

Linking Presentation Documents Using Image Analysis

Berna Erol and Jonathan J. Hull

RICOH Innovations Inc., California Research Center
2882 Sand Hill Road, Menlo Park, CA 94025
+1-650-496-5700
{berna,hull}@rii.ricoh.com

ABSTRACT

Systems for recording presentations are becoming commonly available. Commercial solutions include authoring tools that let users create online presentations by recording audio, video, and presentation slides while a talk is being given. A typical collection of presentation recordings may contain hundreds even thousands of recordings, making it very difficult to retrieve particular presentations and find specific points in a presentation. This paper describes a retrieval technique that utilizes the digital camera pictures taken during presentations. An enabling image matching algorithm is also described. The algorithm utilizes the text and the layout of presentation slides captured in different multimedia streams and documents. Experimental results show that our method yields to a high retrieval accuracy.

1. INTRODUCTION

As digital cameras are becoming widespread, it is becoming more common for attendees to take pictures of interesting slides in a presentation. More often than not, these images become useless because the context and the reason for taking these images are forgotten later. We proposed a retrieval technique that allows users to submit presentation slide images as a query to a collection of presentation recordings to retrieve the audiovisual recording of the presenter talking about those particular slides. This is illustrated in Figure 1. This way, the presentation attendee can refresh his/her memory about some particular presentation slides or share these with others more effectively.

Most digital cameras attach time-stamps to the image they capture. If it is known which presentation recording is related to each captured image, then these time stamps can

be utilized to make the linking between the digital camera image and the captured audiovisual presentation recording. However, in many cases it may not be practical for an attendee to keep track of the presentation sessions that he took each picture in. Especially in the cases when more than one presentation sessions taking place in parallel. In our retrieval application, we overcome this problem by using content-based linking. In the literature, others also proposed content based linking of multimedia streams captured during a presentation. Mukhopadhyay et al. proposed in [1] to match the content of HTML pages that contain presentation slides to the low-resolution video that also includes the presentation slides. Their method is based on first dilating and binarizing the segmented slide images and frames to highlight the text regions, and then using the Hausdorff distance to compute the similarity between the text lines. Their method requires that the slide region be accurately segmented. In [2], Chiu et al. proposed automatically linking multimedia data with a DCT-based image matching of the slide content. They propose to match the contents of scanned handouts, screen capture and presentation video. Their method is mostly suitable for matching high quality and high-resolution slide images and the performance of their method may degrade if the images are low-resolution and are not accurately segmented. Partial occlusion or the presence of blur also degrades its performance. In our proposed method, content-based linking is done by matching the text content and the layout of digital camera image to those of JPEG images captured by a presentation recorder. Our method does not require that the slide region is segmented and it accurately matches slide images even in the presence of motion blur and occlusion.

In the next section we first give a brief overview of our presentation room setup. In Section 3 we describe the proposed content matching algorithm. Experimental results and conclusions are presented in Section 4 and Section 5, respectively.

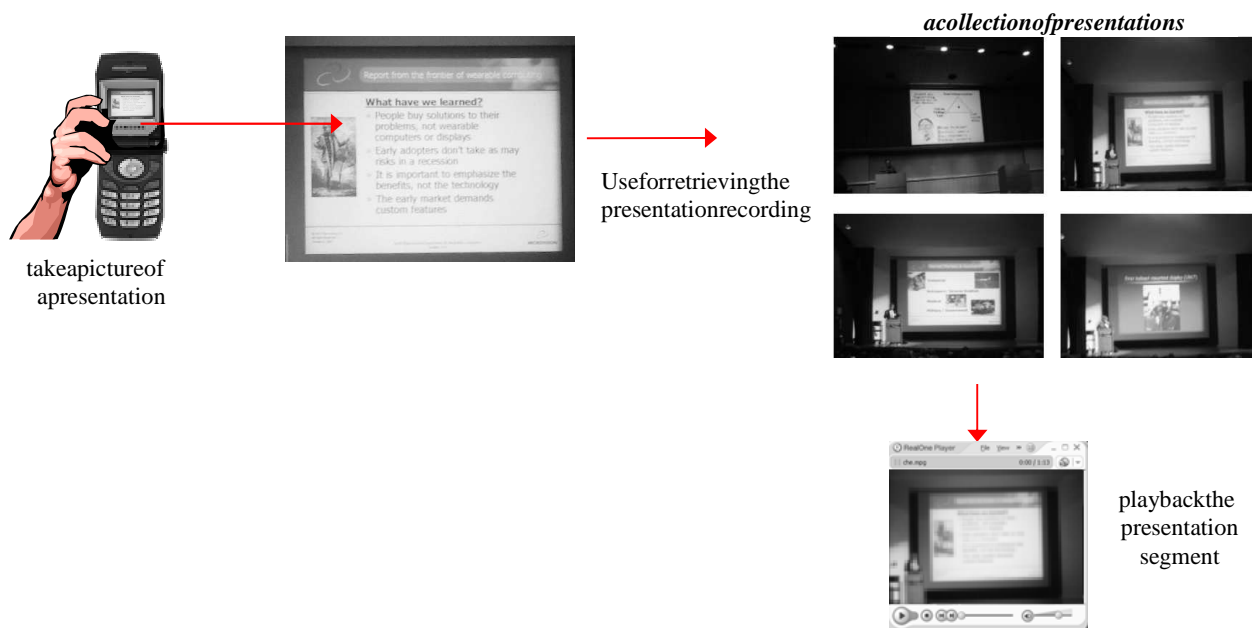


Figure 1. Indexing into presentations with a digital camera picture.

2. MULTIMEDIA CONFERENCE ROOM SETUP

Our presentation room is equipped with an omnidirectional audiovisual meeting recorder and a presentation recorder. The meeting recorder captures 360-degree video at 30 frames per second and view selection is performed based on the sound directions [4]. The presentation recorder automatically captures what is displayed on the presentation screen/projector with the timestamps. To support a wide range of resolutions, the VGA output of the presenter's machine is connected to a scan converter where the VGA output is converted to an NTSC signal, captured by a frame grabber and saved in JPEG format at 640 × 480 resolution. The conversion from digital to analog and analog to digital results in some quality degradation in the image but this allows us to guarantee that the presentation recorder can capture video key frames from any laptop. The output of the meeting recorder and the presentation recorder are synchronized by time-stamps with post-hoc clock-skew correction.

3. IMAGE MATCHING ALGORITHM

We utilize a number of image features for matching the contents of a digital camera image to that of a presentation stream. Slides from the same presentation typically have similar color histograms, dominant colors, etc. Therefore, most color features are not strong discriminators in these

images. On the other hand, slides in a presentation contain different text, combinations of text lines, layouts, images, and graphics. Our experiments showed that the text content and the layout of the text lines of a slide are strong discriminatory features for slide images. Therefore, we utilize the text layout and the text content of slide images to link the presentation stream to digital camera images of slides. A general flow of our algorithm is shown in Figure 2. The next section explains how we extract and match text line layout and text in slide images.

3.1 Text Layout Extraction/Matching

Text regions in images contain strong horizontal and vertical edges. First a Sobel operator is applied to the image to obtain edge magnitudes. After edge magnitudes are computed for an image, an edge is detected only if the edge magnitude is larger than a threshold. Because the background/foreground contrast in a slide image may be obtained by color contrast as well as luminance contrast, edges are detected using R, G, B, and Luminance color components separately. An edge pixel is detected if an edge pixel is identified in any of the color spaces. For each pixel location, a value, E_{xy} , is computed by accumulating the number of edge pixels in a neighborhood window $K \times L$. The pixels that have E_{xy} values that are larger than an adaptive threshold are marked as pixels that belong to a text region. Connected component analysis and aspect ratio information is then used to identify the text regions. Figure 3 shows the identified text regions from a digital camera

slide image. The figure also shows that the edge detection results are significantly better when all the color components are used for edge detection versus using just the luminance component.

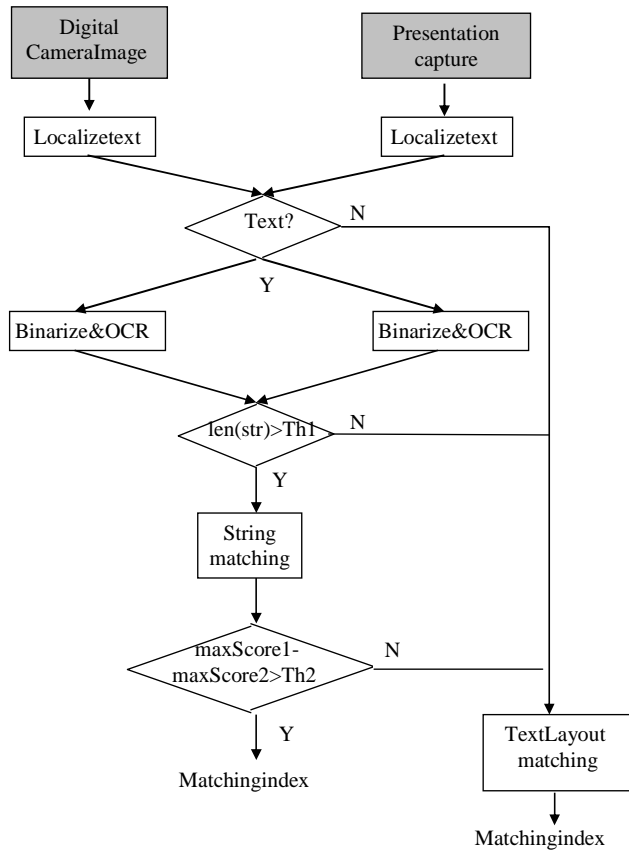


Figure 2. Slide matching with text and layout

After text regions are identified, text line profile is extracted by computing, for each horizontal line in the image, the maximum run of pixels that belong to a text region. In order to compute the similarity distance between the text line profiles of two slide images, first the text line profiles are normalized in both horizontal and vertical directions using the highest peak. Note that this highest peak corresponds to the longest text line in the slide image. The distance between two line profiles are found by aligning the profiles with their peak value and computing the sum of absolute differences.

3.2 String Extraction/Matching

After text regions are identified, a commercial OCR package [5][6] is used to extract text from slide images. OCR packages usually correct for small skew and rotations, therefore we do not address these problems here.

After the images are OCR'd, string matching is performed on the text output to find a similarity score between two slide images. OCR results from slide images captured by different devices can vary widely. For example, the text output extracted from a digital camera image is generally less accurate than that obtained by OCR'ing the screen projection output for the same slide. In most applications one of the capture sources is likely to be more reliable than the other and the OCR results obtained from one of the sources could be close to the ground truth. We take this into consideration when performing string matching and define the string that is obtained from the more reliable source as the ground truth string. The characters obtained for each slide are first concatenated. Then the similarity score between two strings are computed as $d_s = (\ell_g - d_e) / \ell_g$, where ℓ_g is the length of ground truth string and d_e is the edit distance between two strings. Edit distance is computed by counting the number of insertions and deletions required for matching. The punctuations and the extra characters in the string that is extracted from the less reliable source are ignored.

4. EXPERIMENTAL RESULTS

In this section, we present experimental results on the retrieval performance obtained by our proposed algorithm. In the first experimental setup, image matching is performed on the images obtained from the following two sources:

- Presentation slides in PowerPoint that are saved as a series of JPEG images.
- Digital camera: 2-4 MPixel images mostly containing the slide region, with rotation less than ± 5 degrees.

Here, we performed experiments on a presentation recording database containing 41 presentations with 1160 slides. We collected 109 digital camera images to be used for queries in our experiments during 7 different presentations given in our lab. Text extraction is performed on the PowerPoint slides using the Microsoft APIs at the time they are inserted into the database. All the digital camera images contain at least one line of text. These images vary in terms of room lighting, use of flash, distance to the projector, motion blur, occlusion, etc. The process of matching one digital image to a database of 1160 images was performed in under 3 seconds on a 1 GHz PC. Note that segmentation of the slide region from the digital camera image is not required for the matching techniques employed here.

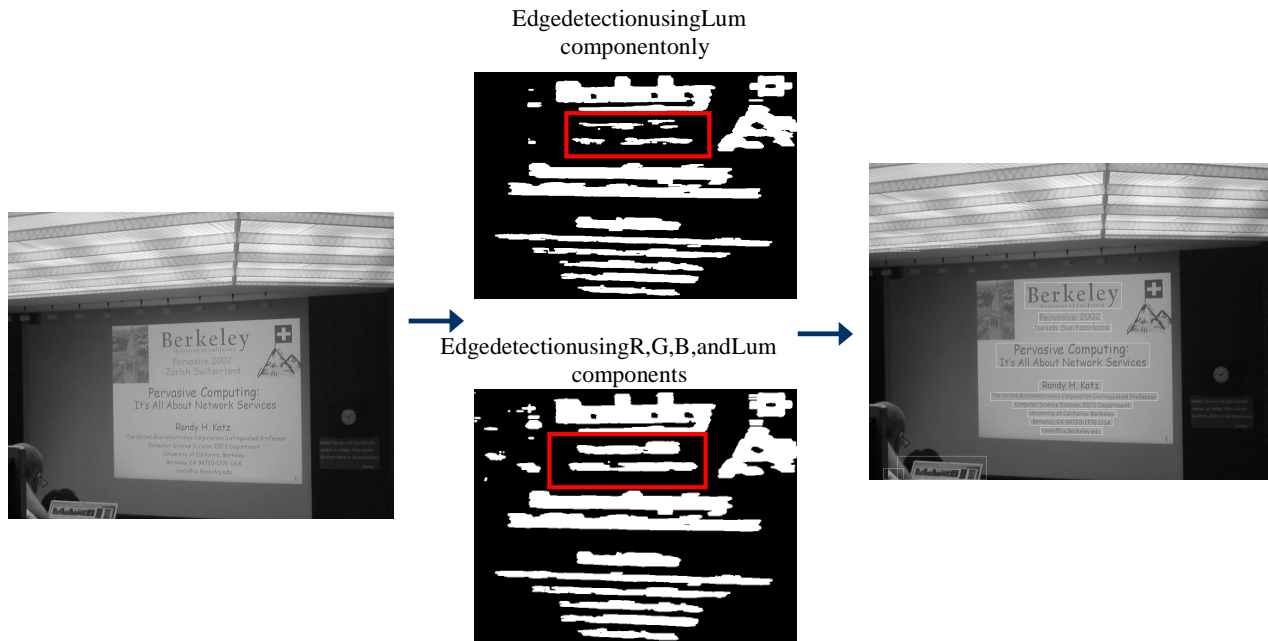


Figure3. Text localization detection in a digital camera image.

We also ran another set of experiments where the slide images captured by the *Presentation Recorder* are retrieved by using the digital camera images. In this case, the database contained 41 presentations with 4814 captured presentation recorder images (i.e. screen capture). Most of these images were presentation slides. Some of these images have the same slide content because a presenter may visit a slide more than once. For each query, the 10 slide images with the highest similarity score are retrieved.

The retrieval results obtained by querying the presentation slides with digital camera images are presented in Table 1. The retrieval rate is defined as the ratio of correctly retrieved relevant images to the number of relevant images in the database. The ground truth images are the presentation slides that match the digital camera images in terms of their slide content. As can be seen from Table 1, the average retrieval rate (recall) obtained by querying the PowerPoint slide database is 95%. The 3rd presentation, where the presentation slides generally contained very few lines of text, yielded the lowest retrieval rate. The average retrieval rate obtained using the Presentation Recorder database is 87%. Some of the presentations contained slides with small fonts (e.g., 2nd and 4th presentations). In these cases, the OCR on the slides captured by the Presentation Recorder yielded low accuracy, resulting in a lower retrieval performance. However, the retrieval performance using the PowerPoint slides was not affected

by this since in this case we extract the text directly from the PowerPoint file instead of using OCR.

Table 1. Retrieval results when digital camera images are used to query presentations.

Pres. no	Number of digital camera images	Ave Retrieval Rate (PowerPoint slides)	Ave Retrieval Rate (Presentation Recorder slides)
1	19	0.94	0.91
2	7	1.00	0.75
3	11	0.91	0.91
4	9	1.00	0.85
5	20	0.95	0.86
6	22	0.96	0.92
7	21	0.95	0.84
ALL	109	0.95	0.87

One should keep in mind that the overall retrieval accuracy of this method greatly depends on the quality (resolution, sharpness, etc.) of the digital camera image as much as the amount of text content in the captured presentation slide. Note that it is possible to improve the retrieval accuracy of slide image retrieval by incorporating information from the digital camera image timestamps.

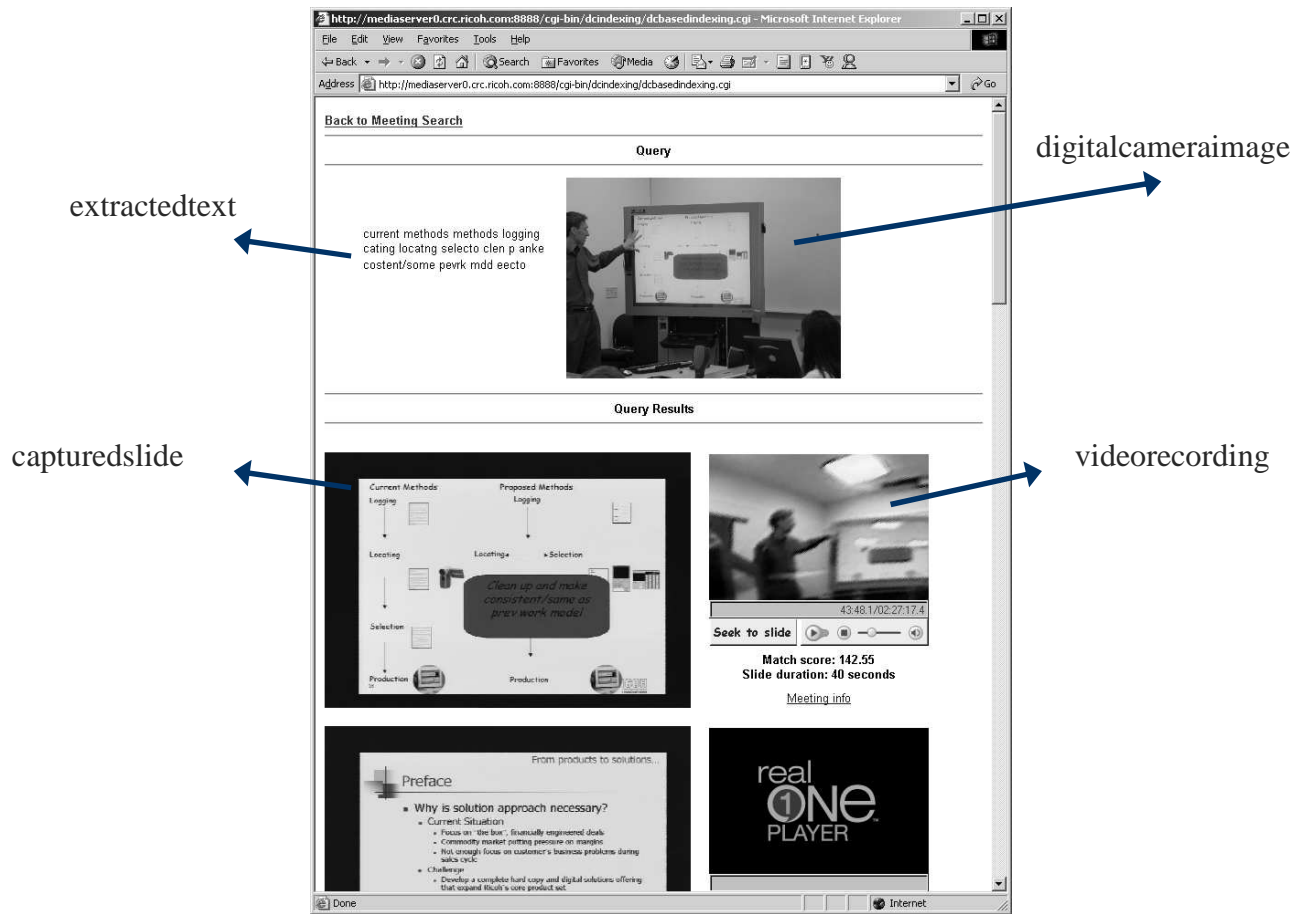


Figure4.Digitalcamera-basedretrievalinterface forpresentationrecordings.

The retrieval results are presented in the interface shown in Figure 4. For a given query image, the matched presentation slides and their corresponding audiovisual recordings are displayed in an order based on the similarity score. User then playback the video recording of the presenter talking about the particular slide that has been captured with the digital camera image.

5. CONCLUSIONS

In this paper, we presented a novel retrieval application and a novel method for linking multimedia streams captured during a presentation. Linking presentation streams and documents through slide image matching opens up possibilities for new ways to index, retrieve, and access meeting/ presentation/ lecture content and, consequently, for many interesting and useful applications.

6. REFERENCES

- [1] Mukhopadhyay, S., and Smith, B., "Passive capture and structuring of lectures", ACM Multimedia pp. 477-487, 1999.
- [2] Chiu, P., Foote, J., Girgensohn, A., and Boreczky, J., "Automatically Linking Multimedia Meeting Documents by Image Matching" proceedings of Hypertext'00, ACM Press, pp. 244-245, 2000.
- [3] Chiu, P., Kapuskar, A., Wilcox, L., "Meeting Capture in a Media Enriched Conference Room," 2nd International Workshop on Cooperative Buildings, pp. 79-88, 1999.
- [4] Lee, D.S, Erol, B., Graham, J., Hull, J.J., and Murata, N., "Portable Meeting Recorder", ACM Multimedia Conference, pp. 493-502, 2002.
- [5] Scansoft Capture Development System, www.scansoft.com
- [6] Transym OCR Engine, http://www.transym.com/