Automatic Creation of Indexed Presentations from Classroom Lectures

Paul E. Dickson, W. Richards Adrion, and Allen R. Hanson
Department of Computer Science
Computer Science Building
University of Massachusetts
Amherst, MA 01003, USA
{pauld,adrion,hanson}@cs.umass.edu

ABSTRACT
This paper describes a system designed to automatically capture classroom events as videos and images. This content is delivered in several ways, most commonly as indexed multimedia presentations but also in real time as notes of classroom events. This content creation system identifies when significant events occur, e.g., material presented by computer and projected on a screen or written on a standard whiteboard, and saves these events as enhanced images. In parallel with the whiteboard capture, a digitally-zoomed video of the speaker is created. The significant event images (from cameras and computers) are used to create an index into the video and the images, video and index are compiled into a Flash presentation. These presentations are used by on-campus or distance students. The event images can also be stored and exported to a Ubiquitous Presenter-style server that provides students with real-time, in-class access. The event images and video are recorded transparently to the lecturer. The lecturer need not make any modifications to teaching style or modality (whiteboard, computer-based presentation, or a combination). The primary focus of this paper is on event image and video capture techniques. The lecture capture system has great benefits for education and we report some initial experience using it in support of computer science curricula.

Categories and Subject Descriptors
H.5.1 [Multimedia Information Systems]: Video; I.4.1 [Digitization and Image Capture]: Sampling; I.2.10 [-Vision and Scene Understanding]: Video analysis; K.3.0 [Computers and Education]: General

General Terms
Algorithms, Design

Copyright 2008 ACM 1-59593-057-4/05/0004 ...$5.00.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

1. INTRODUCTION
The course of education is changing with more technology being introduced as teaching aids. PowerPoint and other presentation technologies are ubiquitous, in-class demonstrations are common-place, and tablet computers are coming into broad favor. Instructors can now create quality lectures with dynamic demonstrations, incorporate materials from textbook and other websites, and access web resources in the classroom. These technologies support reuse and continuous improvement. Notes and presentations can be provided as print materials so that students can focus on the lecture instead of taking notes. Tablet computers offer an opportunity to provide annotations and extemporary notes that can be recorded or even distributed to students with systems such as Classroom-Presenter [1]. When an instructor dynamically alters a lecture in response to student questions or engages students in real-time demonstrations or information access, the pre-printed materials will be inadequate. Tablet computers, screen capture systems, e.g., Camtasia [9], “smart” whiteboards, and human operated (and automated) video recording offer some options for capturing classroom events. No commercial systems, however, give an instructor the freedom to use computer-based presentations, whiteboards, overhead projectors, etc. in any combination without special software, hardware, or training.

For more than a decade, the RIPPLES group at the University of Massachusetts Amherst has developed content delivery systems (MANIC) with a goal of placing no restraint on teaching style, pedagogy or specific technology. In the past, we depended on content captured by human operators in instrumented video classrooms. This was an expensive approach requiring substantial post-production editing, but was effective [7] for providing on-campus students with materials for study and review and for supporting distance education. In this paper we describe a classroom recording system that automatically creates indexed and searchable multimedia presentations using a combination of high-resolution cameras, a computer screen capture device, and a wireless microphone. The output presentations include a video of the lecturer, enhancements of all material presented by computer or written/drawn on a whiteboard, and an index to support easy search and navigation. The captured content is delivered as a Flash presentation as shown in our
is designed to automatically create rich media content. A substantial body of work exists on automatically capturing lectures, with most focusing on automatic camera management in a “seminar format,” [6, 5] and compiling the captured content into “record-and-playback” formats including a video of the lecturer and images of slides. Other work focused on methods of capturing more varied sources than just slides; these will be discussed in comparison with our system in Sections 2.2 and 2.3.

2. ALGORITHM

In the paper, we will describe our system by activity: lecturer video creation, computer capture, and whiteboard capture. Both the computer and whiteboard capture systems use image processing techniques to identify the occurrence of significant events within their individual input streams. The system saves images of these events and the times when they occur. Significant events are considered to have occurred any time the material presented has changed from the previous stored significant event. Significant images are stored once the presented material becomes stable. Images from the high-resolution cameras are used to create the lecturer video and capture whiteboard images. Thumbnails of the images and the times when they occur are used to build the index, which is combined with the full-size images and the video to create the full Flash presentation shown in Figure 1; note that in this illustration we show a numeric rather than thumbnail index.

Within each activity, we describe the results of a number of empirical evaluations of the capture systems. We will describe our recent efforts to carry out usability studies on the GUI, the measures for “significance” (and examples of “significant events”) and how well the multimedia presentation represents the actual classroom.

2.1 System Setup

The basic design of the system is shown in Figure 2. The computer capture device is an Epiphan Systems VGA2USB. The cameras are Point Grey Research Flea2 color cameras with a resolution of 1024x768 pixels. Only the left camera in Figure 2 was functional as this paper was being written and it was run at a resolution of 1024x410 pixels to increase frame rate because of hard drive bandwidth constraints in our test server. Each camera covers 12 feet of horizontal board space, giving us a resolution of just over 8.5 pixels per inch. The server used to capture the camera images was a 3.0 GHz Pentium Dual-Core machine running Windows XP. We are currently upgrading to a faster server better able to handle the processing tasks.

2.2 Computer Capture

The computer capture software performs image processing on a steady stream of input frames provided by the VGA2USB to determine when significant events occur. Using image processing to determine significance enables us to find significant events in any material presented, without constraining the lecturer. The algorithm uses a three-stage process to determine significant images. The first pass runs on each screen image as it is captured and maintains a count of consecutive near-identical images. When the new image is different from the old image, it checks the number of previously stable images and if 3 or more have been stable, stores the previous image and the time it occurred (Figure 3). Un-

![Figure 1: Prototype presentation window.](image)

Research has shown [3] that recorded material presented just when students need it benefits students greatly. Our capture system enables instructors to present material dynamically, e.g., using the whiteboard for a single extemporary point in combination with a computer-based presentation. An instructor could give an entire lecture on a whiteboard and still provide students access to all material presented. Allowing students to review all class material was shown to have a positive influence on learning [2]. This system also allows easy creation of delivery systems that can easily be embedded into electronic text books and Web pages to clarify difficult points.

The focus of this paper is on content creation, specifically on our techniques for identifying and storing significant events, and using these events and event timing to create an index into a video of the lecturer. As noted, our computer capture systems compile these data into an indexed multimedia presentations or send the captured images directly to a Ubiquitous Presenter [11] server allowing students immediate access during the lecture for note taking. For the purposes of this paper, it is useful to distinguish between content capture and content delivery. As noted, we have worked on content delivery systems for over a decade, combining rich media with a substantial tool set for annotation, searching, sharing, and communicating [7]. The downside to MANIC and other “record-and-playback” systems is that they often require human editing to ensure the quality of the presentation does not degrade. Our system...
der testing conditions the screen capture device averaged 2 frames per second (fps), meaning that whatever was displayed on screen would need to be stable for approximately 1.5 seconds to be stable enough to store. This short time criterion for frame stability means that the stored images show the progression of a screen presentation, e.g., a series of animations or annotations. The goal of this pass is to eliminate most of the input images and remove most of the noise from the captured data.

Figure 2: Proposed system setup.

The second pass of the algorithm is used to remove any consecutive duplicate saved significant images. Consecutive duplicate images are sometimes stored by the first pass of the algorithm. For example, noise occurring when no screen change has occurred may cause duplicate images to be captured. The second pass checks for differences in the images from the first pass and eliminates consecutive duplicates. It uses more sophisticated and time-consuming methods of searching for differences than the first pass and eliminates any remaining noisy images. It runs repeatedly until the number of output images stabilizes. A third pass organizes the images into a format more easily compiled into the final delivered presentation.

The computer capture system typically reduced approximately 10,000 input images to around 100 output images, though these numbers varied in our test data due to differing capture conditions and the material presented. The computer capture system was able to capture significant images from lectures using slides, web browsers, text editors, and other programs as shown in Figure 4. All processing is completed within 2 minutes of the end of the lecture. More details of the computer capture system can be found in [4].

While there is no single correct choice for what is significant, one can look at “slide presentations” to develop a measure. Each slide transition (or animation) should be captured without duplicates. Our system captured over 99.8% of all significant transitions (new line, bullet, annotation) that stayed on screen for at least 2 seconds. It therefore did not store images of slides flipped through quickly to get from one section of a lecture to another. We extended this to “inked” presentations, with all annotations captured, but not every pen stroke. We are carrying out usability studies to extend and quantify our notion of “significance” but to date our results appear satisfactory in that no event is missed. The capture software stores 100% of all images that are on screen for at least 2 seconds.

Our system captures significant events in all material run on the computer unlike eClass [2] which is only able to store URLs visited and slides. Mediasite [8] captures images of all screen events that occur, but unlike our system Mediasite also stores many additional unnecessary events like mouse movements. Camtasia Studio [9] creates a dynamic video of all activity that occurs on screen but can only create an index for PowerPoint slides and requires software to be installed on the lecturer’s computer. Our system is somewhat similar to the TeleTeachingTool (TTT) [13] but we do not require lecturers to use a separate teaching server.

2.3 Whiteboard Capture

The whiteboard capture system is similar to the computer capture system, storing images every time material changes and becomes stable. The whiteboard capture system is part of the same software that creates the lecturer video; the input stream is 15 fps from the high-resolution cameras but, for speed reasons, only every 10th image is used for the whiteboard capture phase.

Unlike computer capture, whiteboard capture requires the input image frames to be preprocessed to highlight text and remove the lecturer. As shown in Figure 5, an input frame is segmented into blocks and the brightest 25% of the pixels in each block are averaged. The written material is then highlighted by exaggerating each pixel’s difference from the bright average, and the image is then sharpened. The lecturer is then located and the whiteboard image is created from the highlighted text image by removing pixels in the neighborhood of the lecturer and substituting pixels from the most recent whiteboard image in that area. This process creates a series of whiteboard images that are compared, and if the difference is greater than a threshold, the image is stored. No special process is required to check for image stabilization because the new text will not appear until the instructor is no longer blocking it while writing.

The whiteboard capture system reduces the approximately 80,000 input frames from a typical lecture to an average 72 whiteboard images stored. Of these 72 images roughly 45%
are duplicate images where the only change that occurs is a minor difference such as a shadow appearing/disappearing from the board. The impact of duplication is mitigated by the fact that it is not perceptible in the delivery system. Usability studies are underway to assess whether these extra stored images and the accompanying indexes negatively affect the created presentations.

Of the 45% of all images captured that are duplicates, 36% of the duplicates are caused by changes in the lecturer’s shadow. 12% of duplicate saves were caused by blooming in the camera image captures when an overhead projector was used, 9% were minor differences that were observed but did not constitute changes in the material presented. General noise, pen and eraser movement in the pen tray, color and brightness changes, and unidentifiable noise were recognizable categories of unnecessary saves but none represented more than 5% of the duplicates saved. 27% of the duplicates saved were caused by visible differences that did not relate to material present on the board. Of this 27%, 29% were identifiably portions of the lecturer that the software failed to remove from the saved whiteboard images.

The whiteboard capture software captures all material written on the whiteboard and shows the progression of all material presented. The captured whiteboard images are not the same quality as those created by Microsoft’s meeting recorder [12], but our system can generate images with fewer pixels per inch resolution, a larger area of whiteboard to scan, and worse lighting conditions. Our whiteboard capture system is more robust than that used by Wienecke et al. [10] as a first step in turning text on whiteboards into machine-readable text. The whiteboard material is captured without using commercially available electronic whiteboards that are expensive and often limit writing surface area.

2.4 Video Creation

The lecturer video is created by locating the lecturer in the high-resolution images and cropping regions from these images to compile into a video. Simple differencing between consecutive images can be used to determine lecturer location because only the lecturer is moving. If the number of differing pixels is greater than a threshold, the vertical and horizontal limits of the difference are found (Figure 8), used to determine the location of the lecturer, and the lecturer location (a point approximating the middle of his/her head) is updated.

The image processing software uses an 11-image (2/3-second) pipeline fed by the 15fps video stream. The frames that make up the lecturer video are cropped from the oldest image in the pipeline. After a crop location is established, it is tentatively assigned to the next oldest image in the pipeline. At the next time step the current location of the lecturer is compared with the tentative crop frame. If the lecturer location is close to the center of the crop frame, the same crop frame location is used. If the location is toward the edge of the crop frame, the software starts panning the crop frame at a rate proportional to the separation of the lecturer from the crop frame center. In successive pipeline steps the crop frame will move across each image until the lecturer location is again in the center of the frame. By doing this we are able to use future knowledge of the lec-
turer’s movements to begin panning the camera before the lecturer has actually moved and reduce the number of times the lecturer walks out of the video. Even with this fore-
knowledge the lecturer occasionally walks out of the video window. This system is designed for a single lecturer and will fail if more than one lecturer is present at the front of the room. Combining the created frames with the recorded audio currently is a post-processing step. Unlike Mediasite and Camtasia [8, 9], our system integrates camera control and does not require a lecturer camera feed to be supplied.

2.5 Content Delivery

As we have described throughout the paper, the captured images are integrated with timing information to create an index and then compiled into a Flash presentation (see Figure 1). The ubiquity of Flash makes it ideal for distance education or on-campus applications. Alternatively, we can deliver the images via a modified Ubiquitous Presenter server, although we are working to get the whiteboard image capture closer to real-time.

3. SUMMARY & FUTURE WORK

We have designed a lecture capture system capable of recording any lectures given using computer, whiteboard, or any combination thereof. The system stores all significant (using a intuitive notion of significance) material presented on either whiteboard or computer and uses the images and timing to create an index into the lecture. Our system can simultaneously capture both whiteboard and computer events, unlike the best current capture systems [8, 12] that are limited to one or the other. The system creates a digitally panned and zoomed lecturer video from the full 15fps video stream, the same stream that is sampled to capture whiteboard events. We are performing usability studies on the lecture recordings generated by the system in order to determine the quality of the results generated, and to develop tighter measures of “significance.” We need to speed up whiteboard capture so that the images can be uploaded to our modified Ubiquitous Presenter server in real time as we do with the computer captured images.

Earlier versions of our content delivery systems are effective for distance education or on-campus use. We have used the capture system in a number of classes, but not for a complete semester. Our challenge at this point is to carry out a full experiment using the new content capture and delivery systems to evaluate impact on student learning. This activity is underway in the current semester.

Most importantly, our system is transparent to the lecturer, requiring no training, changes in teaching style or pedagogy, or special software/hardware. We plan to carry out a full assessment of the efficacy of these tools from the perspective of the instructors.

4. ACKNOWLEDGMENTS

This work was partially funded by the National Science Foundation under grants CNS-0749756 and CNS-9979833, and by the University of Massachusetts under an ITC Strategic Initiative Grant.

5. REFERENCES


