EXPANDING THE POINT — AUTOMATIC ENLARGEMENT OF PRESENTATION VIDEO ELEMENTS

Qiyam Tung, Alon Efrat, Kobus Barnard, Ranjini Swaminathan

University of Arizona, Tucson AZ

ABSTRACT

In this paper we present a system that assists users in viewing videos of lectures on small screen devices, such as PDAs. 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/or speech recognition transcript augmented and aligned with WordNet distances. It then magnifies the semantic element using the slide image and the homography between the slide image and the video frame. Our experiment on identifying laser-based events is fairly accurate. Furthermore, a user study suggests that this kind of magnification has potential for improving learning of technical content from video lectures when resolution of the video is limited as is the case when the lecture is being viewed on hand held devices.

Index Terms— lecture, video, magnification, laser, bounding, boxes, 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. In either case, the potential for utilizing this online resource lies in mobile devices such as smart phones and PDAs, which have become powerful enough to watch detailed 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 as the content is crucial to understanding the topic presented. 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 automatically magnifying 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 identify the positioning of each semantic unit on a presentation slide. A key feature is that our method is almost fully oblivious to the format of the presentation slide.
- Identifying the temporal events based on analysis of speech transcript and aligning them to times in the video corresponding to each semantic unit. This event-based segmentation (which acts as a refinement to slide-based segmentation) is of interest because it allows the viewer to browse between these events.
- Similar identification of events based on laser pointer gestures. Our algorithm is quite robust and can correlate a large gamut of gestures to the semantic units they refer to.
- Augmenting the video by backprojecting an enlarged sharp image of this semantic unit, taken from the slide, when and where relevant.

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 have also tested our algorithm that detects when semantic are highlighted by laser points. The results for both experiments are encouraging and 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 to break it into meaningful segments. For example, a lecture video can be indexed by its presentation slides, as shown by Fan *et al.* [4] [5]. Their system, the Semantically Linked Instructional Content project (SLIC), identifies when and where a slide is shown in a video by finding the mapping, a homography, between a presentation slide and a video frame using Scale Invariant Feature Transform (SIFT) points [6]. Using this information, the SLIC system allows the users to browse the lecture by slides. Furthermore, they [7], as well as others ([8] [9]), are often able to find ac-

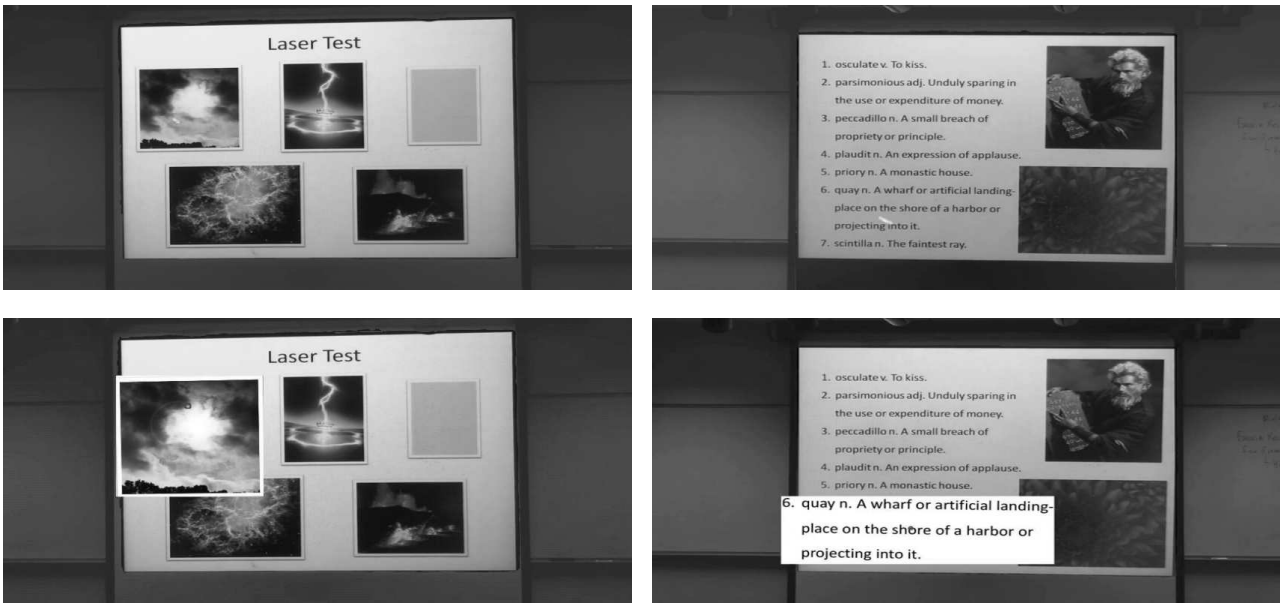


Fig. 1: Two snapshots from videos played with and without magnification. An image, bullet, or word is magnified if triggered by an event, such as a laser gesture.

curate homographies that allow them to project the slide back into the video (for a review on homographies, see Section 4). For our purposes, an additional advantage to having homographies is that we can use this homography to determine where the semantic units are within the video.

There have also been advancements in improving the quality of the video stream aside from backprojection. Cheung *et al.* [10] do this by deblurring the slide in the video, improving the clarity while maintaining the natural look of the video. In [11] and [12], Friedland *et al.* show that their E-Chalk system can improve the readability of chalkboard presentations. The system records the content on a touch-sensitive board and recreates it using vector graphics, creating a clean and sharp video representation of the lecture content. However, even with these enhancements, the lecture material can be hard to see on the small screen of a mobile device.

Mobile devices have long been considered as an important educational tool and much effort and development have been put into mobile learning [13]. Thornton and Houser [14] show that students benefit from using mobile devices as a learning tool. They sent e-mail lessons to Japanese students' phones to promote learning in regular intervals. The response was largely positive. There has also been success in integrating mobile devices into the classroom. In one of the case studies that Dyson *et al.* studied in [15], students participated in a lecture by texting responses to activities using their cell phones. This gave quick feedback on the understanding of the class. These studies suggest a trend towards using mobile devices for educational purposes. Our system will help enhance the understandability of watching a lecture video from a mobile device.

3. IDENTIFYING SEMANTIC UNITS

Our first step is to identify an accurate *bounding box*, which is the set of coordinates for a rectangle, of either a single word, bullet, or image. We have developed a general technique that requires minimal knowledge and assumptions about the format of the presentation files. We have demonstrated it for PowerPoint files, but it can be expanded to suit many other formats, such as KeyNote or OpenOffice presentations.

Microsoft has adopted the Office Open XML (OOXML) format since 2007 [16], which is published as an open standard. Even so, it is difficult to identify the coordinates for each word or image as the coordinates of words are not explicitly specified in the XML format. Instead, we find and modify each semantic unit (words, bullets, and images) so that it has a unique color, effectively identifying their positions. In the next few sections, we discuss in detail how this works.

Finding bullet bounding boxes. We define words to be strings of characters separated by 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 (see figure 2a) in the PowerPoint file. There are two methods of speci-

fying color in a PowerPoint slide file: by using preset color names or by using RGB. We first remove all the preset color attributes and change them to RGB format because it is easier to compare RGB values. Once the format is in RGB, it is possible to identify and give each bullet point a unique color. 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. Thus, we end up with two PowerPoint files whose bullets are uniquely colored. Subsequently, the two sets of slides are exported to images.

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. However, the color of a bullet is only unique among bullets. There could very well be images or background colors that 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 look at our 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 simply 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 archive, 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 archive. 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

Given a video segment corresponding to the use of the slide and its semantic elements, we need to temporally align the elements to when the lecturer discusses those points. We achieve this based on two sources of information: speech and laser pointer use. Having done that, we need to arrange for the magnification of the element in the video frame coordinate system. As noted before, Fan *et al.* [7] found the homographies for slides to frames (and thus when the slide is being shown on the screen). A homography is an operation 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 on a flat surface in the video. The bounding boxes in the slide combined with the homography gives us the information to know where the semantic units are within the video. In the following sections, we will describe how we determine when and how to magnify a semantic unit.

Speech events. When a lecturer speaks, there is a good chance that the words spoken appear in the bullet or are closely related to the words in a bullet. In particular, bullet points tend to contain words that are topically related to what the speaker is saying. When the words in a bullet are read off a slide and are correctly mapped to their corresponding speech words, we can thus obtain times for when a bullet should be magnified.

Swaminathan *et al.* [17] were able to improve speech transcription obtained from automatic speech recognition for lecture videos by noticing that the alignment could be refined by aligning the transcription with words that were in the text of the presentation slides. They observed that presenters often read off their slides and matched the slides' words to the spoken words. This benefits us because when we know which bullet a spoken word belongs to, it also informs us when a bullet is being discussed. We consider a spoken word *aligned* to a bullet if we know what bullet it belongs to.

However, despite this, not all speech words are necessarily aligned to a bullet. This can be caused by two major sources: the speaker rarely says all the words in a bullet verbatim and he or she might utter a few related sentences before and after the bullet. We argue that we can extend the boundaries of speech words related to bullet points by computing their relation to one another. When an unaligned word is between two words that have been aligned to bullets, we call it a *sandwiched* word.

Sandwiched words can be classified into two categories: when the unaligned words are between the same bullet and when they are between two different bullets.

Aligning words between the same bullet. For sandwiched words between the same bullet, we make the simple assumption that these words should belong to the same bullet. Hence, unless we have evidence that other semantic units need to be magnified, we leave the same bullet magnified.

Finding the boundary between two bullets. In the case where the sandwiched words are between two different bullets, we need to find the boundary between the two by comparing the strength of the relationship between the words and the bullets. This tells us when the lecturer switches from talking about one bullet to

the other.

To do this, we use a distance measure based on WordNet [18] to compute the distances between the words in a bullet and a speech word. WordNet is a database for words that are related by *synsets*, which are effectively cognitive synonyms. A lot of work has been done in deriving a general and meaningful way to measure similarity between words. Budanitsky and Hirst compared five different distance measurements of words on several metrics [19]. We chose to use

Lin's word similarity measurement [20] because it generally did well in Budanitsky and Hirst's tests and also because it defines a clear upper bound and lower bound for the similarity between words. Two words are maximally related when their similarity is 1.0. Likewise, when there is no commonality between two words, the similarity is given a measure of 0.0. This makes it easy to gauge how similar two words are. For our experiments, we used the freely available Perl library of WordNet::Similarity, created by Pedersen *et al.* [21]. The library allows the user to compute the distance between two words using a variety of measures, including Lin's measure.

To determine where the boundary between two bullets a and b is, we divide the problem into two subproblems: 1) deciding for each speech word which bullet it belongs to and 2) consequently deciding what the optimal boundary would be.

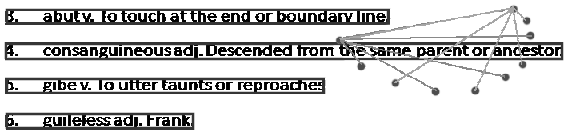
Comparing a speech word to bullets. When determining whether a speech relates to a bullet, we use the following equation to compute the distance.

$$\delta_b(b, s) = \sum_{i=0}^{m} \alpha^i \delta_w(b_i, s) \quad (1)$$

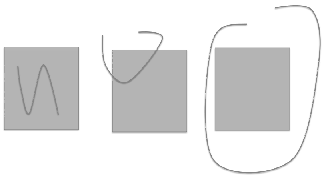
Essentially, the distance between a bullet b and a speech word s, δ_b, is the sum of the individual distances from the m closest bullet words relative to s, denoted in ascending order with b_i. Every bullet word is weighted by α^i, where α is a constant between 0 and 1. For our experiments, we chose α to be 0.5.

Simply summing the distances of all the words will tend to bias the alignment to bullets with more words. As a result, we put the most weight on the bullet word that is closest to the speech word and decrease it for future values through the weight of α^i, which decreases with increasing distance. To determine which bullet it belongs to, we simply take the maximum of the two measures.

Determining the optimal boundary. Once all the words in between the aligned words have been mapped to bullets, we use the following algorithm to determine where the boundary between bullets a and b should be. A boundary is the instance at which the topic switches from bullet a to bullet b. All the words before a boundary j belong to bullet a and all the words that follow it belong to bullet b. Since we already computed which bullet each word belongs to in the previous section, setting a boundary will potentially generate disagreements of which bullet a speech word should be aligned to. We use the disagreements as a cost function, which makes determining the boundary a problem of finding the boundary that minimizes the disagreement. If there is a disagreement for a word s, the cost is simply the similarity measure of s being in the other bullet. Let $bullet(s)$ denote the bullet that s belongs to. Given a sandwiched speech word s before the boundary index j, the cost function $c(a, s)$ of disagreement



(a) The rectangles around the word indicate the bounding boxes and are not part of the original slide image. The points represent a laser dot sequence moving from left to right. Voting on the number of line intersections created by pairs of points is a more robust method of detecting which bullet point to highlight. For the sake of clarity, only a subset of all possible lines are drawn.



(b) The curves represent the path of a laser gesture. Laser gestures can be arbitrary and do not necessarily follow any common geometric shape. For all three cases, our algorithm can still identify which box the laser is highlighting. Here each box represents one of the bounding boxes of the semantic units.

is zero when $a = bullet(s)$. Otherwise, $c(a, s) = \delta_b(b, s)$ when $b = bullet(s)$. The total cost or disagreement is simply the sum of all the disagreements. We then find the boundary index by finding the index that minimizes disagreements out of all possible boundaries.

Laser events. In addition to identifying speech-based events, we also identify events where the laser pointer is used to highlight bullets or images. First, we need to identify the locations of the laser point in every video frame and their corresponding location with respect to the slide's coordinate frame. Our first step is to build on the work of Winslow *et al.*[22]. They find the laser points by identifying the potential bright points on the video frame and fitting these points to curves. Then we apply the homography, which maps the laser pointer to the slide coordinate system, to the laser points so as to compare them to the boxes' coordinates.

Our algorithm uses a voting scheme based on the intersections created by all pairs of laser points (line segments) in a small time interval (see Figure 2a). This provides a notion of the movement of the laser points through the elements. 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. A second advantage is that intersections give a general sense of the area that the laser points cover. Indeed, the area of the arc intersects with the bullet point's box. Even if the curve itself never actually intersects with the box but does circle around it, it will still get votes from the resulting intersections.

For our algorithm, we define a gesture to be a contiguous set of laser points. It is, however, possible that more than one gesture exists in a single set of laser points, so we run our algorithm on a contiguous set of points with a maximum

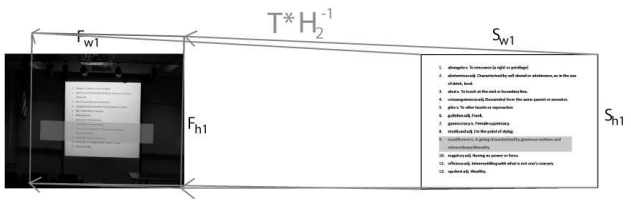


Fig. 2: The highlighted portion shows the location of a bullet point on the slide and its corresponding position on the frame. Without re-centering the bullet, it would shift to the bottom right corner of the frame

length of *INTERVAL*. In our experiments, we chose it to be 1 second.

Magnification of a semantic unit. The homography H we determined as part of the slide-frame alignment process is a mapping from a low-resolution frame to a low-resolution image. To magnify a semantic unit, we need to modify the homography so that it can project and center a semantic unit from a high-resolution image to a high-resolution video frame.

First, we will adjust the scale parameters so that the result is not blurry. This can be obtained with the multiplication of two scale matrices. Assume the homography H maps the slide in an $F_{w1} \times F_{h1}$ frame image to the original $S_{w1} \times S_{h1}$ slide image, which was extracted from the presentation slide. To change this into a homography H_2 that maps a slide from a $F_{w2} \times F_{h2}$ frame image to $S_{w2} \times S_{h2}$ slide image, we use the matrix $H_2 = S_1 \cdot H \cdot mS_2$, where S_2 and S_1 are matrices that scale the slide image and frame to the desired dimensions, respectively. m is the scaling factor for magnification. The inverse, H_2^{-1}, will magnify and place a bullet point on the frame. Next, a translation matrix T is needed to align the centers of the original and magnified bullet (see figure 2). Thus, given a homography H, the homography that magnifies and centers a bullet point from the slide is $T \cdot H_2^{-1}$.

We now have a mechanism for finding events based on both laser points and speech words and can magnify the corresponding semantic unit.

5. EXPERIMENTS

We ran two sets of experiments: one to measure the accuracy of identifying the correct semantic unit through laser gestures and the other to measure the effectiveness of magnification .

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

Setting. We took 9 short videos (approximately 30 seconds each) where 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 (such as circling and pointing to the semantic units). Three graduate students watched the video and created ground truth data, which is the sequence of units highlighted in each video. The ground truth from each student was in perfect agreement.

Results. To test the accuracy of our algorithm, we computed the edit distance (as determined by the Unix program *diff*) between the ground truth sequence and the sequence 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, note that the errors are due to the fact that our laser tracking algorithm loses track of the laser point for a few frames, breaking a continuous gesture into two gestures. Otherwise, our algorithm for identifying the correct semantic unit is exactly the same as the ground truth data.

Usability test. The problem with viewing lecture videos on a handheld device is that regardless of the resolution of the screen, the slide will be difficult to see. We believe that magnification of bullets will alleviate this problem. In this experiment, we randomly show our participants one of two videos, one with magnification and one without. Our hypothesis is that users who see the video with magnified bullet points will be more likely to remember the content of the bullet point as opposed to users who only see the original video.

Setting. To measure the effectiveness of the enlargement, we have 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 contained about ten nouns per slide. Since each slide is shown for a short period (around 50 seconds), this made it difficult to memorize. 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 PDA.

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 a GoogleDoc questionnaire. The participants were given a multiple choice test on the particular definitions of the vocabulary words that were highlighted by the laser pointer in the video. Finally, when they completed the questionnaire, they were directed to a page that explained the purpose of the experiment.

Results. To measure the correctness of each group, we simply 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.

From table 1, 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

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

Table 1: 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.

who did and did not watch the magnified video, which are 72.3% and 46.0%, respectively. Furthermore, assuming that the answers from each group is normally distributed, we can use Welch's t-test to see whether the distribution of scores were statistically significant. The result gave us a p-value of 0.0092, which confirmed that they were indeed. We conclude that participants generally perform much better at remembering the definitions of bullets when they were magnified.

6. CONCLUSION

We have shown that magnification of semantic units are helpful in understanding a lecture video and have developed a method for automatically doing so. We identified semantically meaningful pieces of information from a presentation slide through the use of color matching. We then identify when these pieces of information are referred to. Finally, we magnify them when relevant based on either laser gestures or by relevant speech words. In the future, we hope to be able to enlarge images based on speech by computing the distance between the image and the speech words. This is a challenging problem, but we believe that it will be a very helpful feature.

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, A. Amir, K. Barnard, R. Swaminathan, and A. Efrat, "Temporal modeling of slide change in presentation videos," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 1, pp. I–989–I–992.

[5] 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.

[6] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] 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.

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

[9] 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.

[10] N.M. Cheung, D. Chen, V. Chandrasekhar, S.S. Tsai, G. Takacs, S.A. Halawa, and B. Girod, "Restoration of Out-of-focus Lecture Video by Automatic Slide Matching," 2010.

[11] Gerald Friedland, Ral Rojas, and Ernesto Tapia, "Teaching with an intelligent electronic chalkboard," in *In Proceedings of ACM Multimedia 2004, Workshop on Effective Telepresence*, 2004, pp. 16–23.

[12] Gerald Friedland and Raul Rojas, "Anthropocentric video segmentation for lecture webcasts," *J. Image Video Process.*, vol. 2008, pp. 1–10, 2008.

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

[14] 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)*, Washington, DC, USA, 2004, p. 3, IEEE Computer Society.

[15] 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.

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

[17] Ranjini Swaminathan, Michael E. Thompson, Sandiway Fong, Alon Efrat, Arnon Amir, and Kobus Barnard, "Improving and aligning speech with presentation slides," *Pattern Recognition, International Conference on*, vol. 0, pp. 3280–3283, 2010.

[18] C. Fellbaum, *WordNet: An electronic lexical database*, The MIT press, 1998.

[19] Alexander Budanitsky and Graeme Hirst, "Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures," in *In Workshop on WordNet and other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001.

[20] Dekang Lin, "An information-theoretic definition of similarity," in *In Proceedings of the 15th International Conference on Machine Learning*. 1998, pp. 296–304, Morgan Kaufmann.

[21] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi, "Wordnet::similarity - measuring the relatedness of concepts," 2004, pp. 1024–1025.

[22] Andrew Winslow, Qiyam Tung, Quanfu Fan, Juhani Torkkola, Ranjini Swaminathan, Kobus Barnard, Arnon Amir, Alon Efrat, and Chris Gniady, "Studying on the move: enriched presentation video for mobile devices," in *INFOCOM'09: Proceedings of the 28th IEEE international conference on Computer Communications Workshops*, 2009, pp. 224–229.