This paper is published in Proceedings of International Conference on Advanced Learning Technologies (ICALT) 2016

Pre-Course Key Segment Analysis of Online Lecture Videos

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

Thomas Staubitz
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
14482, Potsdam, Germany
thomas.staubitz@hpi.de

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

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

***Abstract*—In this paper we propose a method to evaluate the importance of lecture video segments in online courses. The video will be first segmented based on the slide transition. Then we evaluate the importance of each segment based on our analysis of the teacher's focus. This focus is mainly identified by exploring features in the slide and the speech. Since the whole analysis process is based on multimedia materials, it could be done before the official start of the course. By setting survey questions and collecting forum statistics in the MOOC "Web Technologies", the proposed method is evaluated. Both the general trend and the high accuracy of selected key segments (*over 70%*) prove the effectiveness of the proposed method.**

Keywords—MOOC, Key Segment, Teacher's Focus, Pre-Course

I. Introduction

Distance learning has developed very fast in recent years, especially in form of MOOC (***M**assive **O**pen **O**nline **C**ourse*). Millions of learners all over the world now can access to this ocean of knowledge created by numerous MOOC providers. However, a major concern about MOOCs is the high dropout rate, which means many learners give up at an early stage of the course and lose the chance to make use of the well-prepared course materials in the later stages.

Similarly, if we take a single lecture video as a subject, the phenomenon of dropout also exists. Researchers in [1] addressed it as "in-video dropout". When a user closes a lecture video before the most important part comes, it is a double loss: missing the knowledge and wasting the time. Many discussions and efforts aim to reduce the dropout or in-video dropout rate[2, 3], which is definitely great, but in this paper we would like to think in another way: if dropout for some users is inevitable, how can we help to maximize their learning achievement before they quit?

Our proposal is to segment the lecture video, analyze their importance, mark the most important ones, key segments in other words, and possibly suggest those learners who tend to dropout to watch these key segments first. If a learner still closes the video after watching the key segment, at least he/she has encountered more important knowledge than watching the other parts of this video. And if the key segment arouses the interest of this learner, he/she may decide not to drop out, which is a better result. Besides, for those devoted learners who originally intend to watch the whole video, marked key segments could also be helpful, at least for replay sessions.

In this paper we focus on how to extract the key segment from lecture videos. We believe a good lecturer knows which part of the lecture is more important and reveals this kind of focus in the materials prepared and the speech given[4]. And a good teacher's focus could be similar to the students' focus[5]. Therefore, we will not utilize the user feedback or server statistics in the analyzing process, but taking them as the ground-truth in evaluation. The key segments will be obtained purely by analyzing multimedia materials, including videos, subtitle files and slide documents. The whole process can be done before the course officially starts, which makes the analysis result of key segments available for all the course participants. The rest of the paper will be arranged as follow: related work in section II, analyzing process in section III and then come the evaluation and conclusion.

II. Related Work

Extracting the key segments from the video is well researched on broadcasting sports videos, and in this context the key segment is generally addressed as "Highlight". The decisive factors could be visually extracted domain-based highlight scene (*like a goal attempt in soccer*)[6], keywords detected from the commentator's speech[7], excitement of the audience involved[8] or replay session parsed from professionally produced sport program[9]. For other types of video, segment importance evaluation is generally taken as a step in video summarization or abstraction. The general idea is to deconstruct videos into shots, extract key-frames from these shots and evaluate the importance of the key-frames by their visual features, timing information or a "Bag-of-Importance" model[10–12].

Lecture video, on the other hand, is something different[13, 14]. It has few scene changes, very few excitement from the audience and almost all the key-frames extracted are visually similar, which disables all the possibilities introduced above. However, He *et al.* have attempted on slide-inclusive lecture videos already[15]. They put audio features, slide transition information and user statistics into general consideration to create an abstraction, but focusing more on context integrity than segment importance. Taskiran *et al.* also contributed to lecture video summarization[16]. They segmented the video by detecting pauses in speech and assigned an importance score to each segment by doing lexical analysis on the words extracted from the speech transcript.

These ideas are inspirational for us. We also need to cope with the scenario when slides are inclusive, just as in [15]. But as already mentioned in Section I, we aim to complete everything "pre-course", which excludes any user stats. Besides, we have no plan to involve NLP-based (***N**atural **L**anguage **P**rocessing*) acoustic or lexical analysis in this approach, but will consider some of them in the future.

III. Key Segment Analysis

A. Data Acquisition

In this approach we segment the video with slide transitions, and address these segments as SUs (***S**lide **U**nits*). The beginning and ending time tags of each SU are extracted from the slide-inclusive video, and then a textual outline of each screenshotted slide image will be created based on its OCR (***O**ptical **C**haracter **R**ecognition*) result[17, 18]. Meanwhile, the digital slide file can also be parsed into outlines per page. By comparing the content, two sequences of outlines can be synchronized and each SU will have the time tags from video and an error-free slide outline from external file. With the time tags, subtitle files can also be segmented and each SU will correspond to a paragraph of speech consisting of several sentences. Now each SU will have following direct parameters:

- Type: *T-SU* (pure textual slide), *NT-SU* (except for the title, there is no text in the slide but only illustrations, such as chart, image, etc.) and *HT-SU* (mixed).
- Duration (d): counted in second.
- O-Words (W_O): total words in the slide outline.
- O-Items (I): total number of title, topics and subtopics in the slide outline.
- S-Words (W_S): total words in speech paragraph.
- Co-Occur (C): total words shared by both slide outline and speech paragraph.

And based on these direct parameters, several indirect parameters can also be generated and applied in further process, which include:

- Speaking Rate: $R_S = W_S/(d/60)$
- Matching Rate: $R_M = C/W_O$
- Explanation Rate: $R_E = W_S/W_O$
- Average O-Item Length: $L_I = W_O/I$
- Average O-Item Duration: $d_I = d/I$

The whole data acquisition process is fully automated. If a lecture video is very long and contains too many SUs, it will be first cut into several independent clips by clustering lexically similar slides[19] or exploring the inter-slides logic[20], and then be treated as several independent videos. SU will still be the unit for key segment analysis.

B. Importance Evaluation of T-SU

Slide based on text is most frequently used for a slide-inclusive lecture. The lecturer lists the things he/she wants to mention in the slide, as an outline of the textbook. However, in the speech the lecturer must offer something more than what is listed in the slides, otherwise there is no point for the video: reading slides can be done by learners themselves.

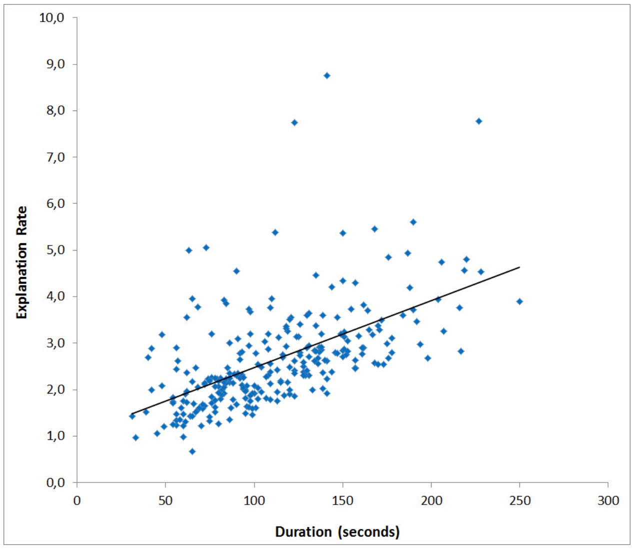


Fig. 1. Ascending Trend of R_E while SU duration increases.

In evaluating the importance of T-SU, we generally focus on the connections between slide and speech information, and also consider the slide structure of the video as an additional factor.

1) Expected Explanation Rate: It is natural to suppose that when the lecturer talks about something important, he/she will try to explain it in more details. A T-SU in such condition will have a comparatively higher explanation rate (R_E). But we believe it is not proper to take the absolute value of R_E as the measurement, because after collecting data from a complete course (*"Web Technologies"*), we found that there is an obvious ascending trend of R_E when SU duration increases, as shown in Fig. 1. Alternatively, we use the data of the whole course to fit a linear trend line. For a SU in this certain course with the duration d, an expected explanation rate $\hat{R}_{E(d)}$ can be calculated. By this effort, the differences between different courses or lecturers can also be distinguished.

Similarly, we are able to obtain another expected explanation rate from the average item length (L_I). It is understandable that when L_I is large, the items in the slide tend to be complete sentences, while a smaller L_I refers that the items are likely to be key-words or key-phrases. Obviously, in order to explain a key-word the lecturer needs to build a complete sentence in the speech, which makes the explanation rate higher. The descending trend line in Fig. 2 (*also by "Web Technologies"*) is an exemplification of this phenomenon and enables us to calculate the second expected explanation rate $\hat{R}_{E(L_I)}$.

Now we intend to connect SU importance with the explanation rate. f_E will be calculated by

$$f_E = R_E - \frac{\hat{R}_{E(d)} + \hat{R}_{E(L_I)}}{2} \tag{1}$$

and taken as the first evaluating factor for T-SU. f_E might be either positive or negative, and we expected the important SUs could get a large and positive f_E.

2) Hypothesis on Speaking Rate and Matching Rate: Lecture speech is generally a kind of solo speech. Especially when shooting the videos for MOOC, many lecturers tend to stand

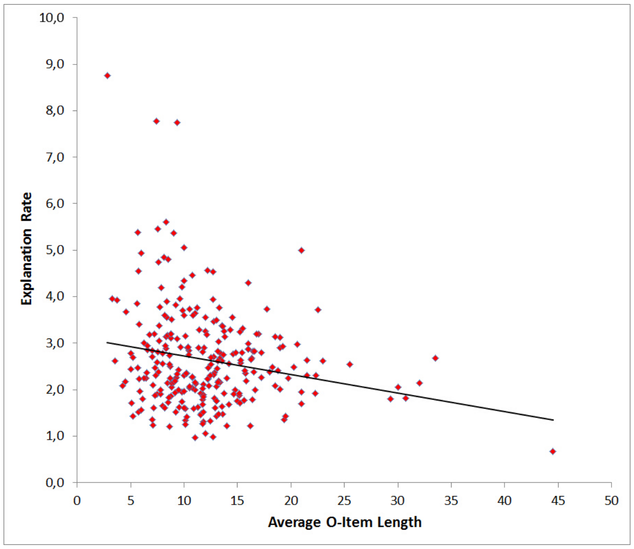


Fig. 2. Descending Trend of R_E while L_I increases.

before a green screen in a studio and talk to a camera[21, 22], rather than in a classroom with real students. In such scenarios, the lecturer is easy to keep a calm mood and speak in a smooth way, which makes the speaking rate stable. However, since most of the lecturers are senior teachers, who are well experienced in teaching, they know where to emphasize. And when emphasizing, slowing down is a frequently used and effective trick[23, 24].

But we cannot simply take the low speaking rate as the evidence for emphasizing, because low speaking rate may result from another reason: pause by hesitation[25]. It is a difficult job to keep talking without actual audience for a long time, even for experienced teachers. When the lecturer gets tired, his/her thinking becomes slow and then unexpected pauses and hesitations will occur. Since we did not implement an acoustic tool to classify them, it is impossible to distinguish a "slow down" event to be either intentional or accidental just by checking the speaking rate.

However, we make a hypothesis that when the lecturer slows down the speaking rate intentionally to emphasize something, the content of the speech should be highly related to the content in the slide, because only important content needs to be emphasized and important content should be included in the slide - it is the reason why people use slides. In this case, the matching rate should be comparatively high. Based on this hypothesis, we introduce the second evaluating factor f_H:

$$f_H = \frac{(R_M - \bar{R}_M) \times 100 - (R_S - \bar{R}_S)}{2} \tag{2}$$

where $\bar{R}_M$ and $\bar{R}_S$ refer to the average values of R_M and R_S of the whole course. We also expect f_H of important SUs to be large and positive.

3) Overview Bonus: Many MOOCs are designed in purpose of science popularization. These courses are not academically advanced, but introduce various possibilities under a certain topic. In these courses, or in some certain stage of these courses, one independent topic can be fully covered in a single short video clip. For this kind of video, overview is the most important part and there is always an overview slide placed in beginning of the video. In our approach if a video clip is not long (*less than 10 minutes*), contains few pages of slide (*less than 10 pages*) and the first slide is an independent slide, which means the titles of first and second slides are obviously different, then we acknowledge the first slide of this video as an overview page and give this SU a bonus (B_O).

By now we can calculate a final "importance value of T-SU": $V_T = f_E + \lambda \times f_H + B_O + \mu$, where λ is a weight to adjust the influence of f_H and μ is a course-based fixed offset to make V_T always positive. Certainly we suppose V_T of important SU, key segments in other words, to be large.

C. Importance Evaluation of NT-SU

In a NT-SU, the slide structure is very simple: a title and a full-page illustration. It could be a chart, a diagram, an image or, in IT courses, a code block. Since there is no available O-Words (W_O), many features we analyzed in T-SU cannot be utilized here, such as the explanation rate and matching rate. Therefore we apply a simple measure: how much information a NT-SU contains, which depends on the S-Words (W_S). We suppose that if a full-page illustration is the description of a core procedure in the technique introduced or a significant exhibition of an important system, the lecturer would explain a lot in the speech, with a large W_S naturally. And for those illustrations which the lecturer just briefly mentions by few words, we consider them as less important. We simply take importance value of NT-SU: $V_{NT} = W_S$.

D. Importance Evaluation of HT-SU

The situation of HT-SU is in between of T-SU and NT-SU. Beside illustrations occupying half the page, there is still a considerable portion of text, which makes all SU parameters available. But because the proportion of information contained by text and illustration is unable to be quantified, explanation rate (R_E) becomes untrustworthy. Instead we implement the average item duration (d_I) as the measurement of how detailed the lecturer explains the content of a certain HT-SU.

On the other hand, similar to NT-SU, we also believe that the importance of a HT-SU is positively related to the amount of information the lecturer gives, including both W_S and W_O. The importance value of a HT-SU will be set as

$$V_{HT} = \frac{W_S + W_O - C}{2} + d_I \tag{3}$$

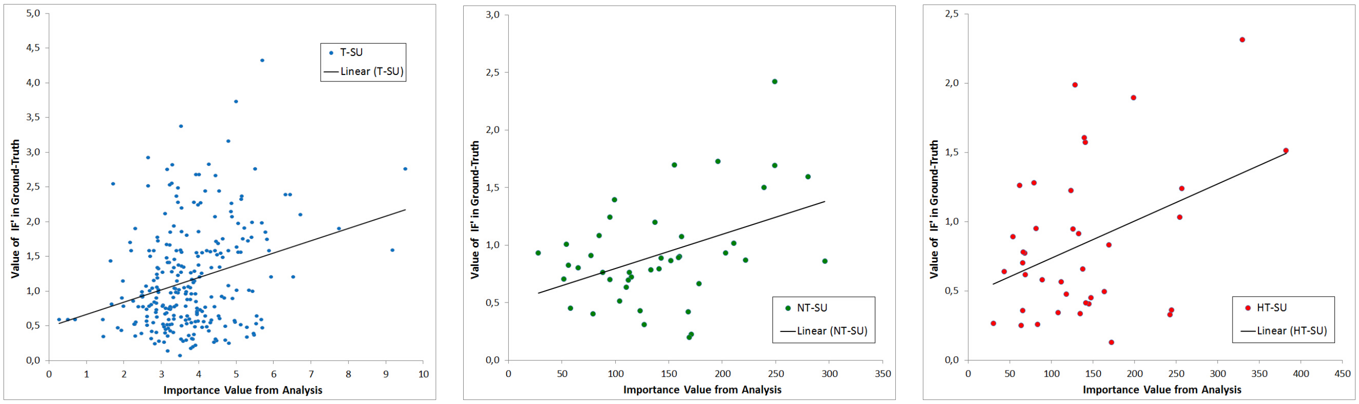
where C is the co-occurrence we intend to remove as redundancy. V_{HT} is supposed to be large when the HT-SU is a key segment.

IV. Evaluation

A. Ground-Truth Acquicision

In order to evaluate the performance of proposed key segment analysis method, we added survey questions in self-tests of the MOOC "Web Technologies". "Web Technologies" is a 6-weeks course in English on openHPI platform[1]. 10022 learners enrolled for the course during the opening time,

[1] https://open.hpi.de/courses/webtech2015



(a) T-SU (b) NT-SU (c) HT-SU

Fig. 3. General Trend Illustration for all 3 Types of SUs

1328 participants attended the final exam and 1179 of them successfully earned a certificate.

We set survey questions for 43 video clips with a total length of 632 minutes, from which 348 SUs are obtained automatically. In the survey we asked the course participants to select a single segment as the most important segment of the correlated video clip. We received over 5000 replies for the first video, and as users dropping out, there were still over 1000 users attending the survey of last video.

For the ith SU in a video having n SUs in total, if u_i users choose it as the most important segment, its importance factor IF_i will be set by

$$IF_i = \frac{u_i}{\sum_{j=1}^{n} u_j} \times n \tag{4}$$

By this calculation the importance of a SU will not be affected by either the different number of SUs in videos or the different number of survey participants. The average important factor of either a video or the whole course is 1.

In addition we also paid attention to the course-attached discussion forum. We believe that the important part of the lecture would intrigue learners to ask questions. So we registered the questions to the related SUs based on content. Each registered question earns a small bonus to the importance factor of the related SU, and the bonus is also balanced since there are obviously more questions in the early stage of the course than the later stage. Suppose the ith SU intrigued q_i questions and η is a coefficient to keep the bonus value proper, the final importance factor for this SU will be:

$$IF'_i = IF_i + \frac{q_i}{\sqrt{\sum_{j=1}^{n} u_j}} \times \eta \tag{5}$$

B. Trend Evaluation

Based on the data collected from "Web Technologies", we set $\lambda = 0.1$, $\mu = 3.5$ and $\eta = 10$. Since V_T, V_{NT} and V_{HT} have different definitions, we will show the evaluation result separately. For all 3 types of SUs, the ground-truth importance IF' is generally increasing while the calculated importance (V_T, V_{NT} *or* V_{HT}) becomes larger, and the ascending trend is obvious, just as illustrated in Fig. 3.

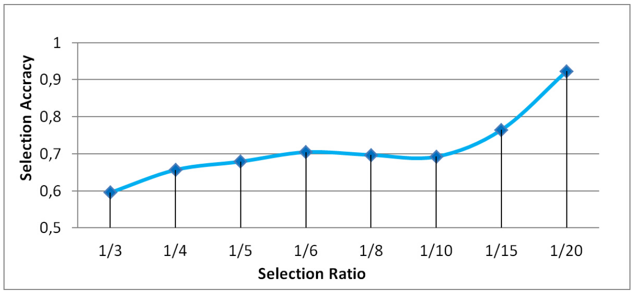


Fig. 4. A ratio of $1/k$ means when sorting all SUs with the calculated importance descending, top $1/k$ SUs will be selected as "key segments". Accuracy generally increases as the selection rate gets smaller.

TABLE I
ACCURACY ANALYSIS ON TOP 1/6 SELECTION

Type	All Segments		Selected Key Segments *(Top 1/6)*			
	Num	A-IF'	Num	A-IF'	Correct	Accuracy
T-SU	268	1.15	44	1.58	31	70.5%
NT-SU	42	0.82	7	1.40	5	71.4%
HT-SU	38	0.83	6	1.13	4	66.7%
All	348	-	57	-	40	**70.2%**

More specifically for T-SU, the hypothesis described in Section III-B-2 independently works not as good as expected, but it is effective to eliminate the T-SUs with low matching rate but high speaking rate. The factor about explanation rate (*in Section III-B-1*) is the foundation of the final result and the overview bonus is a positive boost.

C. Accuracy Evaluation for Selected "Key Segment"

The goal of this proposed work is to recommend the key segments to the learners, in purpose of maximizing their achievement before they drop out or attracting them not to drop out. So comparing to the general trend, the accuracy of the key segments selected is more meaningful practically. If a selected key segment has a ground-truth IF' greater than 1, then it is considered as "Correct". By taking T-SU as example, Fig. 4 shows the accuracy change as the selection ratio varies.

In order to recommend a reasonable proportion of key segments, we intend to take 1/6 as the threshold. Detailed statistics can be found in Table I. For T-SU, NT-SU and HT-SU, the accuracies are 70.5%, 71.4% and 66.7% respectively.

A-IF' represents the "average IF'" of the SUs included. The average of selected segments is obviously larger than the average for all segments, no matter for which class. The general accuracy for all 3 classes is 70.2%.

V. Conclusion

In this paper we proposed a method to select the key segments from online lecture videos purely based on multimedia course materials analysis. Since there is no learner-side statistics required, the whole process can be done pre-course. User feedback was collected in the openHPI course "Web Technologies" for the evaluation, and the result is quite positive. We believe the approach has fulfilled its original aim. In the future we plan to involve more acoustic features into the evaluation or extend from slide unit to sentence unit, in order to provide more accurate key segments to the learners.

References

[1] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller, "Understanding in-video dropouts and interaction peaks in online lecture videos," in *Proceedings of the first ACM conference on Learning@scale conference*. ACM, 2014, pp. 31–40.

[2] C. Meinel, C. Willems, T. Staubitz, and J. Renz, "Reflections on enrollment numbers and success rates at the openhpi mooc platform," *Proceedings of the European MOOC Stakeholder Summit*, pp. 101–106, 2014.

[3] S. Halawa, D. Greene, and J. Mitchell, "Dropout prediction in moocs using learner activity features," *Experiences and best practices in and around MOOCs*, p. 7, 2014.

[4] P. Scott, "Teacher talk and meaning making in science classrooms: A vygotskian analysis and review," *Studies in Science Education*, vol. 32, no. 1, pp. 45–80, 1998.

[5] G. Gibbs and M. Coffey, "The impact of training of university teachers on their teaching skills, their approach to teaching and the approach to learning of their students," *Active learning in higher education*, vol. 5, no. 1, pp. 87–100, 2004.

[6] X. Qian, G. Liu, Z. Wang, Z. Li, and H. Wang, "Highlight events detection in soccer video using hcrf," in *Proceedings of the Second International Conference on Internet Multimedia Computing and Service*. ACM, 2010, pp. 171–174.

[7] Y. Ariki, M. Kumano, and K. Tsukada, "Highlight scene extraction in real time from baseball live video," in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*. ACM, 2003, pp. 209–214.

[8] A. Hanjalic, "Generic approach to highlights extraction from a sport video," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 1. IEEE, 2003, pp. 1–4.

[9] Z. Zhao, S. Jiang, Q. Huang, and G. Zhu, "Highlight summarization in sports video based on replay detection," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 1613–1616.

[10] S. Uchihashi and J. Foote, "Summarizing video using a shot importance measure and a frame-packing algorithm," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 6. IEEE, 1999, pp. 3041–3044.

[11] F. Wang and B. Merialdo, "Multi-document video summarization," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 1326–1329.

[12] S. Lu, Z. Wang, T. Mei, G. Guan, and D. D. Feng, "A bag-of-importance model with locality-constrained coding based feature learning for video summarization," *Multimedia, IEEE Transactions on*, vol. 16, no. 6, pp. 1497–1509, 2014.

[13] H.-P. Chou, J.-M. Wang, C.-S. Fuh, S.-C. Lin, and S.-W. Chen, "Automated lecture recording system," in *System Science and Engineering (ICSSE), 2010 International Conference on*. IEEE, 2010, pp. 167–172.

[14] A. R. Ram and S. Chaudhuri, "Media for distance education," in *Video Analysis and Repackaging for Distance Education*. Springer, 2012, pp. 1–9.

[15] L. He, E. Sanocki, A. Gupta, and J. Grudin, "Auto-summarization of audio-video presentations," in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. ACM, 1999, pp. 489–498.

[16] C. M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. J. Delp, "Automated video program summarization using speech transcripts," *Multimedia, IEEE Transactions on*, vol. 8, no. 4, pp. 775–791, 2006.

[17] H. Yang, M. Siebert, P. Lühne, H. Sack, and C. Meinel, "Automatic lecture video indexing using video ocr technology," in *Multimedia (ISM), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 111–116.

[18] X. Che, H. Yang, and C. Meinel, "Adaptive e-lecture video outline extraction based on slides analysis," in *Advances in Web-Based Learning–ICWL 2015*. Springer, 2015, pp. 59–68.

[19] T. Tuna, M. Joshi, V. Varghese, R. Deshpande, J. Subhlok, and R. Verma, "Topic based segmentation of classroom videos," in *Frontiers in Education Conference (FIE), 2015. 32614 2015. IEEE*. IEEE, 2015, pp. 1–9.

[20] X. Che, H. Yang, and C. Meinel, "Lecture video segmentation by automatically analyzing the synchronized slides," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 345–348.

[21] D. Garcia, M. Ball, and A. Parikh, "L@s 2014 demo: best practices for mooc video," in *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 2014, pp. 217–218.

[22] W. Krauth, "Coming home from a mooc," *Computing in Science & Engineering*, vol. 17, no. 2, pp. 91–95, 2015.

[23] U. Natke, J. Grosser, and K. T. Kalveram, "Fluency, fundamental frequency, and speech rate under frequency-shifted auditory feedback in stuttering and nonstuttering persons," *Journal of Fluency Disorders*, vol. 26, no. 3, pp. 227–241, 2001.

[24] H. Quené, "Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1104–1113, 2008.

[25] D. O'Shaughnessy, "Timing patterns in fluent and disfluent spontaneous speech," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 600–603.