Title: A System for Placing Videotaped and Digital Lectures On-line
Category: Multimedia Authors: Michael N. Wallick, Niels da Vitoria Lobo, Mubarak Shah
Affliations: University of Central Florida, School of Electrical Engineering and Computer Science; Computer Vision Lab
Contact Author: Michael N. Wallick
Mailing Address:
c/o Niels Lobo
University of Central Florida
Computer Science Bldg.
4000 University Blvd
Orlando, FL USA 32817

Phone Number: 407 823-4733
Fax Number: 407 823-5835
email address: michaelw@cs.ucf.edu

A System for Placing Videotaped and Digital Lectures On-line

Michael N. Wallick, Niels da Vitoria Lobo, Mubarak Shah
School of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL 32816
{michaelw, niels, shah}@cs.ucf.edu

Abstract

The idea of distance education, or learning at home has become very popular in recent years. Most major Universities offer classes on-line, and several smaller institutes exclusively offer distance classes. The distance courses can be offered either by mail (videotapes and books) or over the Internet (web pages). The web-based courses are limited by of bandwidth restrictions. This makes complex, interactive, presentation of information virtually impossible. Videotapes are able to present complex information, however the production and distribution cost of videotapes makes the idea very prohibitive. This paper will present a system for compressing video lectures for Internet transmission, without losing classroom information. In addition, we present algorithms that improve the interactivity of the presentation by use of Optical Character Recognition and Speech Recognition technologies.

1 Introduction

Multimedia technology in the classroom and corporate presentations have become very popular in the past several years. These technologies, including overhead and computer aided projectors, are used in just about every lecture and presentation today [3]. The goal of this paper is to take a video of such a lecture and extract the "key frames" from the lecture. Those frames represent a compressed version of the lecture, which can be easily transmitted via the Internet or other means [4, 5], without using massive storage or bandwidth. A digital lecture is one in which the presentation is generated by a computer, such as a PowerPoint presentation. In this case, each slide in the presentation can be used as a key frame. The text and graphics contained in each frame, as well as the audio associated with each frame are extracted for later analysis and retrieval.

One such distance education system is the FEEDS (Florida Engineering Education Delivery System) Program [8]. A professor lectures to a live class, using various multimedia tools, such as a computer presentation, while the lecture is being recorded on videotape. The videotapes are shared with several universities around Florida as well as businesses to use in continuing education programs. While this is an improvement over the text-based classes, there are still several problems with this approach. Problems include the cost of the videotapes, as well as shipping costs in order to distribute the videos. Additionally, the lectures are only archived for about two weeks, because of the physical storage requirements of the videos. All of these limitations can be overcome by distributing the lectures over the Internet. However, transferring the classes over the Internet will either require massive amounts of bandwidth or compression. Standard compression techniques would destroy the quality of the lecture. Our paper proposes a method for compressing video lectures, without losing classroom information. This method leads to a natural partition of the lecture. We then apply Optical Character and Speech Recognition to each partition to mine information from the video tape. That information mining leads to a more interactive lecture.

2 Related Work

In recent years, there has been a boom in distance education. Students are no longer confined to physical meeting places or times. Instead they are free to study at convenient times. Distance education can be carried out in many ways, including over the Internet and videotaped lectures. Recent work [6, 7] has begun to investigate ways in which the Internet can be best used in order to present classes on-line.

In [1], we presented a system which proposes a method for extracting and analyzing the text and

graphics from a still image of an overhead or computer projector. We now present a system which combines the method in [1] with the methods proposed here, in order to mine data from the video tape. We can then even further compress the projector presentation from a traditional class lecture into a simple ASCII file. Since text is very small by comparison to a picture file, the original video of a lecture can be compressed and viewed over the Internet, without a significant loss of detail or information.

Researchers at IBM (Almaden) are currently working on a system known as CueVideo [2] that does have distance learning applications. CueVideo is a system that is designed to be able to segment a videotaped lecture based on overhead projections and spoken audio by the lecturer. The system can then compress the video based on the segmentation for easy downloading over the web. However, their implementation of the system requires several constraints, including that the lecturer "turn-in" their overhead slides for separate processing. In a real world situation most lecturers would not be willing to do this.

3 Overview

The remainder of this paper will discuss the various methods used to mine data from the lecture (either video tape or digital) in order to be able to place the lecture on-line. This will include finding the key frames in the video. After that, we will introduce the concept of a "lecturelet" in order to extract written text, graphics and the audio text from the lecture. At each point we will discuss any differences in the method between using a videotape and a digital presentation.

4 Key Framing

If we are using a videotape input, then our system assumes that there is some video camera pointed towards a projection screen, with no obstructions between the screen and the camera. There is no restriction on what is generating the projection (i.e., an overhead projector, slide projector, computer projector, etc.). This is the exact setup of the FEEDS classrooms as discussed earlier.

Since the information contained in a lecture is not changing rapidly, and many times not at all (such as a videotaped PowerPoint Presentation, where the same slide can be displayed for several minutes), it is not necessary to transmit or store 30 frames per second. Instead, we compress the lecture by use of key frames. In this context, a key frame is defined to be the first frame of video containing new information. Some examples of key frames in a computer generated presentation are a new line of text, a new graphic or a new "slide" altogether. In the case of a lecture where the presenter is writing and that writing is in turn displayed on the screen, a key frame is when new information is added to the projection or when the presenter points to a particular area of the projection. The following is the algorithm that we use to determine if a frame is a key frame or not:

1. The first frame is a key frame
2. While receiving video input do
   (a) perform difference operation between the previous frame and current frame
   (b) if difference operation returns true, perform difference operation on current frame and most recent key frame
   (c) if second difference operation returns true, found key frame, record time that key frame is detected

The difference operation is as follows:

$$PicDiff = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} PicDiffIsBig(Pic1_{ij} - Pic2_{ij})}{m \times n} \times 100$$

Where Pic1 and Pic2 are the two images in question, m and n are the dimensions of the pictures. PicDiffIsBig is a predicate that is 1 iff

$$|Pic1_{ij} - Pic2_{ij}| \leq tol$$

where *tol* is some number close to zero. The difference operation returns true is PicDiff is greater than 7.5, which we have experimentally found to represent significant change.

This algorithm works extremely well for all types of projections. In the case of a non-computer generated presentation, the lecturer may point to a certain region of the projection; depending on the amount of motion, this may or may not be detected as a key frame. We have not found any other cases of noticeable loss of classroom information.

Since this step requires the computer to keep only three frames in memory, and this process runs quickly, it can be performed as an online operation. The video signal simply needs to be connected to the computer, and at the conclusion of the lecture the key frames have been extracted. In order to replay the lecture,

the key frames are displayed in order for the correct amount of time, along with the audio from the lecture. The compression of the lecture depends on how many key frames were generated. On average, a 1 hour lecture was about 50 megabytes, however this varied from 20 to 80 megabytes, with most of the storage cost being the audio.

Figure 1 shows four consecutive key frames that were extracted from a lecture in which the presenter was writing the information that was being displayed on the projector screen. A new key frame is extracted approximately after each word is written. Figure 2 shows four consecutive key frames in a videotaped lecture where the presenter used a PowerPoint computer generated presentation. A key frame is extracted each time the presenter shows a new slide.

5 Digital Presentations

Often the presentation will be in a digital form (PowerPoint) either in addition to or instead of a videotape. If we have a digital version of the presentation, we can use the slides in the presentation as key frames. In order to do this, we need to have the timing information for each slide, as well as the audio from the lecture. Given all of this information, we would end up with an identical result as if we were to use a videotape.

6 Lecturelets

The result of the key framing algorithm is that the video has been segmented into a discrete unit of lecture, or "lecturelet." Looking at each lecturelet individually can yield more information than looking at the entire lecture as a single unit. This section explains some of the operations that can be performed on a lecturelet in order to extract information.

The first operation that can be performed on lecturelets is the union. When all of the lecturelets from one lecture are unioned together (with order being preserved), we get back the entire lecture. Other operations from set theory, such as difference, intersection, etc. can also be defined on the information contained in each lecturelet. An intersection can tell how much information lecturelets have in common. By removing one or more components of the lecturelet (such as order of appearance), a difference operation can tell if two or more lecturelets are the same, or the same information is being repeated.

6.1 Extracting Video Text and Graphics

Each lecturelet has one image (the key frame) associated with it. By applying Optical Character Recognition we are able to extract the text that is contained in each key frame. Since the key frame will contain unconstrained information, we will need to apply some method to extract the text and graphics. This is done by the following:

1. Find edges in image
2. Determine connected components for the edges
3. Perform segmentation on each connected component
4. Process each segmented component through OCR to determine text or graphic
5. Perform Optical Character Recognition over entire set of text regions

The method presented in [1] describes this process in greater detail. Using this method, we are able to extract the text and graphics that are contained in a key frame image. By applying this method to each lecturelet, we can even further compress the contents of the lecture, down to simple ASCII text ("video text") and small graphics. In addition, this sets up one means of searching through the lecture based on content, a task that could not be performed with a video tape. If a digital presentation is used, then the task of extracting can be greatly simplified, as we can get the text and graphics directly from the digital presentation.

6.2 Extracting Audio Text

A search of the video text is not the only search that can be performed. We have also used the temporal information contained in each lecturelet to break apart the audio of the lecture. We than perform speech recognition on the audio for each lecturelet. The result is a file containing all of the recognized speech during that portion of the lecture. Since speech recognition is usually unable to perform correct grammar parsing we applied filtering to the speech file. All of the punctuation marks (periods, commas, etc.) are removed, as well all common words such as "and," "if," and "but." The words that remain after the filtering are associated with the lecturelet as "audio text." The same operations can be performed on the audio text as with the video text.

6.3 Expanding Lecturelets

Each lecturelet has associated with it: order of appearance, display time, key frame image, audio, video text and graphics, and audio text. This definition is, however, open-ended and can be supplemented as new methods for extracting information from video are introduced. Each new component can add to the functionality of the lecturelet design and the interactivity of the on-line lecture.

7 Conclusion

This paper presented a method for improving distance education over the Internet, by use of a key framing compression scheme. Videotaped lectures can be compressed without losing classroom information, and still be small enough to be placed on-line. We can also use the slides of a digital presentation as key frames, and extract the same information as we do with a videotape. In addition, we presented the idea of a lecturelet, which allows a lecture to be divided into discrete units, and have information easily extracted from each lecturelet, rather than the lecture as a whole. The lecturelet idea naturally supports additional components as new methods for extracting information from video is discovered.

References

[1] Michael N. Wallick; Niels da Vitoria Lobo; Mubarak Shah, "Computer Vision Framework for Analyzing Projections from Video of Lectures." *Proceedings of the ISCA 9th International Confernece for Intellegent Systems*, Louisville, KY, June, 2000.

[2] Syeda-Mahmood, T.; Srinivasan, S.; et. al. "CueVideo: A System for Cross-Modal Search and Browse of Video Databases." *Computer Vision and Pattern Recognition (CVPR)*, Hilton Head Island, South Carolina, June 2000.

[3] Burhalew, Chris Dr. and Porter, Alan, "The Lecturer's Assistant." *SIGCSE Bulletin*, Volume 26, Number 1. Phoenix, Arizona, March 1994.

[4] Robler, Toms; Fernndez, David; et. al. "Using Multimedia Communication Technologies in Distance Learning." *SIGCSE Bulletin*, Volume 29 Number 23. Uppsala, Sweden, September 1997.

[5] Murray W. Goldberg and Sasan Salari, "An Update on WebCT (World-Wide-Web Course Tools) - a Tool for the Creation of Sophisticated Web-Based Learning Environments." *Proceedings of NAUWeb '97 - Current Practices in Web-Based Course Development*, Flagstaff, Arizona June 1997.

[6] Ma, Wei-hsiu; Lee, Yen-Jen; et. al. "Video-Based Hypermedia for Education-on-Demand." *IEEE Multimeda, Volume 5, Issue 1.* Jan - March 1998. Pages 72 - 83.

[7] Siddiqui, K.J.; Zubairi, J.A. "Distance Learning Using Web-based Multimedia Environment." *Proceedings of Working Conference on Academia/Industry Research Challeges 2000* 2000, Pages 325-330.

[8] Florida Engineering Education Delivery System http://feeds.engr.ucf.edu

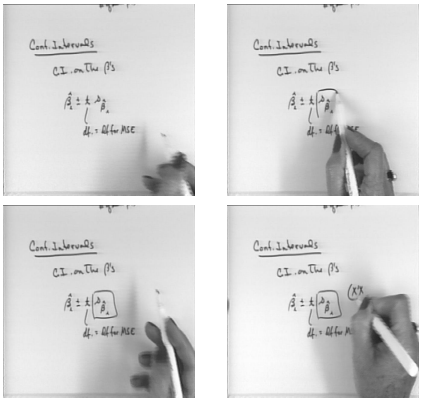

Figure 1: Four consecutive key frames from a presentation in which the lecture was writing the information to be displayed. Each key frame occurs approximately after one new word was written.

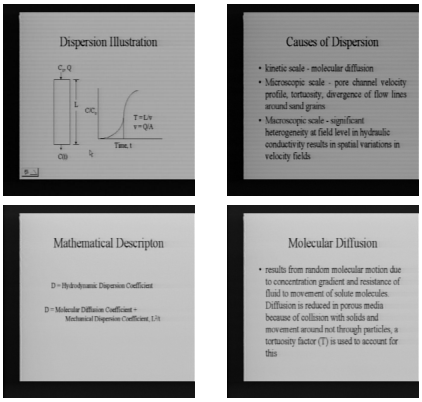

Figure 2: Four consecutive key frames from a presentation that used a PowerPoint computer generated presentation. Each key frame occurs once a new slide is shown.