Lecture Video Segmentation by Automatically Analyzing the Synchronized Slides

Xiaoyin Che
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3,
14482 Potsdam, Germany
xiaoyin.che@hpi.uni-potsdam.de

Haojin Yang
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3,
14482 Potsdam, Germany
haojin.yang@hpi.uni-potsdam.de

Christoph Meinel
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3,
14482 Potsdam, Germany
christoph.meinel@hpi.uni-potsdam.de

ABSTRACT

In this paper we propose a solution which segments lecture video by analyzing its supplementary synchronized slides. The slides content derives automatically from OCR (*Optical Character Recognition*) with an approximate accuracy of 90%. Then we partition the slides into different subtopics by examining their logical relevance. Since the slides are synchronized with the video stream, the subtopics of the slides indicate exactly the segments of the video. Our evaluation reveals that the average length of segments for each lecture is ranged from 5 to 15 minutes, and 45% segments achieved from test datasets are logically reasonable.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods, Indexing methods*; K.3.1 [**Computers and Education**]: Computer Uses in Education—*Distance learning*

Keywords

Video Segmentation, Slides Content Analysis, OCR

1. INTRODUCTION

With the prosperity of online video lecture, the knowledge can break through the wall of campus and benefit the people around the world. However, it is not comfortable for people to stare at a computer screen for a long lecture, especially when they just have interest in some parts of the lecture rather than the whole. As a result, to segment the lecture video becomes nessecary.

Manual segmentation has no doubt the best result, but it is also a huge consumption for the time or/and money. And due to the specialty of lecture videos (*commonly no scene change*), the segmentation methods for natural videos[1, 4] are no longer proper. Instead, many research for lecture video segmentation base on blackboard writing[2, 6], speech recognition[5, 7] or additional lecture transcript[3]. But unfortunately, people use blackboard less recently while speech, regardless of the accuracy problem, is always too flexible to be concluded as certain topics. And additional transcript exists only in some special cases. All those facts drive us to look for some new ideas of segmenting lecture videos.

In recent years, many lecturers use slides when giving lectures, instead of blackboard. Meanwhile many E-Learning systems also include the slides as well as the lecture videos. Slides are generally the outline of the lecture or presentation. By analyzing slides content properly, the whole lecture or presentation can be logically divided into several subtopics, in other words, segments. If the slides are synchronized with the relevant lecture video, no matter recorded in a second video stream(*tele-TASK*[1] *e.g.*) or just stored as a sequence of time-related pictures (*VideoLectures.NET*[2] *e.g.*), the slides segments can be perfectly treated as the video segments.

Based on this idea, we propose a new segmentation method for lecture videos by analyzing the synchronized slides content from OCR results. OCR technology enables us to extract textual data from pictures or videos, and for the slides of lecture video, the accuracy is fairly high that over 92% characters and 85% words can be extracted correctly and saved in text-lines, which contain both text data and location info[8]. Then we can reconstruct the content structure of each slide. After that we will first figure out whether there are special slides can be recognized as subtopics border, which is the foundation of global segmentation. Furthermore, we also search for index-pages or generate virtual index-pages by exploring the shared words between titles of continuous slides, which enables partial segementation. Finally a default time segmentation process is included.

The rest of this paper will be organized as follow: Section 2 illustrates our detailed solution while Section 3 and Section 4 show the evaluation results and conclusion respectively.

2. DETAILED SOLUTION

2.1 Solution Framework

Figure 1 illustrates the framework of proposed solution. With OCR results as input, the whole procedure is fundamentally divided into two parts: Intra-Slide Reconstruc-

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2508115.

[1] http://www.tele-task.de/

[2] http://videolectures.net/

Published as: Che, X., Yang, H., & Meinel, C. (2013, October). Lecture video segmentation by automatically analyzing the synchronized slides. In Proceedings of the 21st ACM international conference on Multimedia (pp. 345-348). ACM.

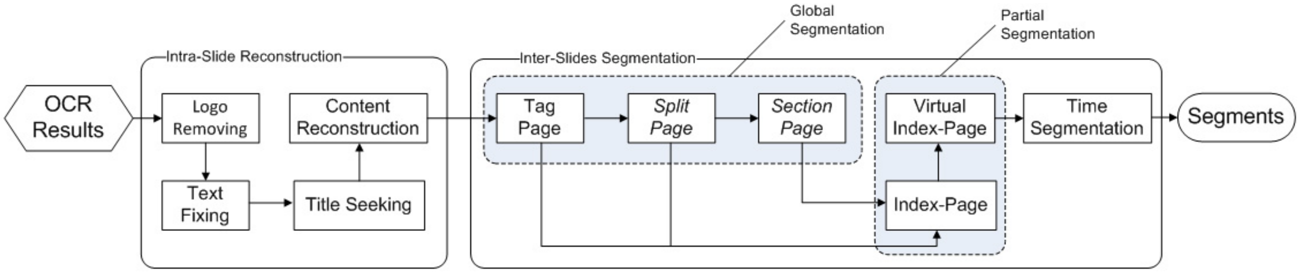


Figure 1: Diagram of proposed solution framework

tion and Inter-Slides Segmentation, which will be further explained in Chapter 2.2 and Chapter 2.3.

2.2 Intra-Slide Reconstruction

2.2.1 Preparation

Firstly, we will try to remove the logo. When existing, the logo appears in the same position of almost every slide, commonly in a corner. But it is totally meaningless or even harmful for our solution.

Next, we will try to fix the error occurred in OCR results. Text-lines will be checked by splitting into words. If the average word length is shorter than 2 characters, this text-line will be deleted entirely. Otherwise, a text-line can also be shortened by eliminating ill-recognized words with an abnormally long length or containing too much symbols.

All further steps will build on the remaining texts, along with their location info.

2.2.2 Title Seeking

Title is the most important content for a slide. Generally, the title has 3 features: bold, locating in the upper part of the slide and being separated with other texts. Sometimes, the title can be so long that have to occupy multiple physical rows. And because of the OCR accuracy problem, text in one physical row may probably be recognized as several text-lines when the gap between two words is large. In addition, potential subtitle will be included as a part of the title in our approach. So we will search for at most 3 text-lines as candidates. Finally they will be sorted by the location logic and combined together as the title.

When seeking a title candidate, a lot of factors have to be considered. But if a text-line can match all the requirements below, it seems very likely to be selected as a title candidate (*with slide resolution 1024×768*):

1. *Higher than the average or 30 pixels.*
2. *Vertically locates in the top 1/3 of the slide.*
3. *Horizontally locates in the left 2/3 of the slide.*
4. *Not closer than 10 pixels to any border.*
5. *If it is not the only text-line matching requirements above, not far from the previous one, neither vertically nor horizontally.*

2.2.3 Content Reconstruction

To organize a group of text-lines in order is a big challenge, especially when the slide contains pictures, charts or tables. But first of all, we will connect those long descriptions occupying multiple physical rows, just like what we do when seeking title. Then we will go through all the text-lines to explore potential hierarchy among them, based on their locations and sizes. As a result, up to 3 levels will be found, according to the custom that most presenters write their slides content from the left horizontally and from the top vertically.

2.3 Inter-Slides Segmentation

2.3.1 Global Segmentation

The aim of global segmentation is to segment the lecture or presentation by its main structure. Most presentations are indeed consisted by several subtopics, but unfortunately only some of them have apparent signs which can be considered as subtopic borders. In this step we attempt to figure out all possible such 'border' and generate segments based on them. We name this kind of segment as GLS (*Globally Logical Segment*).

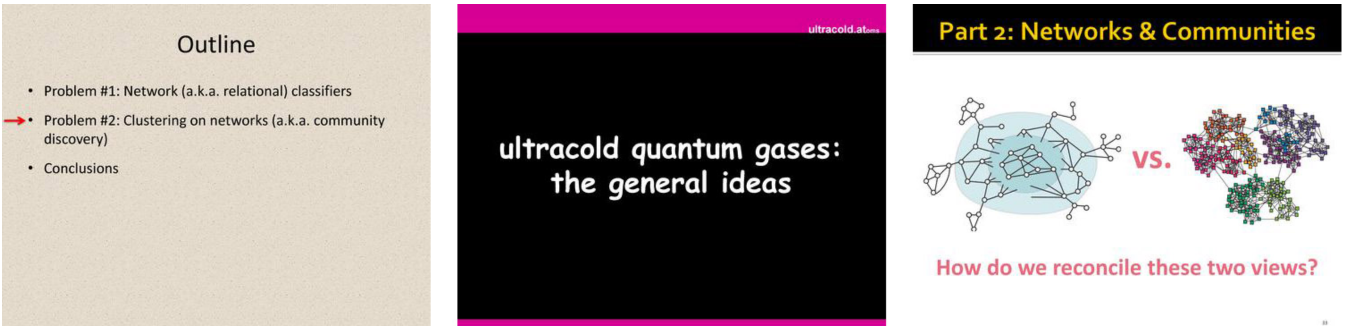
In our research, 3 kind of slide can be identified as GLS basis: tag-page, split-page and section-page. This order also affects the priority when more than 1 kind of GLS basis exists. And at least two slides of same kind should be found before they get utilized as GLS basis. Figure 2 shows examples of these 3 kind of special slides.

Tag-Page:

A tag-page in fact is an outline of the whole slides, with a special title such as 'Agenda', 'Topics' or 'Outline', and its content containing most or all the subtopics. For each time, one certain text-line is highlighted to indicate that this subtopic will be discussed in the following slides until the next appearance of the tag-page. The OCR process cannot recognize the feature of 'highlighted topic' and a tag-page may contain more text-lines than the real number of subtopics (*sub-subtopic e.g.*), but the correct subtopic must be found as the annotation to this GLS. As a result, to get a correct mapping relation between the multi-slides sections and their corresponding subtopics in the tag-page becomes vitally important.

In our research, we try to match the subtopics in the tag-page to the titles in following common slides. However, we have to be very tolerant about the matching requirements, approximately 50%, because in content the subtopics are much more general than the following titles. Even so, not all sections can be matched to a subtopic, so we have to use the already matched ones as reference and assign those 'free' sections around them. Finally several GLSs will be achieved, noticed that the last slide with special title such as 'Thanks' will be separated from the last GLS.

Published as: Che, X., Yang, H., & Meinel, C. (2013, October). Lecture video segmentation by automatically analyzing the synchronized slides. In Proceedings of the 21st ACM international conference on Multimedia (pp. 345-348). ACM.



Outline

- Problem #1: Network (a.k.a. relational) classifiers
- Problem #2: Clustering on networks (a.k.a. community discovery)
- Conclusions

ultracold.at

ultracold quantum gases:
the general ideas

Part 2: Networks & Communities

vs.

How do we reconcile these two views?

Figure 2: Examples of tag-page, split-page and section-page

Split-Page:

Split-page is another kind of widely used 'border' by the lecturer or presenter. Generally the only content in a split-page is a prompt of the following slides, and for most cases, locates in the center of the slide rather than the title position. In other words, a split-page doesn't have a title after our intra-slide reconstruction process, which is a main prerequisite when we search for it.

The text we get from the split-page, obviously, is the best annotation text for this GLS. So there is no such 'matching' procedure required like what we do when dealing with tag-page.

Section-Page:

Not like tag-page or split-page above, section-page is more than a border. A section-page has all the features a common slide may have, expressing definitions, explaining algorithms or showing pictures. The only difference is in the title, which contains some special border words such as 'Part 1:', 'Theme two' or 'Topic III' e.g. Only if when we find at least two slides with special title words in same format, then they can be acknowledged as section-pages.

The section-page title, after removing the special border words, will be applied in annotation.

2.3.2 Partial Segmentation

For those presentations without recognizable subtopic border, or the slides locating before the first GLS, we propose a partial segmentation process to explore the logical correlation among several neighboring slides. Under partial segmentation, PLS (Partially Logical Segment) will be found out, by which some slides with continuous or relevant content can be gathered together. Compared to GLS, PLS is less convincing, but still reasonable. In our research, partial segmentation process affects in 2 steps: index-page and virtual index-page.

Index-Page:

Index-page looks like partial tag-page. The content of an index-page is a preview of a sequence of following common slides, so it is natural to combine all these slides together as a PLS. The method of index-page searching is also similar to the tag-page, by exploring the similarity of text-lines in the potential index-page and titles from following slides. But the threshold of 'matching' here is higher, at least 75%, because this kind of partial correlation is always direct.

Due to the independency of index-page, every slide in a non-GLS presentation, or before the first GLS, has a chance to be an index-page. We have to go over all of them with a floating search window. If over half text-lines in the current can match some following titles, this slide will be marked as index-page and its text-lines will connect those corresponding slides. And similar to tag-page, those unmatched slides will be assigned to the text-lines by position. After that, the search window will jump to the end of the newly generated PLS. If the current slide is not an index-page, the search window will just move to the next.

Finally, one text-line in the index-page may connect to a single following slide or a sequence. And the title of the index-page will be directly adopted as the annotation text of the PLS.

Virtual Index-Page:

Virtual index-page derives from a series of continuous slides sharing some words in their titles. In this case, those slides are very likely describing similar topics, and can be packaged as a whole. In our approach, we only consider nouns longer than 3 characters as candidates, and the noun must appears in half slides inside the counting interval, and at least 3 times, to be finally adopted.

This process is also running inside a floating window, just like the one searching real index-page. And to improve the accuracy, both the plural and the punctuation problem get well solved. At the end, the annotation text of the PLS generated from the virtual index-page will consist of all the possible shared words.

2.3.3 Default Time Segmentation

A time segmentation procedure is reserved to apply for the rest of the presentation except the GLSs and PLSs, by which all segments will not be too long. The length of TS (*Time Segment*) depends on the average length of logical segments in the same presentation, or else, if there is neither GLS nor PLS found, a time segment should not longer than 1/4 of the whole presentation. The annotation text of a TS adopts the title of the slide in this segment with longest duration.

3. EVALUATION

To evaluate the result of our approach, the diversity of test videos is very important. So we collect 10 extra lecture or presentation videos as an additional dataset, along with 10 videos from public dataset (*Videos with slides only*). All videos have different lecturers and both their contents and styles are totally independent with each other.

In our experiments, we just focus on our own method rather than do comparison between methods, according to 2

Published as: Che, X., Yang, H., & Meinel, C. (2013, October). Lecture video segmentation by automatically analyzing the synchronized slides. In Proceedings of the 21st ACM international conference on Multimedia (pp. 345-348). ACM.

Table 1: Segments Overview of Public Dataset

Video	Borders	Segs	GLS	PLS	TS	AveLength
d-g-u	Yes	4	3	0	1	13:29
k-e-d	No	7	0	1	6	05:53
k-k-s	No	6	0	1	5	08:41
k-r-n	Yes	5	3	0	2	12:27
k-s-n	Yes	3	2	0	1	15:02
s-d-s	No	4	0	2	2	11:45
s-e-r-c	Yes	4	2	0	2	14:12
s-k-s	No	5	0	0	5	09:26
s-l-n	Yes	5	2	1	2	09:44
s-w-i	Yes	7	2	2	3	05:06

Table 2: Segments Overview of Additional Dataset

Video	Borders	Segs	GLS	PLS	TS	AveLength
5626	No	5	0	1	4	08:05
6011	No	4	0	0	4	07:11
6031	No	5	0	2	3	06:06
6098	Yes	6	4	0	2	07:58
6102	Yes	6	3	1	2	03:19
6104	No	5	0	1	4	07:12
6106	Yes	5	4	0	1	06:22
6196	No	8	0	4	4	07:20
6201	No	3	0	2	1	16:26
6212	Yes	6	2	1	3	10:21

reasons. Firstly, we have to manually achieve ground truth data before comparing, but this may be too subjective to convince people. Secondly, it is not fair to compare results if they derive from completely different resources, especially when the accuracy of OCR is much better than ASR (*Automatic Speech Recognition*), which is well acknowledged.

The overview segmentation result of the public and additional dataset are depicted in Table 1 and Table 2 respectively. For public dataset, we use the initials of the folder as the video name in 'Video' column, while a series number for the video from additional dataset. 'Border' means whether the synchronized slides of this video contain obvious subtopic border, which is obtained manually. Next we present the numbers of total segments, GLS, PLS and TS generated by our solution. And finally an average length of the segments in each lecture video is calculated.

From the segments overview we can easily find out that a presentation gets commonly split into 3∼8 segments, with the average length controlled inside 5∼15 minutes. It means that a presentation will never be cut too fragmentary, for which each segment will keep plenty of information as complete knowledge points, and the duration of a segment is comfortable for the learners behind the screen, neither too long to feel tired, nor too short to be confused.

Based on the test datasets, the proposed segmentation method is well capable in exploring the subtopics border when they exist (*10 in 10*). According to the further analysis showed in Table 3, over 50% segments we got from this kind of videos is GLS, in addition with almost 10% PLS, makes the segmentation result highly logical. For other videos, there are also over 1/4 segments achieved as PLS, from in fact comparatively lineal organized or discrete consisted presentations. In general, nearly half segments derived from our 20-videos dataset are logical.

Table 3: Segments Ratio Analysis

Video Type	GLS	PLS	All LS	TS
with Border	52.9%	9.8%	62.7%	37.3%
no Border	-	26.9%	26.9%	73.1%
All	26.2%	18.4%	44.6%	55.4%

And to be noticed, each presentation automatically contains one TS because the first slide is always a front cover and never a part of any logical structure. And many presentations will have one more TS at the end due to the large probability that the presenter will use the final slide to express the gratitude. Regarding these two facts, the actual ratio of logical segments should be even higher.

4. CONCLUSION

The proposed solution for lecture video segmentation has been proven effective by the evaluation results. Mainly by comparing text, we have successfully explored lots of logical correlation between slides and apply them into segmentation process. To go further in the future, the method of analyzing must involve more artificial intelligence factor, which is what we will attempt next.

5. REFERENCES

[1] V. Badrinarayanan, I. Budvytis, and R. Cipolla. Semi-supervised video segmentation using tree structured graphical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1, 2013.

[2] G. Friedland and R. Rojas. Anthropocentric video segmentation for lecture webcasts. *Journal on Image and Video Processing*, 2008:9, 2008.

[3] A. S. Imran, L. Rahadianti, F. A. Cheikh, and S. Y. Yayilgan. Semantic tags for lecture videos. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on*, pages 117–120. IEEE, 2012.

[4] I. Koprinska and S. Carrato. Temporal video segmentation: A survey. *Signal processing: Image communication*, 16(5):477–500, 2001.

[5] M. Lin, J. F. Nunamaker Jr, M. Chau, and H. Chen. Segmentation of lecture videos based on text: a method combining multiple linguistic features. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, pages 9–pp. IEEE, 2004.

[6] M. Onishi, M. Izumi, and K. Fukunaga. Blackboard segmentation using video image of lecture and its applications. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 4, pages 615–618. IEEE, 2000.

[7] N. Yamamoto, J. Ogata, and Y. Ariki. Topic segmentation and retrieval system for lecture videos based on spontaneous speech recognition. In *European Conference on Speech Communication and Technology*, pages 961–964, 2003.

[8] H. Yang, M. Siebert, P. Luhne, H. Sack, and C. Meinel. Lecture video indexing and analysis using video ocr technology. In *Signal-Image Technology and Internet-Based Systems (SITIS), 2011 Seventh International Conference on*, pages 54–61. IEEE, 2011.