Automatic Digest Generation by Extracting Important Scenes from the Content of Presentations

Hieu Hanh LE[1] Thitiporn LERTRUSDACHAKUL[2] Tetsutaro WATANABE[3]
Haruo YOKOTA[3,4]

[1] Department of Computer Science, Falculty of Engineering, Tokyo Institute of Technology
[2] Application Lab, Software R&D Group, Ricoh Company, Ltd.
[3] Department of Computer Science, Graduate School of Information and Engineering,
Tokyo Institute of Technology
[4] Global Scientific Information and Computing Center, Tokyo Institute of Technology
{hanhlh@de.cs.titech.ac.jp,thitiporn.lertrusdachakul@nts.ricoh.co.jp,tetsu@de.cs.titech.ac.jp,
yokota@cs.titech.ac.jp}

Abstract

Recently, combining a video recording of a presentation along with the digital slides used in it has become popular in E-learning and presentation of archives. For users of the archives, it is useful to preview a digest of such content to grasp the atmosphere and/or an outline of the presentation. This paper proposes a method of automatic digest generation by extracting important scenes from the presentation content. The extracted scenes are chosen based on several factors such as frequency and specificity of words, scene duration and order. Finally, the effectiveness of the proposed methods are evaluated by comparing with testers' answer sets for actual lectures.

1 Introduction

In recent years, the combination of digital slides used in a presentation along with a video recording of the presentation has been increasingly used in E-learning or for presentation-archival purposes. The development of software to easily create web-publishable content of presentations, such as MPMeister [8] developed by Ricoh Co. Ltd., has accelerated this trend.

To improve the effectiveness of E-learning and archiving, it is important to provide functions for searching for appropriate scenes described by given keywords. For this type of content, a scene is defined as a part of video bounded by changes in the slides. In previous studies, we have proposed a number of methods to distinguish the scene described by given keywords from other scenes containing the same slide, caused by reuse or backtracking by the presenter. We developed a system named UPRISE (Unified Presentation Slide Retrieval by Impression Search Engine) [11], and demonstrated the effectiveness of the system's methods [12].

However, in actual situations, one cannot presume that users can always choose good keywords to search for the desired portions in the video. Additionally, there has been a high demand for a brief summary of the presentation before watching the whole content. Although a number of studies on video summarization have been done, their results could not be directly applicable to our application because of the differences in content characteristics.

In this paper, we propose a method for generating a digest of a presentation by concatenating important scenes extracted from the content created by MPMeister. Here, we assume that scenes showing the representative topic of the presentation are important. To judge the representativeness, we pay attention to factors such as frequency and specificity of words in slides, duration, and order of scenes, without analyzing the semantics of slides. Obviously, the semantics is important but still hard to be analyzed correctly. We then evaluate the proposed method by comparing scenes selected by our experimental system with testers' selections to verify the effectiveness of the approach.

This paper is organized into the following four sections. First, we take an overview of related work in section 2. Then, we propose our method in section 3, and report experiments and discuss the results in section 4. Concluding remarks and possibilities for future work are given in section 5.

2 Related Work

In the field of summarizing text documents, there are two major approaches: those that extract sentences that best match a given user query [9], and those showing the general topic of the document [1, 3]. Although our objects, i.e., presentation content, are different from simple text documents, concepts developed for text summarization are also useful for our approach.

There has also been much previous research on speech [2] and video summarization [7, 10]. One of the studies most related to our approach was done by Liwei et al. [4]. However, the slide content that we focus on in this paper has not been considered yet.

We have previously proposed a unified presentation content search system named UPRISE [11] (Unified Presentation Slide Retrieval by Impression Search Engine) focusing not only on the textual information in slides, but also on the duration and order of scenes. Later, we improved our precision by integrating speech [6] and laser pointer [5] information. Although the purpose of this study is the extraction of important scenes representing the presentation topics, we utilize the techniques developed in the UPRISE research project.

3 Method Proposal

In this paper, we propose a method of automatic digest generation by extracting important scenes from presentation content created by MPMeister. With the words and timing information extracted from MPMeister, we propose a digest generation method based on several important scene extractions, which consists of the following three steps: 1) preprocessing, 2) important scene extraction, and 3) digest generation.

3.1 Preprocessing

First, we pick up the text content of the slides used in the scenes, and the timing information from the recorded presentation. Second, considering slide traverses and errors in operation, we filter out scenes with durations less than three seconds. In addition, we remove exceptional scenes that may have extremely long duration, but no meaningful information, such as scenes in which students silently carry out exercises in the lecture. In this paper, we tentatively set this threshold time to ten minutes, but a more suitable time or other feature, such as distinguishing the exercise, could be considered in the future.

Because we use Japanese lectures as experimental data, we apply the Japanese language morphological analysis Sen to extract only the meaningful words from the text content. Finally, we use Japanese-English Ontology to discover the same English and Japanese words in meaning and treat them as the same words.

3.2 Important scene extraction

3.2.1 Assumptions

Suppose that a presentation L_l is one of the recorded presentations in the database. Then, $S = \{s_1, s_2, \ldots, s_N\}$ represents the set of all slides used in the presentation L_l. In our methods, we focus on the frequency of words in slides. Therefore, each slide s_i can be expressed as an array of words $s_i = [v_{i_1}, v_{i_2}, \ldots, v_{i_m}]$. Additionally, let $W = \{w_1, w_2, \ldots, w_M\}$ be the set of words that appear in all slides. Obviously, $v_{i_j} \in W$. Furthermore, the presentation L_l can be represented by an array of scenes determined by slide transitions, $L_l = [c_{l_1}, c_{l_2}, \ldots, c_{l_n}]$. The scene c_{l_k} corresponds to one of the slides in S, which can be expressed as $c_{l_k} = [v_{i_1}, v_{i_2}, \ldots, v_{i_m}], \exists i : s_i \in S$. Note that because of backtracking or reuse, a slide can appear in more than one scene.

3.2.2 Calculation of scene importance

Based on the idea of "The concept that is repeatedly mentioned is an important concept", all of these formulas focus on word frequency. In this process, we also consider the structure of the slide by using weights $p(v)$ of the word v:

$$p(v) = \begin{cases} \rho_t & v \text{ appears in the title} \\ \rho_{b_l} & v \text{ appears at indent level } \mathrm{b_l} \\ 0 & v \text{ does not appear} \end{cases} . \quad (1)$$

Considering the other factors that effectively distinguish scenes in E-learning content, we select four influential factors: scene duration, scene order, word specificity (such as idf), and number of words appearing in a slide. We propose five formulas to put those factors into the importance calculation, where the influence of the scene duration factor is changed by a parameter and those of the other three factors are implemented by changing combinations.

Scene duration importance I_d Suppose that the longer a scene is, the more detailed the concepts in this scene are explained, so we rate the importance of this scene highly. Therefore, as in UPRISE, we propose a formula for I_d that considers the duration of scenes:

$$I_d(c_{l_k}, \theta) = I_p(c_{l_k}) \cdot t(c_{l_k})^{\theta}, \tag{2}$$

where $t(c_{l_k})$ is the duration of scene c_{l_k}, θ is the time parameter, and I_p can be calculated as:

$$I_p(c_{l_k}) = \sum_{w_x \in c_{l_k}} (\sum_{c_y \in L_l} (\sum_{v_z \in c_y, v_z = w_x} p(v_z))). \tag{3}$$

Scene order importance I_{dc} When a word appears not only in a certain scene, but also in the neighboring scenes, it can be seen that this word is explained well in this scene. Based on this idea, as in UPRISE, we rate the importance of such scenes highly. The formula I_{dc} captures this:

$$I_{dc}(c_{l_k}, \theta, \delta, \varepsilon_1, \varepsilon_2) = \sum_{j=\gamma-\delta}^{j=\gamma+\delta} I_d(c_{l_k}, \theta) \cdot E(j-\gamma, \varepsilon_1, \varepsilon_2), \tag{4}$$

where δ is a window-size parameter that determines how many neighboring slides are taken into account. $E(j-\gamma, \varepsilon_1, \varepsilon_2)$ specifies the effect of neighboring scenes in the context window and is calculated as follows:

$$E(x, \varepsilon_1, \varepsilon_2) = \begin{cases} \exp(\varepsilon_1 x) & (x < 0) \\ \exp(-\varepsilon_2 x) & (x \geq 0) \end{cases} .$$

Note that the smaller ε is, the more likely neighboring slide information is used.

Word specificity importance I_{dr} When focusing on the importance of a word itself, such as idf (Inverse Document Frequency), it is thought useful to consider the word frequency in other scenes. Therefore, we propose a formula I_{dr} to capture the specificity of words:

$$I_{dr}(c_{l_k}, \theta) = \sum_{w_x \in W} \frac{\mathrm{app}(c_{l_k}, w_x)}{\sum_{c_y \in L_l} \mathrm{app}(c_y, w_x)} \cdot I_{pl}(c_{l_k}) \cdot t(c_{l_k})^{\theta}, \tag{5}$$

where, $\mathrm{app}(c_y, w_x)$ is a function that counts the number of appearances of word w_x in scene c_y, and I_{pl} is calculated according to:

$$I_{pl}(c_{l_k}) = \sum_{v_x \in c_{l_k}} p(v_x). \tag{6}$$

Scene order and word specificity importance I_{drc} By combining the factors above, we obtain a formula expressing both scene order and word specificity:

$$I_{drc}(c_{l_k}, \theta, \delta, \varepsilon_1, \varepsilon_2) = \sum_{j=\gamma-\delta}^{j=\gamma+\delta} I_{dr}(c_{l_k}, \theta) \cdot E(j-\gamma, \varepsilon_1, \varepsilon_2). \tag{7}$$

Table 1. Summary of factors considered in the proposed formulas

Formula	Considered factors				
	word frequency	scene duration	scene order	spec of word	#words in slide
I_d	○	○			
I_{dc}	○	○	○		
I_{dr}	○	○		○	
I_{drc}	○	○	○	○	
I_{df}	○	○			○

Number of words appearing in slide importance I_{df} Through the above proposed formulas, we can see that the more words there are in a slide, the higher the importance of the scene using this slide. However, where there are few words in the slide, it can be suggested that those words are better explained compared with others. Therefore, these kinds of scenes should be given high importance. To achieve this idea, we propose

$$I_{df}(c_{l_k}, \theta) = \frac{1}{\sum_{w_x \in W} \mathrm{app}(c_{l_k}, w_x)} \cdot I_d(c_{l_k}, \theta). \tag{8}$$

The factors considered in the proposed formulas are summarized in Table 1.

3.2.3 Important scene extraction method

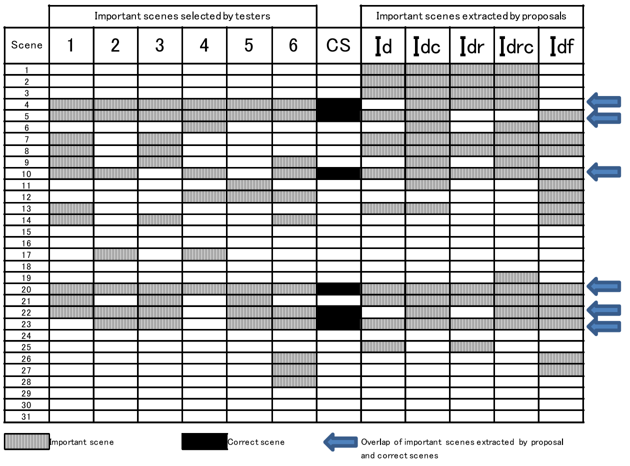
By applying each formula to calculate the importance of scenes, we then extract those scenes that have greater than average importance as important scenes.

3.3 Automatic digest generation

To make the digest video, we truncated the extracted important scenes and concatenated them. The durations of the truncated scenes are determined by the ratios of importance of the scenes.

4 Experimentation and Evaluation

The purpose of our research is to create an automatic digest generation method from presentation content, and so far we have evaluated the important scene



	Important scenes selected by testers							Important scenes extracted by proposals					
Scene	1	2	3	4	5	6	CS	Id	Idc	Idr	Idrc	Idf	
1								▨	▨	▨	▨		
2								▨	▨	▨	▨		
3								▨	▨	▨	▨		
4	▨	▨	▨	▨	▨	▨	■		▨	▨	▨		⇐
5	▨	▨	▨	▨	▨	▨	■	▨	▨			▨	⇐
6				▨					▨		▨		
7	▨		▨					▨	▨	▨	▨	▨	
8	▨		▨					▨	▨	▨	▨	▨	
9	▨		▨			▨			▨		▨		
10	▨	▨		▨		▨	■	▨	▨	▨	▨	▨	⇐
11					▨				▨			▨	
12				▨	▨	▨						▨	
13	▨							▨	▨			▨	
14	▨		▨			▨						▨	
15													
16													
17		▨		▨									
18													
19											▨		
20	▨	▨	▨	▨	▨	▨	■	▨	▨	▨	▨	▨	⇐
21	▨		▨		▨			▨	▨	▨	▨	▨	
22	▨	▨	▨		▨	▨	■		▨		▨	▨	⇐
23		▨	▨		▨	▨	■	▨	▨	▨	▨	▨	⇐
24													
25								▨		▨			
26						▨						▨	
27						▨						▨	
28						▨							
29													
30													
31													

▨ Important scene ■ Correct scene ⇐ Overlap of important scenes extracted by proposal and correct scenes

Figure 1. Difference between important scenes extracted by testers and our proposal

extraction methods that we proposed in this paper by two-tester-based experiments. The first experiment was aimed to verify the effectiveness of our proposed method. The second one was to evaluate the impact of the two parameters, ρ_t and θ, which are used in our proposed five formulas.

4.1 Effectiveness verification

4.1.1 Experiment circumstance and method

We applied our proposed formulas (2),(4),(5),(7),(8) to a lecture in Japanese that had taken place at our university.

Next, we asked six testers who have not previously seen the original video of the lecture to choose the important scenes that they would like to include in a digest. Then, we define a correct set (symbolized as CS), which contains those scenes that were chosen by more than four testers. Here, we set values of parameters used in proposed formulas as $\rho_t = 5$, $\rho_{b_l} = 1$, $\forall b_l$, $\theta = 1$, $\delta = 3$, $\varepsilon_1 = 5$, and $\varepsilon_2 = 0.5$.

4.1.2 Experimental results and discussion

The experimental results are shown in Figure 1. This figure indicates that almost all of the scenes included in CS can be found by our proposal. Therefore, our important scene extraction methods are thought to be effective for generating digests.

4.2 Evaluation of impact of factors

In this experiment, we evaluated the impact of the two parameters ρ_t, which is the weight given to a word appearing in the title of the slide, and θ, which determines the effect of scene duration.

4.2.1 Experiment circumstance and method

In this experiment, we used recordings of four other lectures that took place at our university.

¿From the previous experiment, we selected two representative testers and let them choose scenes that they would like to include in a digest from the lecture video. Then, the CS was defined as those scenes chosen by both testers. Next, we evaluated the impact of the two parameters of ρ_t and θ by F measure (symbolized as F), which is calculated from the CS and the set of important scenes extracted by our proposal. The greater F is, the better the performance of the system.

At first we changed the value of ρ_t from 1 to 10 in steps of 1 and calculated F to evaluate the impact of ρ_t. Then, for each lecture and formula, we changed the value of θ from 0 to 5 in steps of 0.5. Note that in this last step, we set ρ_t to that which resulted in the highest F in the previous step.

4.2.2 Experimental results and discussion

The experimental results of the average value of F are shown in Figure 2 and Figure 3, as a function of ρ_t and θ, respectively.

First, Figure 2 indicates that for almost all formulas, it can be seen that the values of F were higher when $\rho_t > 1$. Therefore, for extracting important scenes, the consideration of ρ_t is somewhat important; however, the results are not too sensitive to this parameter.

¿From the experimental results shown in Figure 3, it is easily recognized that F is lowest when $\theta = 0$. It indicates that the treatment of scene duration is very important in the proposed approach for extracting important scenes. However, the results also indicate that there was a trend of worsening F as θ grows large. Thus, in order to extract important scenes from presentation content, the impact of duration time of scenes should not be too large.

Last, but not least, Figure 2 and Figure 3 also show that method I_{df} was the best for extracting important scenes, so the consideration of number of words appearing in the slides was the most effective.

5 Conclusion and Future Work

In this paper, we propose methods for generating a digest of a presentation using digital slides, which concatenates important scenes extracted from all scenes in the presentation. To select the important scenes, we

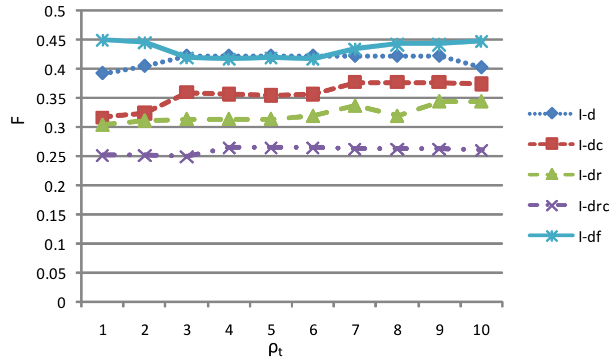


Figure 2. The average value of F as ρ_t varied from 1 to 10

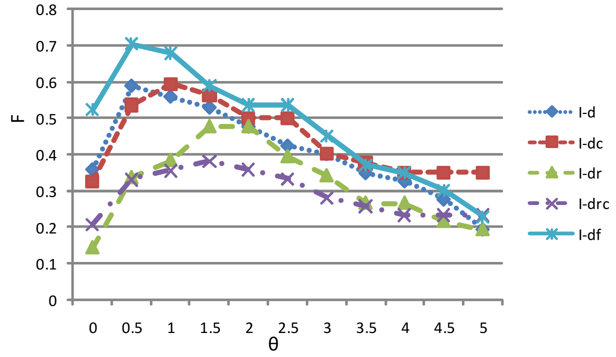


Figure 3. The average value of F as θ varied from 0 to 5

define a number of formulas focused on factors in the presentation content: the frequency of words appearing in slides, scene duration, and the order of scenes. We then developed an experimental system and evaluated the extraction method by comparing scenes selected by the system with those selected by testers from recordings of lectures that took place in our university. The experimental results indicated that the consideration of two factors of scene duration and number of words appearing in a slide were influential.

In the future, we plan to do the following:

- Perform a more detailed verification of the proposed method using a higher volume of presentation content.
- Consider other Information, such as speech and animation, to segment the scenes into smaller units.
- Integrate the ontology by considering the relationship between words and contriving semantically aware filtering methods to remove noise from the presentation content.

Acknowledgment

This work is partially supported by a Grant-in-Aid for Scientific Research of MEXT Japan (#19024028), Tokyo Institute of Technology 21 COE Program “Framework for Systematization and Application of Large-Scale Knowledge Resources”, and CREST of JST (Japan Science and Technology Agency).

References

[1] R. Barzilay and M. Elhadad. Using Lexical Chains for Text Summarization. In *Proc. of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*, 1997.

[2] H. Chiori. *A Study on Statistical Methods for Automatic Speech Summarization*. PhD thesis, Tokyo Institute of Technology, 2002.

[3] Y. Gong and X. Liu. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25, 2001.

[4] L. He, E. Snocki, A. Gupta, and J. Grudin. Auto-Summarization of Audio-Video Presentation. In *Proc. of the seventh ACM international conference on Multimedia (Part 1)*, pages 489–498, 1999.

[5] W. Nakano, T. Kobayashi, Y. Katsuyama, S. Naoi, and H. Yokota. Treatment of Laser Pointer and Speech Information in Lecture Scene Retrieval. In *Pro. of the Eighth IEEE International Symposium on Multimedia (ISM'06)*, 2006.

[6] H. Okamoto, W. Nakano, T. Kobayashi, S. Naoi, H. Yokota, K. Iwano, and S. Furui. Presentation-Content Retrieval Itergrated with the Speech Information. In *IEICE Transactions on Information and System*, pages 209–222, 2007.

[7] M. J. Pickering, L. Wong, and S. M. Rüer. *Image and Video Retrieval*. Springer Berlin / Heidelberg, 2003.

[8] Ricoh Japan. MPMeister II — Ricoh Japan. http://www.ricoh.co.jp/mpmeister.

[9] M. Sanderson. Accurate User Directed Summarization from Existing Tools. In *Proc. of the 7th International Conference on Information and Knowledge Management (CIKM98)*, 1998.

[10] J. Wang, C. Xu, E. Chng, and Q. Tian. Sport Highlight Detection from Keyword Sequences using HMM. In *Proc. of IEEE International Conference on Multimedia and Expo, ICME '04*, volume 1, pages 599–602, 2004.

[11] H. Yokota, T. Kobayashi, T. Muraki, and S. Naoi. UPRISE: Unified Presentation Slide Retrieval by Impression Search Engine. *IEICE Trans. on Info. and Syst.*, E87-D(2):397–406, 2 2004.

[12] H. Yokota, T. Kobayashi, H. Okamoto, and W. Nakano. Unified Contents Retrieval from an Academic Repository. In *Pro. of International Symposium on Large-scale Knowledge Resources*, 2006.