through slides (e.g., the Mediasite system [18] and Microsoft Producer [19]).

The approach taken by Zhang et al. [5] approaches assumes that the RGB video signals from the lecturer's computer to the video projector can be "tapped", and the video frames can be matched against the slides' images. This is a rather reliable method, compared to methods that search for the slides in a video from a camera, as the signal has much less distortion. It also assumes the scenario common in, for example, conferences, in which lecturers use their own computers, while the projectors are usually fixed.

The approaches discussed so far typically require dedicated hardware and/or software systems that are engaged in advance of the presentation. This has two limitations. First, it adds some overhead on the capture process. Second, it does not help exploit the large stores of on-line video from disparate sources. To address this limitation, methods to automatically match slides-to-video frames have been proposed. Most of of the early work can be described as a two-step matching process: slide extraction followed by slide identification. For example, Mukhopadhyay et al. [20] developed a system for structuring multimedia content in which slides and audio information are automatically synchronized with video. The system relies on a fixed camera and a predetermined transformation established using the four corner points of the projector during the system installation. Adding flexibility, Sydeda-Mohmood [24] proposed locating slides in videos using an illumination-invariant descriptor built upon the background color of slides. This method can detect slides appearing anywhere in the frames. The spatial layout geometry of the identified slide regions is then exploited to recognize slides. Liu et al. [15] also detect quadrilaterals in frames to extract slide regions, but apply a different, pixel based, method to identify the slides.

Erol et al. [8] developed a method that links slide images captured during presentations to the source files that generate the slides. Their method first classifies slides into four different types, and depending on the type, the slide is recognized by combining one or more methods that include edge histogram matching, line profile matching, string matching and layout matching. More peripheral to our topic, we note that Behera et al. [4] developed a method to spot slide change events based on analyzing the visual stability of a sequence of frames. However, no further slide recognition is performed in that work. Finally, speech transcripts alone [28] and audiovisual features [6] have been exploited to synchronize slides with video. Most of the approaches described above have in common that they were developed to handle one particular type of video style.

The approach used in the SLIC (Semantically Linked Instructional Content) system [25], is distinct from previous approaches that begin with transformation estimation followed by slide identification. The method used by SLIC [11, 9] simultaneously solves for the transformation and slide identity. To do so, SIFT (Scale invariant feature transform) [17] keypoints extracted from both the slides and the

video frames are matched subject to the constraint of consistent homography using RANSAC [12]. This approach is fully automatic and can handle video capture systems with few limits on the motion (pan, tilt, zoom and even motion) of the cameras.

To gain robustness, SLIC combines the spatial matching with camera event detection, and a dynamic HMM based on camera events. In addition, the spatial matching is done in several steps. In this first phase, frame-to-slide homographies (i.e., the projection transformation between the slides captured by camera and their original ones) are found for some slides (the easier ones) by keypoint matching. Using this information, all frames are classified into three categories: *full-slide* (the entire frame shows the slide content), *small-slide* (the frame contains both a slide area and a substantial portion of the scene back-ground), and *no-slide*. This classification, and associated homography estimates for all frames are then used to match all frames quickly using *local keypoint matching* which uses the estimated homographies to limit the possible matches to geometrically plausible ones. A second phase detects camera event between each pair of consecutive frames by using the computed homographies and the frame types classified in the first phase. Finally, in the third phase, the visual features, temporal information and the camera events are incorporated into a dynamic HMM to provide an optimal sequence of slides matching the frame sequence.

Besides being a very effective way to achieve the alignment, the homographies extracted while doing so have many uses. Note that the simultaneous method allows the homographies to be extracted even when the slide is zoomed in, and corners are not available. This is in contrast to the work of Gigonzac et al. [13] that focusses on video enhancement opportunities enabled by knowing the homographies, and uses detected corners to determine the homography.

The method developed by Wang and Kankanhalli [29] share with SLIC the usage of SIFT feature points and RANSAC. However, they place more emphasis on handling cases where the use of SIFT points alone creates problems. First, SIFT is defined on gray-scale level, so differences based on colors could be overlooked. Secondly, their system better addresses slides with animations, and, by using color gradients, also corrects slides that appear blurry. The latter case occurs for example when the camera is focused on the lecturer, and the screen is in the background. Similar to SLIC, after finding the matching keypoints, this methods also uses an HMM to further disambiguate the temporal matching — estimating which slide is shown during each video frame.

1.3 Characterizing the slide channel

The slides channel emerges once video frames are linked to the slides that were used in the presentation. Once this is done, text extracted from the slides can

be used as semantic handles into the content. This is especially effective for two reasons. First, words can be extracted from the slides reliably compared with the obvious alternative of using automated speech recognition. This is especially true of technical words which are, ironically, the most important ones to get right. Second, slides are typically designed to summarize content, and words extracted from them are more likely to be good words for indexing.

The slides channel further enables segmenting the video into semantic chunks that are searchable and browsable by the extracted words. Further, for browsing, the images of the slides themselves provide effective icons for the chunk, as used in the SLIC system.

Finally, we consider the full use of the slides channel to include knowing the geometric mapping between the slide image in the frames, which is a linear transformation in homogeneous coordinates, specifically a homography [14]. Knowing this geometry means that user pointing gestures with respect to the video frame can be interpreted in the coordinate frame of the slide. Hence one knows, for example, what slide word is clicked, or which slide region should be magnified Further, slides can be backprojected into the video to improve its quality, or allow for better compression for studying the presentation with low bandwidth.

1.4 Applications of the Slides Channel

Video Browsing: Assisting a scholar in the browsing process is an excellent example in which cross-modality expresses itself:

We are accustomed to searching and finding text, or even presentation slides, on the web that are of interest to us. However, comparable capability in the case of video—notably educational videos—does not exist. The familiar (and sometimes awkward) process of finding textual information by retrieving many pages, and then sifting through the hits to find items of interest, completely breaks down in the case of video, even if transcripts are available and indexed. We usually have to either watch the video in its entirety or try fast-forwarding in our attempt to locate the information in which we are interested. More sophisticated video search tools, such as CueVideo [1], provide only modest help, and are not available to most users.

By presenting the slides to the students, and enabling them to browse through the slides and watch the video from the points at which these slides are discussed, SLIC is able to give the student efficient access to relevant material. Please refer to Figure 1.2. This is helpful for a video of a single lecture, and even more so when the student is browsing slides of multiple lectures — e.g. an entire course. Moreover, a textual search of the text of the slides is far more accurate than a textual search of a transcript of the course.

Slide backprojection: The term *backprojection* in this context describes the process of substituting the slide image in the video frames with high resolution

images from external slides. Subsequently, text and details of the slides that are too small and/or too blurry in the original video would appear much clearer and readable after backprojecting the slides. To successfully apply backprojection, the SLIC system relies on the high accuracy of homography described above between video frame pixels and external slide pixels. For each pixel $F_{i_{(x,y)}}$ in the frame region of frame i of the video, the system interpolates a color from the slide image S_j. Figure 1.2 demonstrates the improvement in the quality and readability of the original frame (bottom), compared to the backprojected frame (top). Note that the projected slide region is sharper and the color is improved.

There are several ways in which backprojection can be accomplished. One is to generate a new video that the user can access, and treat it as if it is a "regular" video. This method enables careful adjustment of pixel colors. The method is simple from the user's standpoint. It can also assist in colors correction to best fit the original slide appearance. A slightly different method was presented in [30], based on sending to the user the (images of the) slides, the video of the lecture and the homography for each frame. An application on the client (user) side, for example SMIL (Synchronized Multimedia Integration Language) plays the video, while presenting the image of the slide as part of the frame seen by the user. The settings of the slide in the frame — its location and scales are determined by the homography. This method yields more visual artifacts, but yields a more detailed and sharper image. This method is more suitable for mobile applications, where low bandwidth limits the communication. It allows for higher compression while ensuring high quality where it is needed. The loss of detail from aggressive compression can be directed to less significant parts of the video such as the audience in the classroom, while the slide appearing in the video will stay sharp. Therefore, this technique significantly improves delivery and display of educational video in mobile devices with limited bandwidth availability and small screens such as PDAs or Smartphones.

Application to on-demand magnification: SLIC uses the slide-frame homographies to provide a magnifier utility, which provides a sharply magnified version of the slide region around a point clicked on by the user. This is very useful when the details of the slides appearing in the video are too blurry or too small. Notice that getting such detail from a slide image in some other window is often a poor alternative, as the user must then switch between the two and determine the correct location in the slide, losing the connection to the video context. Instead, in the SLIC system, once the user clicks on a specific point of the video player, its coordinates are back-transformed to the slide, providing the center of the region of (the image of) the slide of interest, which is then displayed with the desired enlargement factor. This is distinctly different than attempting to magnify the lower quality video and/or attempting to sharpen it (e.g. by the magnifying glass utility of Windows XP). With this approach, the only limitation on sharpness is the level

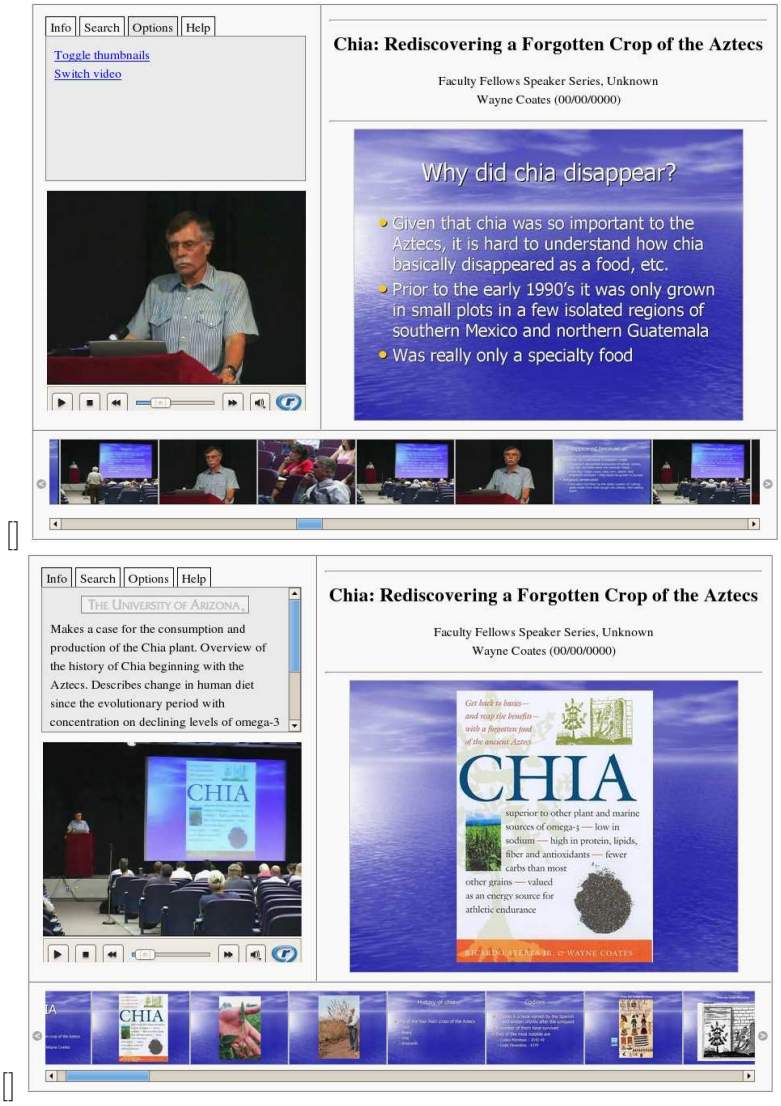

Figure 1.1: Two ways of browsing videos [26]: a) by keyframes and b) by slides. Notice the difference at the bottom thumbnails slider. Slide changes provide a semantic video segmentation and are therefore more desirable for video browsing than keyframes extracted by shot boundary detection.

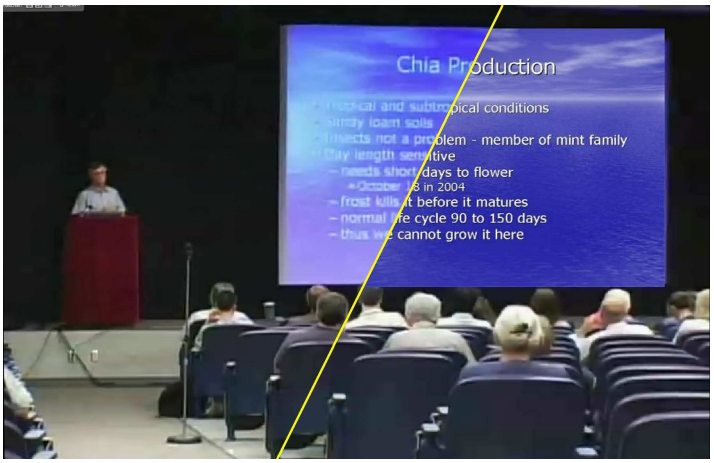

Figure 1.2: Comparison between the original video frame, vs. frame after backprojection of the slide. The later is clearly much more readable

of detail present in the original slide.

Laser Pointer Motion Semantics Laser pointers are frequently used by speakers in lectures to indicate a specific area of interest on a slide. Understanding where the laser pointer is directed on the video frame and hence the presentation slide can give us a good idea of what the main topic of focus is at that instant in time. The detection of laser pointer light is therefore of value for a good system for searching and browsing lectures.

In [30] the authors describes how to track the laser pointer robustly, and the spectrum of applications it enables. Since input video comes from external sources, scene brightness and color is unknown and frame differencing is used to remove these variations. For each frame of the video, a second frame is computed as the difference between the corresponding frame and a pixel-wise average of a set of proceeding frames in the original video. A median filter on brightness approximately equal in size to the laser pointer is then passed over each frame of the difference video. The brightest pixel from the resulting image is then selected. Curve fitting is used both for noise filtering and as input for following steps.

The applications of the tracking include: (i) Seeking the instance in the video where a queried keyword is *emphasized* by the lecturer, and (ii) it is also used for enhancing, changing background colors, or enlarging ("zooming-in") portions of the relevant slides. The enlargement effect is especially useful for displaying on a small screen (e.g. smartphones) or helping visually impaired students.

1.5 Cross-Modality and the Internet.

All applications mentioned in the previous section fail to utilize the real potential of studying in an environment so rich of data as the internet. Seeing the variety of data sources on the web, we face two challenges. One is how to be an efficient consumer of this data, so a student watching an educational video could follow links that we would embed in the video and in the video browser, accurately finding most-relevant resources. The second challenge is to be good providers of the data withing videos created by ourselves and others in the community. Having the video deposited and accessible for view does not go far enough. We want to ensure that interested scholars can find exactly what they are looking for. What is more, we want the scholar to have easy access to video on their topic in alternative contexts (e.g., math presented in the context of math classes, versus application oriented discussion) and styles of presentation. The search for relevant video segments should not be harder than a search for relevant documents. As an analogy, we do not expect a scholar to need to skim through an entire online document that we have created when she seeks only a small portion of it. So, why should we expect her to listen to an entire lecture if she only needs knowledge from a short segment?

We show next how the usage of the "slide channel" can address both challenges.

Linking and Selectively Browsing Educational Videos from Multiple Sources: Once an educational video is processed by a system like SLIC, it can be linked to other online-available educational materials. First, keywords extracted from the video's slides form searchable terms that can be linked, for example, to standard internet sources such as Wikipedia and electronic books. Results could be further filtered and refined using speech recognition tools and context similarity. These links are embedded in the videos so that students browsing them can efficiently link to keyword-based content.

Second, the matching can link to other relevant educational videos that are available online. These videos may be located, for example, on the educational channels of Youtube, through Google searches of publicly available lectures, or via contributions to a repository that could potentially be formed by SLIC users willing to share videos with one another. Given the universe of potentially relevant videos, we have two tasks: to identify which videos are substantively relevant, and to automatically link to segments of relevant videos with appropriate supplementary content. Both tasks are easier if slides on which videos are based are separately available, since matching slides to other slides is much more robust than matching videos directly. Thus it seems quite feasible to make use of the slides channel, even without having the slides available as a separate data source. Here slides would be extracted and linked simultaneously. This is a research direction that we are actively pursuing.

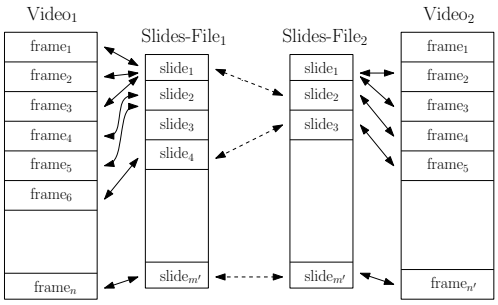

Figure 1.3: This example demonstrates the potential of accurately linking frames of different videos once each video is linked to its slides file. Here we assume that the matching between Video$_1$ to Slides-File$_1$, and between Video$_2$ to Slides-File$_2$ are computed using for example SLIC, while the matching between Slide-File$_1$ and Slide-File$_2$ is computed based on textural and image matching. Hence implicitly obtaining the matching between Video$_1$ to Video$_2$. So for example in this illustration frame$_6$ in Video$_1$ is matched to frame$_5$ in Video$_2$.

Once slides are matched, the cross-modality nature of the materials provide data that is used to rank the overall relevance of a particular video. The two sets of slides from the original video and the target online video are then matched to each other. From these slide-matchings, we are capable of directly matching video segments, and embedding links within the material the user is browsing.

Making Videos Accessible via Internet Search Engines By piggybacking SLIC search mechanisms onto common search engines such as Google or Bing, content indexed by SLIC can be made widely and seamlessly available to internet searchers. Currently, a user searching for educational content will at best receive a link to powerpoint slides, if available on the web, or less commonly, a link to an entire video whose metadata happens to contain the search term. The user must then manually browse the slides and/or video to find relevant content, a tedious and potentially fruitless endeavor. The slides channel from SLIC solves the problem of indexing and locating relevant content by leveraging the natural partition of a video into meaningful *video segments* based on the slides the video contains.

SLIC can be extended to enable accurate search for video segments stored within the SLIC from anywhere on the web. A user can simply submit a query, and receive an appropriate video segment, with the accompanying slide in reply. The mechanism that allows search engines to index SLIC content is that each video has a website with meta-data indexing video segments and the text of the associated slide. Once a search engine such as Bing or Google invokes a SLIC webpage, the website in turn retrieves the text of the query from the search engine, and passes the query to SLIC. Next, SLIC uses its internal query mechanism to find the most appropriate

video segment and play it. From the user's perspective, this process is transparent: she simply enters a query in a search engine, and is taken directly to the relevant segment of the SLIC website, where the video will begin automatically playing. The user will also be able to see and browse neighboring segments, and enjoy the other useful properties of SLIC mentioned above. In short, a searcher anywhere on the web can be just one click away of viewing otherwise difficult to locate and highly relevant educational content.

References

[1] "Cuevideo." [Online]. Available: http://www.almaden.ibm.com/projects/cuevideo.shtml

[2] G. D. Abowd, C. G. Atkeson, A. Feinstein, C. E. Hmelo, R. Kooper, S. Long, N. N. Sawhney, and M. Tani, "Teaching and learning as multimedia authoring: The classroom 2000 project," in *ACM Multimedia*, 1996, pp. 187–198. [Online]. Available: citeseer.ist.psu.edu/abowd96teaching.html

[3] A. Amir, G. Ashour, and S. Srinivasan, "Automatic generation of conf. video proceedings," in *Journal of Visual Communication and Image Representation, JVCI Special Issue on Multimedia Databases*, 2004, pp. 467–488.

[4] A. Behera, D. Lalanne, and R. Ingold, "Looking at projected documents: Event detection document identification," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2004.

[5] Y. R. C. Zhang, J. Crawford and L. He, "An automated end-to-end lecture capturing and broadcasting system," in *ACM Multimedia*, 2005, pp. 808–809.

[6] Y. Chen and W. J. Heng, "Automatic synchronization of speech transcript and slides in presentation," in *International Symposium on Circuits and Systems (ISCAS)*, 2003, pp. 568–571.

[7] T. S. Consortium, "Maing the grade: Online education in the united states, 2006," in *http://www.sloan.org/publications/survey/survey06.asp*, 2006. [Online]. Available: http://www.sloan.org/publications/survey/survey06.asp

[8] B. Erol, J. J. Hull, and D. Lee, "Linking multimedia presentations with their symbolic source documents: algorithm and applications." in *ACM Multimedia*, 2003, pp. 498–507.

[9] Q. Fan, A. Amir, K. Barnard, R. Swaminathan, and A. Efrat, "Temporal modeling of slide change in presentation videos," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[10] Q. Fan, K. Barnard, A. Amir, and A. Efrat, "Accurate alignment of presentation slides with educational video," in *IEEE International Conference on Multimedia & Expo (ICME)*, 2009. [Online]. Available: http://kobus.ca/research/publications/09/icme-09-bundle.pdf

[11] Q. Fan, K. Barnard, A. Amir, A. Efrat, and M. Lin, "Matching slides to presentation videos using sift and scene background matching," in *8th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006.

[12] A. Fischler, M. and C. Bolles, R., "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. of the ACM*, vol. 24, pp. 381–395, 1981.

[13] G. Gigonzac, F. Pitie, and A. Kokaram, "Electronic slide matching and enhancement of a lecture video," *IET Conference Publications*, vol. 2007, no. CP534, pp. 9–9, 2007. [Online]. Available: http://link.aip.org/link/abstract/IEECPS/v2007/iCP534/p9/s1

[14] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[15] T. Liu, R. Hjelsvold, and R. Kender, J, "Analysis and enhancement of videos of electronic slide presentations," *IEEE International Conference on Multimedia and Expo (ICME)*, 2002.

[16] M. Liwicki and H. Bunke, *Recognition of Whiteboard Notes: Online, Offline and Combination*. World Scientific Pub Co Inc, 2008.

[17] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 20, pp. 91–110, 2004. [Online]. Available: citeseer.ist.psu.edu/lowe04distinctive.html

[18] MediaSite, "Mediasite," in *http://www.mediasite.com*, 2004.

[19] MSproducer, "Microsoft producer: - a capturing and authoring tool for distance learning," in *http://www.microsoft.com/windows/windowsmedia/technologies/producer.mspx*, 2003.

[20] S. Mukhopadhyay and B. Smith, "Passive capture and structuring of lectures," in *ACM Multimedia (1)*, 1999, pp. 477–487. [Online]. Available: citeseer.ist.psu.edu/485766.html

[21] J. M. Rowe, L. A. andGonzelez, "Bmrc lecture browsers," in *http://bmrc.berkekey.edu/frame/projects/lb/index.html*, 1999. [Online]. Available: http://bmrc.berkekey.edu/frame/projects/lb/index.html

[22] T. Russell, *The No Significant Difference Phenomenon as Reported in 355 Research Reports, Summaries and Papers*. North Carolina State University Press, 1999.

[23] C. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[24] T. F. Syeda-Mahmood, "Indexing for topics in videos using foils," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. II: 312–319.

[25] J. Torkkola, Q. Fan, R. Swaminathan, A. Winslow, K. Barnard, A. Amir, A. Efrat, C. Gniady, and S. Fong, "The slic video browsing system demo." [Online]. Available: http://vision.cs.arizona.edu/SLIC

[26] ——, "The slic video browsing system demo," 2005. [Online]. Available: http://vision.cs.arizona.edu/SLIC

[27] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment – A modern synthesis," in *Vision Algorithms: Theory and Practice*, ser. LNCS, W. Triggs, A. Zisserman, and R. Szeliski, Eds. Springer Verlag, 2000, pp. 298–375. [Online]. Available: citeseer.ist.psu.edu/triggs00bundle.html

[28] F. Wang, C.-W. Ngo, and T.-C. Pong, "Synchronization of lecture videos and electronic slides by video text analysis." in *ACM Multimedia*, 2003, pp. 315–318.

[29] X. Wang and M. Kankanhalli, "Robust Alignment of Presentation Videos with Slides," in *Advances in Multimedia Informatioin Processing-Pcm 2009: 10th Pacific Rim Conference on Multimedia, Bangkok, Thailand, December 15-18, 2009. Proceedings*. Springer-Verlag New York Inc, 2009, pp. 311–322.

[30] A. Winslow, Q. Tung, Q. Fan, J. Torkkola, R. Swaminathan, K. Barnard, A. Amir, A. Efrat, and C. Gniady, "Studying On The MoveEnriched Presentation Video For Mobile Devices," 2009.