Expanding the Point — Automatic Enlargement of Presentation Video Elements [*]

Qiyam Tung Ranjini Swaminathan Alon Efrat Kobus Barnard
University of Arizona
1040 E. 4th St.
Tucson, AZ 85721
{qtung, ranjini, alon, kobus}@cs.arizona.edu

ABSTRACT

We present a system that assists users in viewing videos of lectures on small screen devices, such as cell phones. It automatically identifies semantic units on the slides, such as bullets, groups of bullets, and images. As the participant views the lecture, the system magnifies the appropriate semantic unit while it is the focus of the discussion. The system makes this decision based on cues from laser pointer gestures and spoken words that are read off the slide. It then magnifies the semantic element using the slide image and the homography between the slide image and the video frame. Experiments suggest that the semantic units of laser-based events identified by our algorithm closely match those identified by humans. In the case of identifying bullets through spoken words, results are more limited but are a good starting point for more complex methods. Finally, we show that this kind of magnification has potential for improving learning of technical content from video lectures when the resolution of the video is limited, such as when being viewed on hand held devices.

Categories and Subject Descriptors

K.3 [**Computing Millieux**]: Computers and Education – *Computer-assisted instruction*

General Terms

Algorithms

Keywords

lecture, video, magnification, laser, speech

1. INTRODUCTION

Many universities offer video lectures as a way to bring classes to students who cannot physically attend courses. Examples include MIT OpenCourseWare [1], Stanford on iTunes [2], and UC Berkeley Extension Online [3]. Such online materials also benefit students who can attend classes as lecture videos are helpful for reviewing concepts. There is potential educational value in using mobile devices, such as cell phones, to watch these videos. However, as lecturers increasingly rely on electronic slides (e.g., PowerPoint) to present their topics, it also becomes important that the user should be able to read the slides in the video. This problem is particularly important when the lecturer attempts to draw students' attention to a specific semantic unit (word, bullet, or image) using laser pointers or by speaking about it. We therefore propose to automatically magnify those elements as the video is presented to the user so that the text is easily readable and therefore understandable. Our contributions are as follows:

- Identifying **semantic units** in each slide, such as bullet points, groups of bullets, and images.
- A method for robustly identifying the position of each semantic unit on a presentation slide.
- Identifying events, which are a lecturer's attempt to draw attention to a semantic unit, based on analysis of speech transcript and aligning them to times in the video corresponding to each semantic unit. We consider a spoken word *aligned* to a bullet if we know what bullet it belongs to.
- Identifying events based on laser pointer gestures.
- Augmenting the video by backprojecting an enlarged sharp image of this semantic unit, taken from the full-resolution slide.

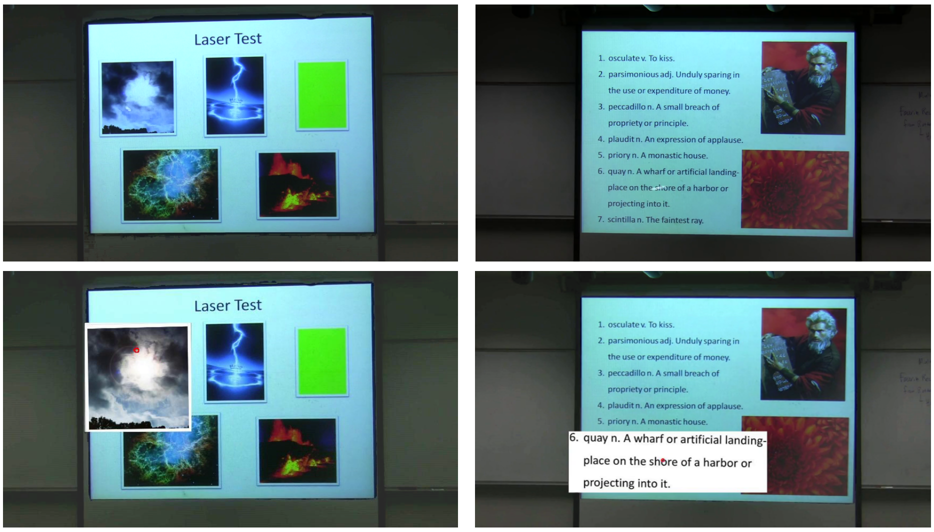
We have demonstrated the usefulness of our technique towards increasing readability of lecture videos by exposing two randomly selected groups of students to two videos, one with magnification and one without. We found that the average scores of those who watched the magnified video were higher and statistically significant. We have also tested our algorithm that detects when semantic units are highlighted by laser points and found that the identified semantics units were very similar to those identified by humans. These experiments are detailed in Section 5.

2. RELATED WORK

Several methods have been proposed for improving the quality of understanding for lecture videos. An hour-long video can be hard to navigate. One of the ways to make lecture videos more useful is by indexing it by its presenta-

[*] Area chair: Massimo Zancanaro

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...$10.00.



Laser Test

1. osculate v. To kiss.
2. parsimonious adj. Unduly sparing in the use or expenditure of money.
3. peccadillo n. A small breach of propriety or principle.
4. plaudit n. An expression of applause.
5. priory n. A monastic house.
6. quay n. A wharf or artificial landing-place on the shore of a harbor or projecting into it.
7. scintilla n. The faintest ray.

Laser Test

1. osculate v. To kiss.
2. parsimonious adj. Unduly sparing in the use or expenditure of money.
3. peccadillo n. A small breach of propriety or principle.
4. plaudit n. An expression of applause.
5. priory n. A monastic house.
6. quay n. A wharf or artificial landing-place on the shore of a harbor or projecting into it.

Figure 1: Two snapshots from videos played with (bottom) and without (top) magnification. A semantic unit (image, bullet, or word) is magnified when triggered by a laser gesture. A semantic unit is also magnified for a length of time when a word from the bullet is read.

tion slides, as shown by Fan *et al.* [4]. Their system, the Semantically Linked Instructional Content (SLIC) project, identifies slides in a video by finding the mapping, a homography, between a presentation slide and a video frame. Using this information, the SLIC system allows the users to browse the lecture by slides. Furthermore, they [5], as well as others ([6] [7]), are often able to find accurate homographies that enable them to project the slide back into the video, making the slide look clean and sharp. For our purposes, having accurate homographies allows us to determine where the semantic units are within each frame of the video. With this, we can identify when this semantic unit is being highlighted by a laser pointer as they now share the same coordinate system.

Even with backprojection, lecture material can hard to see on the small screen of a mobile device. This is a significant concern as mobile devices have long been considered as an important educational tool and much effort and development have been put into mobile learning [8]. Thornton and Houser [9] show that students benefit from using mobile devices as a learning tool. They sent e-mail lessons to students' phones to promote learning in regular intervals. They have found that 93% of the students found it a useful teaching method. There has also been success in integrating mobile devices into the classroom, as Dyson *et al.* show in [10]. Students participated in a lecture by texting responses to activities using their cell phones, giving quick feedback on the understanding of the class. These studies suggest a trend towards using smartphones for educational purposes. Our system will help enhance the understandability a lecture video.

3. IDENTIFYING SEMANTIC UNITS

In order to magnify a bullet point, we need to know where the semantic unit is within the slide. We assume the slide's position within the video frame is known to us, either by manual alignment, or using methods such as [5]. Our first step is to identify an accurate *bounding box* of a single word, bullet, or image in the slide. We do this by analyzing the original presentation files, such as Microsoft PowerPoint files. Microsoft has adopted the Office Open XML (OOXML) format since 2007 [11], which is published as an open standard. However, this format does not specify the coordinates of words or images, so we have developed a method for finding the locations based on modifying text color described in detail below, effectively identifying their positions. We have also developed a general technique that finds the boxes by parsing the presentation file, which requires minimal knowledge and assumptions about the format. If the presentation file's semantic units, such as text and images, can be found and colored, this technique can be expanded to suit many other forms. We have demonstrated it for PowerPoint files, but it is applicable to similar formats such as KeyNote and OpenOffice presentations.

Finding bullet bounding boxes. We define words to be strings of characters separated by blank spaces. A *bullet point* is similar to a word as it is an item in a list whose items start after the typographical symbol of a bullet or any other numbering scheme. Due to space constraints, we will describe the bounding boxes algorithm for just bullets. The process for words is similar.

First, we create uniquely colored bullet points in the PowerPoint file. Note that the bullets in the original presentation are not necessarily uniquely colored, so we change each bullet point's color again to create a second version with a different set of unique RGB values. The results are two PowerPoint files whose bullets are uniquely colored. The two sets of slides are exported to images. We emphasize that the coloring of the slides is only a means of identifying the bounding boxes and does not affect the slides presented to the user.

We now identify the coordinates of the corners of the bullet point's text by comparing the two corresponding images. For each image, we retain a bullet-color correspondence. Note that it is not possible to robustly find the coordinates of a bullet point with just one image. It is possible that some images or background colors match the bullet's RGB values. This observation motivates comparing each corresponding pixel of the two augmented images. When we find a color difference, we reference the table of bullet-color correspondences and identify which bullet it belongs to. This guarantees that we will find the pixels of a bullet point because only bullet points will be colored differently. Then, for each bullet, we find the minimum and maximum x and y coordinates to derive its bounding box.

Finding image bounding boxes. To find the bounding boxes for images, we adopt a similar technique. In the PowerPoint file, the images are stored in their original form. How the image is actually presented (i.e., cropped, scaled, etc.) is specified elsewhere within the PowerPoint file. This allows us to substitute an original image with a monochromatic image of arbitrary size and still have it retain the original position and size. Once this is done, we can follow the same algorithm for images as we did for bullets.

4. IDENTIFYING AND MAGNIFYING EVENTS

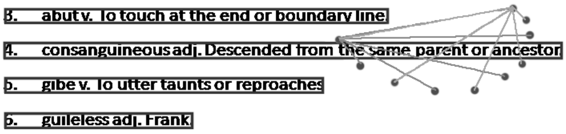
Next, we describe how to identify at which frames a semantic unit is being discussed in the lecture video. We achieve this based on speech and laser pointer use. Once identified, we magnify the element in the video frame coordinate system. We use the algorithm described by Fan *et al.* [5] to find the homographies for slides to frames and thus the time at which the slide is shown in the video. A homography is a transformation that maps points between two planes as seen by a projective camera. In other words, this operation describes the relationship between a slide and its projection in the video. The bounding boxes in the slide combined with the homography specify where the semantic units are located within the video frame coordinate system. In the following sections, we will describe how we determine at what time to magnify a semantic unit.

Speech events. In a lecture, the speaker may utter words that appear in a bullet of the slide. When the words in a bullet are read from a slide and are correctly mapped to their corresponding spoken words, we obtain times (i.e. video frame number) for when a bullet should be magnified.

Swaminathan *et al.* [12] show how to align speech transcript to the words on slides in the context of improving the transcript. This was done by creating an HMM for expected phoneme sequences for each slide, allowing for words to be skipped or additional words to be inserted, which often occurs as speakers embellish their main points. This benefits us because once we know which bullet a spoken word belongs to, it also informs us the time at which a bullet is being discussed.

From our experiments, we use the timing information accompanying the speech transcript and follow the algorithm outlined by Swaminathan *et al.* to create time boundaries for each bullet. The boundary for each bullet is created by using the minimum and maximum timestamps of all words in a bullet.

Laser pointer events. In addition to identifying speech-based events, we also identify events where the laser pointer is used to highlight bullets or images. We use the technique used by Winslow *et al.* [13] to find the laser points. Then we use the homography transformation to map the laser point to find where it appears in the slide's coordinate system.



3. abut v. To touch at the end or boundary line

4. consanguineous adj. Descended from the same parent or ancestor

5. gibe v. To utter taunts or reproaches

6. guileless adj. Frank

Figure 2: This figure illustrates how our algorithm works. The rectangles around the word indicate the bounding boxes (not part of the original slide). The points represent a laser dot sequence moving from left to right. For the sake of clarity, only a subset of all possible lines are drawn.

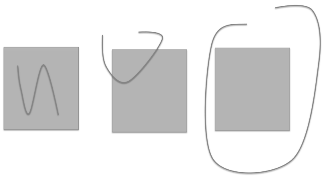


Figure 3: The curves represent examples of the paths of a laser gesture. The laser gestures can be arbitrary and do not necessarily fall inside the box. For all three cases, our algorithm can still identify which bounding box the laser is highlighting.

Our algorithm uses a voting scheme based on line intersections to determine the semantic unit being discussed. Given a laser *gesture*, a small time interval of continuous laser points, we compute a set of line segments from all the pairs of laser points (see Figure 2). This provides a notion of the movement of the laser points in the video. For example, the first two points (leftmost points in the figure) fall outside of the box. However, the segment between the two points intersects the box, so the algorithm counts that as a vote. Furthermore, this method approximates the area of a convex polygon that contains all the laser points. Thus, this method can be thought of as a method of calculating the area intersections of two polygons. In the rightmost example in Figure 3, the laser points determine a path that curves around a semantic unit's bounding box. Even though the points never fall within the box, it will still get votes from the resulting line segment intersections.

5. EXPERIMENTS

We ran three sets of experiments measuring how well the spoken words are aligned to text bullets, how accurately our algorithm identifies semantic unit through laser gestures, and how effective magnification is in learning.

Speech alignment. In this experiment, we tested how well our method can identify the time boundaries of a bullet (or part of a bullet) when it is read off the slide. Since our algorithm can only identify bullets whose slide words appear in the transcript, we ran our algorithm on 3 videos in which the lecturer often read off the words of a slide.

We compared the estimated bullet time boundaries to the boundaries created manually, which was done by identifying which bullet words were read from the slide. We use precision and recall to estimate how good the results are. Specifically, precision is $\frac{t_p}{t_p+f_p}$ and recall is $\frac{t_p}{t_p+f_n}$, where t_p, f_p, and f_n are true positives, false positives, and false negatives, respectively. The number of true positives is obtained by intersecting ground truth and estimated time boundaries and the remaining milliseconds of the estimated bullet time boundary are considered false positives. Similarly, false negatives are the number of milliseconds of a non-bullet time interval incorrectly predicted by our algorithm.

Table 1 shows that the average precision and recall are 32% and 37.6%, respectively. While the results may not be good enough to make magnifying semantic units based on speech practical by itself, it provides useful information for more complex methods, such as identifying bullets through topic similarity.

Laser pointer event test. In this experiment, we tested our algorithm's accuracy of identifying semantic units with laser pointers.

We took 9 short 30-second videos in which a presentation slide with bullets and images were shown. In the video, the lecturer used the laser pointer to highlight these semantic units with simple gestures like circling and pointing to the semantic units. Three graduate students watched and created a ground truth sequence of semantic units highlighted for each video. The ground truth from each student was in perfect agreement.

To test the accuracy of our algorithm, we computed the edit distance (using the Unix *diff* program) between the ground truth and the order generated by our algorithm. The error rate is defined as $error = \frac{e}{l}$, where e is the number of edits and l is the length of sequence of semantic units. There were a total of 8 edits out of a sequence of length 59, which gives us a error rate of 13.6%. However, the errors were due to the fact that our laser tracking algorithm lost track of the laser point for a few frames, changing a single gesture into two gestures. Otherwise, our algorithm yields the same semantic unit sequence as the ground truth data.

Effectiveness of Magnification. One problem with viewing lecture videos on a handheld device is that the small screen size makes the contents of the slide difficult to see. We believe that magnification of bullets will alleviate this problem. By enlarging the bullet based on laser-based events, we simultaneously help the viewer find the relevant bullet and also make it easier to read. Our hypothesis is that users who see magnified bullet points will be more likely to remember the content of the bullet point as opposed to users who do not. To test this, we randomly show our participants one of two videos, one with magnification and one without.

To measure the effectiveness of the enlargement, we created a questionnaire by sampling GRE-level nouns. We showed each participant a video of two slides containing definitions of these uncommon nouns (e.g., "gynecocracy"). Each slide had around 10 nouns and was shown for 50 seconds each, making it difficult to memorize all the definitions. To focus the participant's attention to particular nouns, a lecturer would use a laser pointer to highlight them. The font and screen size were chosen so as to simulate a typical smartphone. Students randomly viewed either the original video or a video in which enlargement was performed on the highlighted bullets. Once they finished watching the video, they were automatically redirected to a questionnaire on the definitions of the highlighted nouns.

To measure the correctness of each group, we counted the percentage of total correct answers, In our experiments, there were a total of 40 responses. 23 of those saw the original video and 17 saw the magnified video.

	Precision	Recall
Talk 1	0.290	0.375
Talk 2	0.369	0.409
Talk 3	0.317	0.343
Average	0.320	0.376

Table 1: The precision and recall of bullet time alignment.

	No Magnification	Magnification
Total Correct	74	86
Total Incorrect	87	33
Score	0.460	0.723

Table 2: The table lists the data from the user study. It is partitioned into the group that watched the video with magnification and the group that did not.

From table 2, we see that participants who viewed the magnified video answered more questions correctly and made fewer mistakes. This is reflected by the scores of the users who did and did not watch the magnified video, which are 72.3% and 46.0%, respectively. Assuming that the answers from each group are normally distributed, we use Welch's t-test to show that the scores are statistically significant. The p-value of 0.0092 confirms that participants generally perform much better at remembering the definitions of bullets when they were magnified.

6. CONCLUSION

We demonstrated a method for magnifying relevant semantic units. Our experiments suggest that our method of identifying semantic units by laser pointer is accurate insofar as human judgment is concerned and when the laser points themselves have not been missed. Our method of identifying semantic units based on speech is less accurate, but is a good first step to truly relating spoken words to bullets. Finally, our user study shows that our method of enlarging semantic units can potentially help users remember the lecture contents.

7. REFERENCES

[1] "MIT OpenCourseWare," 2010, `http://ocw.mit.edu/OcwWeb/web/home/home/index.htm`.

[2] "Stanford on iTunes," 2010, `itunes.stanford.edu/`.

[3] "UC Berkeley Extension Online," 2009, `http://learn.berkeley.edu/`.

[4] Quanfu Fan, Kobus Barnard, Arnon Amir, Alon Efrat, and Ming Lin, "Matching slides to presentation videos using sift and scene background matching," in *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, 2006, pp. 239–248.

[5] Quanfu Fan, Kobus Barnard, Arnon Amir, and Alon Efrat, "Accurate alignment of presentation slides with educational video," in *ICME'09: Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, 2009, pp. 1198–1201.

[6] X. Wang and M. Kankanhalli, "Robust alignment of presentation videos with slides," in *PCM '09: Proceedings of the 10th Pacific Rim Conference on Multimedia*, 2009, pp. 311–322.

[7] G. Gigonzac, F. Pitie, and A. Kokaram, "Electronic slide matching and enhancement of a lecture video," in *Visual Media Production, 2007. IETCVMP. 4th European Conference on.* IET, 2008, pp. 1–7.

[8] J. Attewell and C. Savill-Smith, “Learning with mobile devices: research and development,” *mLearn 2003 book of papers*, 2003.

[9] Patricia Thornton and Chris Houser, “Using mobile phones in education,” in *WMTE ’04: Proceedings of the 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE’04)*.

[10] L.E. Dyson, A. Litchfield, E. Lawrence, R. Raban, and P. Leijdekkers, “Advancing the m-learning research agenda for active, experiential learning: Four case studies,” *Australasian Journal of Educational Technology*, vol. 25, no. 2, pp. 250–267, 2009.

[11] “ECMA-376,” 2008, `http://www.ecma-international.org/publications/standards/Ecma-376.htm`.

[12] R. Swaminathan, M. E. Thompson, S. Fong, A. Efrat, A. Amir, and K. Barnard, “Improving and aligning speech with presentation slides,” *Int. Conf. Pattern Recognition (ICPR) 2010.*

[13] A. Winslow, Q. Tung, Q. Fan, J. Torkkola, R. Swaminathan, K. Barnard, A. Amir, A. Efrat, and C. Gniady, “Studying on the move: enriched presentation video for mobile devices,” in *2nd IEEE Workshop on Mobile Video Delivery (MoViD), in conjunction with INFOCOM*, 2009.