Searching for classes of visual content in electronic lectures

Peter Ziewer
Technische Universität München, Germany
ziewer@in.tum.de

Martin Ebner, Christian Safran,
Wolfgang Slany
Technische Universität Graz, Austria
{martin.ebner, csafran,
wolfgang.slany}@tugraz.at

Abstract

Lecture recording allows the creation of electronic learning material for local and distance education. Flexible screen recording techniques can capture virtually any material displayed during a presentation. Unlike other screen recorders, our TeleTeachingTool already offers slide-based navigation and full text search as meaningful retrieval features. Our paper describes how to locate remembered content, such as images, diagrams, or graphs, and furthermore how to find similar passages and related content in large databases of recorded lectures by classifying navigational indices. We introduce a classification scheme using characteristic parameters based on the analysis of color histograms of the recorded screens.

1. Introduction

Lecture Recording is conserving presentations for later playback and provides additional learning material that can be created rather cheaply and quickly and therefore is called lightweight content creation [4]. In order to provide full-fledged electronic lectures as meaningful learning material, the provision of useful navigational and retrieval features is essential [4].

Presentation recorders using symbolic representation, e.g., Authoring on the Fly (AOF) (University of Freiburg) or Lecturnity (imc AG), store structured documents and thus achieve slide-based navigation as well as full text search. The more flexible approach of the screen recording technique, which is used by Camtasia (TechSmith Corporation) and TeleTeachingTool (TU München, http://ttt.in.tum.de/), digitally grabs the content of the presentation machine's desktop on a pixel basis and therefore can capture not only slide presentations but rather any application (e.g., presentation software, editors, and browsers) including pointer movements, animations and annotations (e.g., notes and sketches drawn with an electronic pen). This flexible technique is often criticized for omitting document structures and textual content [6] that provide useful and, as postulated by [4], necessary advanced playback features. However, we have shown in [11] how to regain document structure and achieve slide-based navigation and full text search by automated analysis of the pixel-based recordings. We thus have diminished the major drawbacks of screen recorders when compared to symbolic recorders.

Up to now students could either search by specifying a textual search pattern or by surveying a large number of thumbnails only. With the work described in our present paper we introduce visual retrievability, i.e., the localization of classes of graphical content such as images, diagrams, or graphs within large databases of recorded lectures. Analogous to our previous work this additional feature should be integrable without manual post-processing and all required data structures should be automatically derived directly from the pixel-based recordings.

At first we motivate our approach according to the didactical background. Afterwards we shortly describe TeleTeachingTool that is the basis for our research. Chapter 4 explains the basic idea of content prediction via color histograms and furthermore states characteristic parameters and filtering techniques to classify text slides, photos, diagrams, graphs, or other content. We then discuss the application from the point of view of students and propose ideas for future work.

2. Didactical background

"When is an illustration worth ten thousand words?" – This famous question was answered in the early 1980's by a number of psychologists [7]. Baddley [1] described in his working memory model of our brain that we remember through, on the one hand, a phonological loop for speech-based memories and acoustic information, and on the other hand a

visuo-spatial sketchpad for visual information. Considering these theories Paivio [9] founded his dual coding theory and proved it with several experiments. He showed that words can be learned much easier if a visual association has been given to the testing person [2]. Thus pictures will aid the learning process and will be remembered better than text alone. Pictures can thus be interpreted as abstract anchors that help to structure the learning material.

Firstly Mayer [8] verified in a number of studies that in terms of using multimedia productions for learning, combining text or speech data with visual information is optimal. Secondly didactically prepared pictures help learners to remember essential parts of lectures [5]. Finding pictures without any knowledge about exact details or descriptions thus can positively influence the learning process.

Since it is hard to precisely specify a suitable visual search pattern we suggest that users can react to answers from the system and adapt their query in order to find better results, thereby further enhancing the learning experience. Our system is geared at allowing this kind of interaction.

3. TeleTeachingTool

Since the results presented in this paper are derived from, although not limited to, our research with TeleTeachingTool (TTT), we will give a short overview about the system. TTT is a freely available, cross-platform lecture recording and broadcasting environment that offers flexible screen recording enhanced with audio and video. It supports various operating systems and allows the parallel use of arbitrary applications, including the teacher's choice of presentation software, animations and browsers. TTT can be seamlessly integrated into an existing teaching environment without influencing the teacher [10].

Unlike other screen recorders, TTT offers slide-based navigation and full text search. The required document structure is automatically derived from pixel-based recordings without access to the original documents and with no need for manual post-processing [11]. The core idea for regaining structure is to automatically create indices as access points to certain positions within the timeline of a recording that allows to subdivide and thereby to structure it. During playback, slide indices are presented as a thumbnail overview. Clicking on a preview image causes an instantaneous replay starting from the corresponding slide. Index screenshots are used to automatically generate lecture scripts in the format of html or pdf documents that include annotations (e.g., notes and sketches drawn with an electronic pen) made during the lectures, thus making them available to students. We furthermore extract a textual search base by applying optical character recognition to the index screenshots and thereby can offer keyword based search presenting slide indices as search results. TTT thus already offers the most important navigational and retrieval features [4][6] to provide full-fledged electronic-lectures. Figure 1 shows TTT displaying an annotated slide and the thumbnails for slide-based navigation.

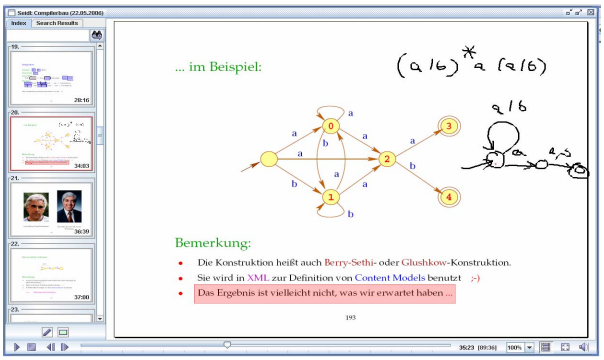

Figure 1: TeleTeachingTool

4. Content prediction by color histograms

In order to extend the retrievability of electronic lectures we classify the navigational indices according to the presented content. Since we do not want to apply complex image analysis and comparison algorithms or time-consuming manual post-processing, we must find simple but characteristic parameters that easily can be derived from the electronic lectures in an automated fashion. For each index we have a screenshot of the recorded desktop. Analyzing these index screenshots reveals that most pixels of simple text slides are colored in the background color, typically more than 90%. On the other hand we have more complex slides or recorded desktop applications that have lower values. Furthermore, the number of used colors differs for highly colored photos, diagrams, and tables.

Since equipollent pixel values are analyzed, it is not explicitly known which color is the background color. At least for ordinary slides the background color is the most frequently used color (abbr. m.f.u.c.), which can be determined by computing the color histogram for each index screenshot. Such a histogram can easily be calculated once and then stored for later retrieval purposes. Note that it is not necessary to store millions of pixel values per screenshot because typically only a few dozen different colors will be used per screenshot and furthermore our algorithm does not require the full histogram as it takes into account only the most

frequently used colors. For this research we have computed the color histograms of over 15,000 index screenshots of 361 recorded lectures. As these lectures were given by six different teachers lecturing various courses throughout several years, the applied database reflects different presentation styles, especially as the teachers not only used different operating systems and presentation software but also additional applications such as simulators, programming editors, or browsers.

All of the given thresholds have been ascertained by iterative approximation, i.e. the values have been increased/decreased until a meaningful classification could be achieved. In order to find suitable thresholds it was often useful to survey the rejected indices that where close to the threshold. Thereby it was possible to find thresholds that do not reject too many matching results. Typically the classification has an error rate of less than 5%.

4.1. Border elimination

While analyzing our index screenshots to determine characteristic values for content prediction we found that for some presentations the characteristic values did not perfectly match but rather were slightly shifted. Surveying the screenshots revealed that headers and footers of slides introduced this undesirable effect. Large headers that use separate header background colors shift the m.f.u.c. by about 5-10% depending on the area covered by the header. Analogous effects arise from a desktop's taskbar, application menus, or the slider bars on either side of a window. Since almost all of the irritating elements are placed on the outer regions of the recorded desktop they can easily be eliminated by computing the color histograms of the inner region only. Therefore our color histogram computation ignores the upper 140 lines of pixels at the top of each screenshot and also cuts the area by 80 pixels on each of the remaining sides (at a resolution of 1024x768 pixels). These values were determined as appropriate for most of the given screenshots. Since the shift of values of the resulting histograms is negligible, it is unproblematic to cut the outer borders of slides that do not have complex headers or footers or to erase parts of slides whenever their borders are actually smaller than the truncated borders.

4.2. Text versus complex slides

Most pixels of plain text slides are colored in the background color. Hence, all color histograms that contain a very high value for the m.f.u.c. are likely to correspond to text slides. Smaller values indicate less background and thus more space for "content" such as text, diagrams, and images. The background color of slides that consist of a few words of text only covers more than 95% of all pixels. With more text this value decreases to 80-90%. If slides are well packed with text that is written in a huge bold typeface, the area colored in the background color can be as low as 75%.

However, not all histograms that show 75-100% background color correspond to text slides. Tables or diagrams that use the same background as the slide also result in similar values. An additional parameter that indicates which slide may be classified as a text slide is the length of the presented text, which is available as search base for the full text search. Obviously the more characters were recognized, the more likely a text slide is shown. Again this is not a sufficient condition because some text slides may show only a headline or a few keywords, and others may show diagrams with textual explanations, both resulting in a similar small amount of characters. Although the retrieval of text slides via content prediction is unlikely, a classification as text slide is meaningful because it can be used as an exclusion criterion for other requests to eliminate the number of results that are presented to the students. Only screenshots with a very large portion of background color and a large number of recognized characters are therefore classified as plain text slides.

4.3. Photo detection

One often remembered element of presentations is a photo. Therefore a good prediction of which screenshot may show a photo is desirable. The color histogram of a photo typically consists of many different colors, all of which cover only very few pixels. If a photo is presented as part of a slide, most of the pixels may still be colored in the background color of the slide. At first we filter all screenshots by the m.f.u.c. and eliminate those indices that are represented by screenshots where more than 75% of all pixels of the screenshot (reduced by the border) are colored in the same color. The threshold of 75% has empirically been shown to work well and implicitly defines an area of at least 25% to be filled with "content" such as text, diagrams, or images. Very small images or images displayed at the border of the slide may not be found but typically will not be very important as otherwise the teacher would have made them bigger and more centered.

Since photos rarely consist of many equally colored pixels, we further filter all screenshots with histograms that contain a second and third m.f.u.c. that exceeds 15% and 7% respectively. These filters eliminate slides that contain diagrams, tables, black-and-white scans, or web pages but almost no photos of notable size. By applying those filters to our database of over 15,000

screenshots, the number of results is reduced to about 600 indices that very likely refer to slides with photos.

Although the described filtering already leads to very good results, it can be further improved by taking into account not only the frequencies of the histogram per color but the sum of the most frequently used colors. If the amount of pixels that are colored in the five most frequently used colors exceeds 90% of all pixels, the corresponding screenshots typically do not display a photo and therefore can also be ignored. Lowering the filter to accept sums up to 80% eliminates screenshots that contain diagrams and some rather small or color reduced images only. Whenever the sum is lower than 50% one can expect with high probability that it contains a photo. Values between 50 and 80% are caused by web pages and complex text slides with images, but also by slides with photos and therefore should not be filtered by default. However, lowering the threshold can lead to better results for some courses. We thus propose user adjustable thresholds. Since the users should not be bothered with technical details, we prefer offering the possibility to stepwise reduce or increase the number of results and implicitly adjust the filtering thresholds accordingly.

Special cases are fullscreen images including screenshots, fullscreen videos, and live demos in programs. Recall that we eliminate the borders of our screenshots to reduce misleading noise. Obviously retrieval of fullscreen images requires the unrestricted histograms of the complete screenshot. Fullscreen photos are those images without any background, i.e., no color covers more than 10% of the overall area and at most one color exceeds 5%.

4.4. Diagrams, graphs and tables

Other classes of objects that are often presented during a lecture are diagrams, graphs, and tables. Such artificially created images, i.e., images created via computer, typically consist of fewer colors than real world photos and thus lead to different characteristics of the corresponding color histograms. The best results are achieved by considering the overall area covered by the most frequently used colors. A characteristic value is the sum of the five m.f.u. colors. Applying a sum filtering with a range of 85-100% eliminates most photos. Additionally we eliminate all indices that are classified as simple text slides according to chapter 4.2.

Surveying our database of screenshots reveals that good results are achieved whenever the coverage of the m.f.u.c. is within a range of 60-85%, the sum of the five top values of the histograms exceeds 85% and slides with more the 500 recognized characters are excluded. If the m.f.u.c. exceeds 85%, then the number of text slides retrieved increases. Therefore a lower threshold for the text length filter of 250 characters should be applied for high background coverage, which is unproblematic because slides with much background and many words hardly offer any space for additional graphs. Note that specifying lower values for the text length filter may eliminate not only plain text slides.

Screenshots with less than 60% background color often present some desktop applications that cannot be distinguished from diagrams or graphs by this approach. Nevertheless such screenshots are special compared to simple text slides and therefore can be visually remembered as well.

A stepwise adjustment of the quality and number of search results can be implemented analogously to our suggestions for the photo search. The range of the m.f.u.c. can be reduced from 40-100% to 60-85% and the threshold of the text length filter can be lowered. Furthermore we can increase the minimum area covered by the second m.f.u.c. in order to list only those elements with a distinct background color such as the graph shown in Figure 3.

4.5 Similarity search

Lecturers can present very diverse content, especially as the flexible technique of screen recording allows arbitrary applications to be recorded. It thus is hard to classify certain groups of graphical content and even harder to determine characteristic parameters to distinguish those groups. However, such content, for instance certain types of graphs or diagrams, often show similarities that are recognizable to the human brain at a glance. Surveying several hundred screenshots or thumbnails of a dozen or more lectures is not acceptable for human beings in realistic settings. Therefore we introduce an approach to solve this problem by implementing a similarity search based upon color histograms.

For a given index screenshot the similarity search should return screenshots and thereby indices with similar histograms, which should present related content with high probability. At first we must determine suitable filtering thresholds. The more results we receive, the more likely we receive dissimilar results. However, a larger range is useful for certain content, for instance if a lecturer uses an editor or a terminal application such as the windows command line interface shown in Figure 2, which is not uncommon in computer science lectures. Obviously the size of the terminal window has a large impact on the area that is covered by the terminal's background color, but a resized terminal nevertheless represents a very similar content. On the other hand we have slides with graphs as presented in Figure 3.

Whenever using the same background as the slide, such graphs often cover no more than 5% of all pixels regardless of the number of knots the graph has. We therefore set the ranges for the similarity search in relation to the histogram of the input screenshot. We found ranges of +/-15%, +/-10% and +/-2% applied to the three most frequently used colors suitable whenever the origin histogram does not show a very high percentage of one color, e.g., when showing a terminal as described above. Whenever the coverage approaches the upper or lower bounds of 100% or 0% respectively, the applied range is set to the minimum distance to the closest bound. For example a histogram of 95%, 3%, and 1% results in a search range of 90-100%, 0-6%, and 0-2% respectively.

Figure 2: Partly visible terminal application and corresponding histograms

The search by percentages alone does not lead to very meaningful results. Applying the search range for the second of the previously given examples results in not only retrieving the searched graphs but also almost all simple text slides. Therefore we additionally compare the color values. Since it is likely that colors may switch their position in the compared histograms, permutations must be tolerated. For instance if the terminal of Figure 2 is resized, black and white may shift positions in the histogram. However, it is not tolerable if a background color that covered more than 80% in the input histograms shifts to another position. A tolerable permutation must thus respect the values of the input histogram.

Due to the specialty of the compared content, this simple similarity search delivers good results whenever applied to screenshots of lectures. The reason is that teachers typically use a certain kind of graph, diagram, or application with characteristic coloring to explain specific topics. The similarity search can, for instance, be used to find similar explanations within the same lecture series, e.g., if examples are consecutively extended throughout a lecture. Searching similar results in lectures of the same course that were recorded in earlier years can be useful whenever a student has not understood a certain explanation of a graph or other topic and he/she wants to check if earlier recordings contain some additional or better explanations.

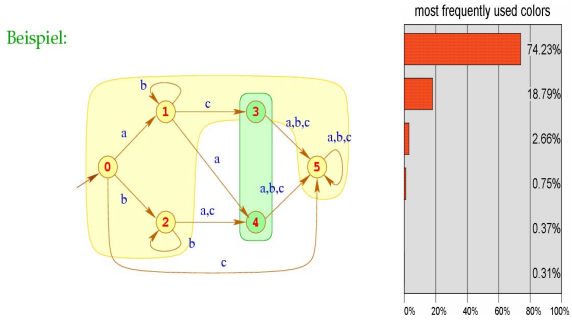

Figure 3: Slide with graph and histogram

5. Application for students

The development of an application for students implementing the previously introduced technologies should focus on potential use cases during the learning process. Based upon such use cases, proposals for possible features can be given.

A first application is the search for pictures in the lecture recordings for the reason described in Section 2. Pictures can serve as visual anchors for learning material and are remembered more easily than written words [2]. A second possible use case is the search for graphs in the course material. While pictures rather help to remember certain parts of a lecture, the search for graphs allows the retrieval of contents of a lecture. A third possibility is to search for fullscreen graphics. These include screenshots, fullscreen images and fullscreen videos, as well as live demos in programs. In order to find a summary of picture, fullscreen, and graph slides, together with other slides that were not recognized by above filters, it is possible to search for “non-text” slides. To that end the filter for pure textual slides described in Section 4.2 is inversed.

All the above mentioned applications are available through a search interface, either as a search on the whole set of available course material or just within a

series of lectures. A separate application is the search for similar slides. For each slide found by one of the filters (or taken from a listing of all slides or the slides of a series of lectures), a “show similar” link is provided. It returns a list of slides with similar histograms and can be used to find slides showing the same application or related graphs. The results can be adapted by a “more exact – less exact” slider that changes the threshold for the allowed rankings of the individual colors in the histogram.

We have not done an evaluation of our approach yet, because we have no data about long term usage of the system. However, based on the reasons stated in Section 2 and the low error rate of the presented results, we assume that finding certain graphical content within electronic lectures can be improved with our approach whenever textual search fails. Especially the similarity search mostly presents good results and thus offers access to related content very easily.

6. Conclusion and Future Work

In this paper we have described how color histograms can be used to locate visual content within large databases of electronic lectures. We have determined characteristic parameters for different classes of content. Finding photos is a relative easy task. Due to the vast variety of different graphs, diagrams and tables used within presentations, finding such objects is more difficult. Nevertheless the number of screenshots that must be surveyed by the user can be noticeably reduced with our approach. Furthermore, we have introduced a similarity search to locate related content for a given input index. Although the presented results are not perfect, visual retrieval offers students additional means to locate certain graphical content within electronic lectures whenever textual search fails. Our interactive approach offers students the possibility to adjust their query according to the presented results if necessary. Since the required data structures are derived directly from the pixel-based recordings without manual post-processing, the suggested retrieval methods can be applied to any electronic lectures, even pixel-based recordings.

One drawback of the current approach is that slides with multiple background colors, background images, or color gradients in the background falsify the results. The first group is interpreted as graphs, while the second and third are identified as pictures. In order to improve the recognition rate, the detection and removal of background pixels for the calculation of the color histogram will be added in a next step. The necessary techniques to identify the image background are used in video processing, i.e., in video surveillance systems. Possible methods to detect the image background pixels that can also be used in our approach are for example described by [3]. However, the problem is much simpler in our case as the picture quality does not change as compared to video input and only few pictures have to be processed.

References

[1] A. Baddely, "Working memory. Life Sciences", 1998, p. 167–173.

[2] G.H. Bower, M.C. Clark, A.M. Lesgold, and D. Winzen, "Hierarchical retrieval schemes in recall of categorized word lists", Journal for Verbal Learning and Verbal Behavior, 1969, 8:323–343.

[3] P. Gil-Jiménez, S. Maldonado-Bascón, R. Gil-Pita1, and H. Gómez-Moreno, "Background Pixel Classification for Motion Detection in Video Image Sequences", in Lecture Notes in Computer Science, Volume 2686/2003, p. 1041ff.

[4] P.-Th. Kandzia, G. Kraus, and Th. Ottmann, "Der Universitäre Lehrverbund Informatik - eine Bilanz", *Softwaretechnik-Trends 24:1*, Gesellschaft für Informatik, Feb. 2004, p. 54–61.

[5] A. Krapp and B. Weidenmann, "Pädagoschische Psychologie", Urban & Schwarzenberg, Psychologie Verlags Union, Weinheim, Beltz, 4 edition, 2001.

[6] T. Lauer and Th. Ottmann, "Means and Methods in Automatic Courseware Production: Experience and Technical Challenges", World Conference on E-Learning (E-Learn'02), Montreal, Canada, Oct. 2002.

[7] J. Levin and R.E. Mayer, "Understanding illustrations in text", In: Britton, B., Woodward, A., and Brinkley, M., editors, Learning from Textbooks, 1993, p. 95–113, Erlbaum, New York.

[8] R.E. Mayer, "Illustrations that instruct", In: R. Glaser, editor, Advances in Instructional Psychology, volume 4, 1992, p. 253–284, Erlbaum, New York.

[9] A. Paivio, "Abstractness, Imagery, and Meaningfulness in Paired- Associate Learning", Journal of Verbal Learning and Verbal Behavior, 1965, 4:32–38.

[10] P. Ziewer and H. Seidl, "Transparent Teleteaching", 19th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCILITE), Auckland, New Zealand, Dec. 2002, (2), p. 749–758.

[11] P. Ziewer, "Navigational Indices and Full Text Search by Automated Analyses of Screen Recorded Data", World Conference on E-Learning (E-Learn'04), Washington, D.C., Dec. 2004.