論文 / 著書情報
Article / Book Information

Title(English)	Presentation Scene Retrieval Exploiting Features in Videos Including Pointing and Speech Information
Authors(English)	Takashi Kobayashi, Wataru Nakano, Haruo Yokota, Koichi Shinoda, Sadaoki Furui
Citation(English)	Proc. Symposium on Large-Scale Knowledge Resources(LKR2007)., Vol. , No. , pp. 95-100
発行日 / Pub. date	2007, 3

Presentation Scene Retrieval Exploiting Features in Videos Including Pointing and Speech Information

Takashi Kobayashi†, Wataru Nakano‡, Haruo Yokota†‡, Koichi Shinoda‡ and Sadaoki Furui‡

† Global Scientific Information and Computing Center, Tokyo Institute of Technology
‡ Dept. of Computer Science, Grad. School of Info. Sci. & Eng., Tokyo Institute of Technology
{tkobaya@gsic, wnakano@de.cs, yokota@cs, shinoda@cs, furui@cs}.titech.ac.jp

Abstract

Unified presentation contents are widely used for e-learning and/or e-training. There is a strong need for efficient search mechanism for unified presentation contents. We have proposed a unified presentation contents search mechanism named UPRISE (Unified Presentation Slide Retrieval by Impression Search Engine), and have also proposed a method to use laser pointer and speech information in lecture scene retrieval. In this paper, we discuss how to exploit various features in lecture videos such as the laser pointer and speech information for the efficient retrieval method. We evaluate our approach by using actual lecture videos and presentation slides.

Index Terms: Video Scene Retrieval, e-learning, Information Integration, Speech Information, Pointing Information, Metadata.

1 Introduction

Unified presentation contents, which consist of multimedia contents, such as presentation slides, video and documents, are widely used in a variety of contexts, such as e-learning and e-training. Many systems to store and distribute such unified presentation contents have been proposed [10, 3, 1].

In order to adapt to the increasing number of contents, there is a strong need for efficient search mechanism for unified presentation contents. In particular, it is important for e-learning systems that users can not only retrieve suitable contents but also find a scene they should start to see effectively.

There are several existing works investigating cross-over retrievals for lecture contents [9, 5, 4, 2]. However, some of them do not realize actual retrieval methods and systems. [2] provides a lecture passages retrieval system using transcription by speech recognition. However, they do not consider the case of backtracking or reuse of slide materials.

We have proposed the UPRISE (Unified Presentation Slide Retrieval by Impression Search Engine) [16, 7, 17, 12, 11, 13] which combines slides in a presentation and a video recording the presentation, and retrieves a sequence of desired presentation scenes from archives of the combined contents.

In UPRISE, we modeled a unified presentation contents as a sequence of scene divided by switching slides. Figure 1 shows our model of unified presentation contents. Each scene has metadata extracted from various information such as slide, speech and pointing information. The UPRISE retrieves a relevant scene by weighting schemata considering keyword position in the slides, duration of a scene, and context in the presentation by using metadata. In our previous work[16], we have shown that the precision of the UPRISE is much better than that of ordinary *tf-idf* based approaches.

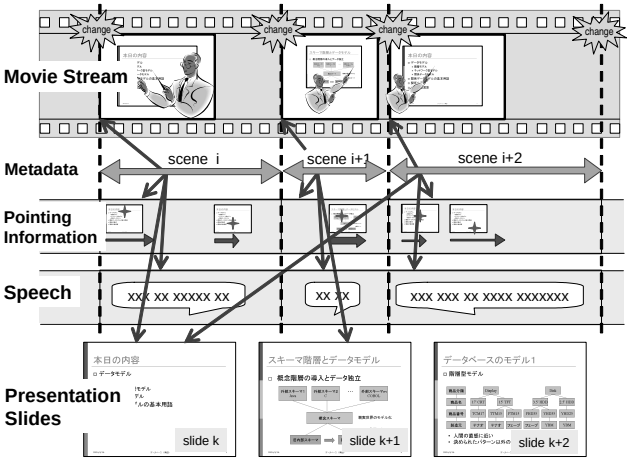

Figure 1. Unifying a presentation video and slides

We have proposed an integration method of the speech information in the lecture videos into the search functions to improve the precision of scene retrieval [13]. In [13], we have proposed four ways to integrate the influence of speech into the scene search functions of the UPRISE. The experimental results indicate that the speech information is effective to improve the precision.

We have also proposed methods to reflect the influence of laser pointer on the weighting schemata[12] and three improved weighting schemata by using methods to filter the irrelevant pointing information based on keyword occurrences in slides and speech[11].

In this paper, we discuss how to exploit various features in lecture videos such as the laser pointer and speech information for the efficient retrieval method. We introduce our information integration approach and weighting schemata with laser pointer information and speech information. Moreover, we evaluate our approach by using actual lecture videos and presentation slides.

The reminder of this paper is organized as follows. In Sect. 2, we introduce the base weighting schemata to retrieve

the unified contents we have proposed. Then we discuss the treatment of the speech information in lecture scene retrieval in Sect. 3 and introduce weighting schemata using the laser pointer information in Sect. 4. Sect. 5 reports experiments and results using actual lecture materials. We summarize the paper's main points in the final section.

2 Base Weighting Schemata in UPRISE

2.1 Considering Slide Structure Information

We consider the structure of each lecture slide. If a given word appears in the title of the slide or less indented lines, the value of the position impression is high, whereas if the keyword only appears in deep indented lines, the value is lower. The expression used to calculate the weighting schemata combined with the slide structure is:

$$I_p(s,k) = \sum_{l=1}^{L(s)} P(s,l) \cdot C(s,k,l)$$

where s denotes an identifier of the objective scene, k a target keyword, $L(s)$ the total number of lines in a slide used in scene s, $P(s.l)$ a function of the assigned point in the line l in the slide for scene s, and $C(s,k,l)$ a function of counting keywords k in the line l of the slide for scene s.

2.2 Addition of Scene Duration Information

Duration information is useful in distinguishing multiple appearances of the same slide caused by backtracking or reuse by the lecturer. To reflect the duration information of scenes to the weighting schema, we have proposed the duration-impression indicator. The value of the position-impression indicator is modified by the presentation time with a duration parameter:

$$I_d(s,k,\theta) = T(s)^\theta \cdot I_p(s,k),$$

where $T(s)$ denotes the time used for scene s, and θ is the duration parameter for changing the influence of the time factor.

2.3 Addition of Context Information

We have combined the information of slide appearance sequence to reflect the influence of context on the weighting schemata, which accumulates values of the duration-impression indicator within a presentation window indicated by a window-size parameter δ:

$$I_c(s,k,\theta,\delta,\varepsilon_1,\varepsilon_2) = \sum_{\gamma=-\delta}^{\delta} E(\gamma,\varepsilon_1,\varepsilon_2) \cdot I_d(s+\gamma,k,\theta)$$

where $E(x,\varepsilon_1,\varepsilon_2)$ is a function to specify the effect of neighboring scenes in the context window. The effect of distance between scenes is decided using the exponential function with distance-effect parameters of ε_1 and ε_2.

We describe these parameters $(s,k,\theta,\delta,\varepsilon_1,\varepsilon_2)$ as Φ in this paper.

2.4 Considering Rareness of a Term

Considering the rareness of a term is important for the weighting schemata in information retrieval. For *tf-idf* , the inverse document frequency, *idf*, is the rareness factor which indicates the rareness of a term. While the appropriateness factor, *tf*, can be calculated simply in a document, the range for calculating the rareness factor is variable. For example, in scene retrieval for lecture videos in universities, there are several types of target range, such as a class, a course, courses in a school, and courses in the university.

We have proposed $iafr(k,\lambda)$, the inverse keyword k appeared scene frequency in range specified by λ, as a dedicated rareness factor for scene retrieval [17].

$$iafr(k,\lambda) = log\frac{\text{\# of scenes in range } \lambda}{\text{\# of scenes having keyword } k \text{ in its slide}}$$

We can apply $iafr(k,\lambda)$ to all weighting schemata we previously discussed as the rareness factor.

3 Exploiting Speech Information

We have proposed a method for using the speech information in a video [13]. If a target keyword not only appears in the slide for a scene but also is frequently uttered in the scene, the scene has to be highly ranked because the keyword is explained well in the scene.

In [13], we have proposed $skc(s,k)$, the number of utterance target keyword k in scene s. We have also proposed weighting schemata to reflect the influence of speech by combination *skc* and I_c.

Since the slide sequence information and speech information are independent, we should consider how the influence of these information spreads to the neighbor scenes. We have proposed the separated context consideration methods and the integrated context consideration methods. Moreover, to deal with speech recognition errors, we have proposed two criteria for the calculation of the influence of speech information. The first criterion is that the influence of speech information is calculated only when the uttered keyword appears in a slide of the scene($I_p \neq 0$). As the second criterion, we relax the condition of the appearance range to neighbor scenes($I_c \neq 0$). As the combination of above consideration, we have proposed four indicators.

First indicator is $I_{c+skc/p}$, which reflect the influence of speech information only when the keyword appear in slides of the scene and is uttered in same scene. The definition of $I_{c+skc/p}$ is as follows:

$$I_{c+skc/p}(\Phi,\psi) = \begin{cases} I_c(\Phi) + \psi \cdot T(s)^\theta \cdot skc(s,k) & (I_p \neq 0) \\ I_c(\Phi) & (I_p = 0) \end{cases}$$

where ψ is the spoken-keyword-count parameter to change the effect of speech on the rating.

As next indicator $I_{c+skc/c}$, we relax the condition of the ap-

pearance range to neighbor scenes.

$$I_{c+skc/c}(\Phi,\psi) = \begin{cases} I_c(\Phi)+\psi\cdot T(s)^\theta\cdot skc(s,k) & (I_c\neq 0)\\ 0 & (I_c=0)\end{cases}$$

On the other hand, $I_{c+skc/c}$ is another variation of the first one. We consider the influence of speech information of neighbor scenes when the keyword appears in the scene.

$$I_{c[p+skc/p]}(\Phi,\psi) = \begin{cases} I_{c[p+skc]}(\Phi,\psi) & (I_p\neq 0)\\ I_c(\Phi) & (I_p=0)\end{cases}$$

$$I_{c[p+skc]}(\Phi,\psi) = \sum_{\gamma=s-\delta}^{s+\delta} E(\gamma-s,\varepsilon_1,\varepsilon_2)\cdot T(\gamma)^\theta\cdot\{I_p(\gamma,k)+\psi\cdot skc(\gamma,k)\}$$

As fourth indicator, we have proposed $I_{c[p+skc/c]}$ which reflects the influences of speech information of neighbor scenes when the keyword appears in neighbor scenes:

$$I_{c[p+skc/c]}(\Phi,\psi) = \begin{cases} I_{c[p+skc]}(\Phi,\psi) & (I_c\neq 0)\\ 0 & (I_c=0)\end{cases}$$

We have also proposed $isfr(k,\lambda)$, the inverse keyword k spoken scene frequency in a range λ, as another rareness factor dedicated to speech information[13].

$$isfr(k,\lambda) = log\frac{\#\text{ of scenes in range }\lambda}{\#\text{ of scenes uttering keyword }k}$$

To apply rareness factor to an indicator I_x, we combine it with $isfr$ and $iafr$ using ψ.

$$I_x\cdot iafr\cdot isfr(\Phi,\psi,\lambda) = \begin{cases} I_c(\Phi)\cdot iafr(k,\lambda) & (I_p\cdot I_c=0)\\ I_c(\Phi)\cdot iafr(k,\lambda)+\psi\cdot I_x\cdot isfr(k,\lambda) & (otherwise)\end{cases}$$

4 Exploiting Pointing and Speech Information

4.1 Extraction of Laser Pointer Information

In [12], we first extract the laser pointer information by using the method of extracting radiant information of the laser pointer as coordinates in the slide by an image analysis technique[6]. We extract subscenes from a scene to make each subscene contains a continuous pointer information. Since there is ambiguity regarding target keywords caused by shaking or any other habit of the lecturer, we distribute the possibilities of a hit by the pointer in a subscene to the neighborhood lines and make their sum equals to 1. $H(l,q)$ denotes the hit probability of line l in subscene q of scene s. We make $H(l,q)$ related to the distance between the line l and the point. We then propose an indicator $phd(s,k)$ by multiplying the possibilities of a hit by the duration of each subscene.

$$phd(s,k) = \sum_{q_i\in s}\sum_{l=1}^{L(s)} H(l,q_i)\cdot T(q_i)$$

To reflect the effect of laser pointer information on the weighting schemata, we modify Ic by adding $phd(s,k)$ to the term for the scene duration, and we denote it as $I_{c[d+phd]}$:

$$I_{c[d+phd]}(\Phi,\omega_d) = \sum_{\gamma=s-\delta}^{s+\delta} E(\gamma-s,\varepsilon_1,\varepsilon_2)\cdot I_{d[d+phd]}(\gamma,k,\theta,\omega_d)$$

$$I_{d[d+phd]}(s,k,\theta,\omega_d) = \{T(s)+\omega_d\cdot phd(s,k)\}^\theta\cdot I_p(s,k),$$

where ω_d is the pointer-hit-duration parameter that change the effect of the duration of a hit by the laser pointer.

We have evaluated the weighting schemata combined with the laser pointer information using an actual lecture material in [12], and the experimental results indicated that the laser pointer information is effective in improving the precision of retrieval.

4.2 Pointing Information Filtering based on Keyword Occurrence in Slides and/or Speech

Since the laser pointer pointing has several meanings for example, to emphasize the keyword, to illustrate a relationship between multiple concepts, and ambiguously pointing. We have proposed two filtering methods to moderate the influence of irrelevant laser pointer information for scene retrieval [11].

As the first filtering method, we have proposed "p-filter", in which the laser pointer is ignored except when all keywords in a query exist in the slide. We have suggested that the laser pointer pointing to a line not containing all target keywords have few meaning for the query. We have defined $phd_{/p}(s,k)$ as $phd(s,k)$ considering the condition:

$$phd_{/p}(s,k) = \begin{cases} phd(s,k) & \prod_{k\in K} I_p(s,k)\neq 0\\ 0 & \prod_{k\in K} I_p(s,k)=0\end{cases}$$

where K is a set of keywords in a query.

We modify $I_{c[d+phd]}$ by replacing $phd(s,k)$ by $phd_{/p}(s,k)$, and we denote it as $I_{c[d+phd/p]}$ as follows:

$$I_{c[d+phd_{/p}]}(\Phi,\omega_d) = \sum_{\gamma=s-\delta}^{s+\delta} E(\gamma-s,\varepsilon_1,\varepsilon_2)\cdot I_{d[d+phd_{/p}]}(\gamma,k,\theta,\omega_d)$$

$$I_{d[d+phd_{/p}]}(s,k,\theta,\omega_d) = \{T(s)+\omega_d\cdot phd_{/p}(s,k)\}^\theta\cdot I_p(s,k)$$

When a laser pointer hits target keywords in a slide and keywords are not spoken by a lecturer in the scene, we can consider that the lecturer do not use the laser pointer for emphasizing the keywords. To moderate the effects of such laser pointer information, we have proposed "s-filter" in which the laser pointer is ignored except when the keywords are spoken by the lecturer in the scene. We have defined $phd_{/s}(s,k)$ as $phd(s,k)$ considering this condition:

$$phd_{/s}(s,k) = \begin{cases} phd(s,k) & skc(s,k)\neq 0\\ 0 & skc(s,k)=0\end{cases}$$

We replace $phd(s,k)$ by $phd_{/s}(s,k)$ in $I_{c[d+phd]}$, and add $skc(s,k)$. We denote it as $I_{c[d+phd/s]}$:

$$I_{c[d+phd_{/s},p+skc_{/p}]}(\Phi,\omega_d,\psi) = \sum_{\gamma=s-\delta}^{s+\delta} E(\gamma-s,\varepsilon_1,\varepsilon_2)\cdot I_{d[d+phd_{/s},p+skc_{/p}]}(\gamma,k,\theta,\omega_d,\psi)$$

$$I_{d[d+phd_{/s},p+skc_{/p}]}(s,k,\theta,\omega_d,\psi) = \{T(s)+\omega_d\cdot phd_{/s}(s,k)\}^{\theta}\cdot\{I_p(s,k)+\psi\cdot skc_{/p}(s,k)\},$$

Since a keyword's occurrences in slides and in speech are independent, we have proposed the weighting factor that considers both of the two filtering methods. We have defined an indicator as $phd_{/ps}(s,k)$

5 Experimental Evaluation

5.1 Experimentation Setting and Data

We evaluate our proposed methods to exploit speech and pointing information using a series of videos from two actual lecture courses on databases and on computer architecture respectively.

For the experimentations, we run 124 retrievals using different sets of keywords. These keywords consist of 78 keywords about computer architecture and 46 keywords about databases. Testers decide only one relevant scene as the answer scene corresponding the keywords.

We use the open-source speech recognition software Julius [1] to derive the speech information from the lecture videos. As the language and acoustic model for the speech recognition, we use a general model generated using CSJ[8] in [15]. We add some words from the lecture slides into the dictionary for speech recognition, which has not been originally included in it.

As the evaluation measure, we use mean reciprocal rank (MRR). MRR is commonly used for evaluation of the question-answering system such as TREC[14]. The definition of MRR is as follows:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{ranking by the i-th retrieval}}$$

where N is the number we retrieve.

5.2 Evaluation of Speech Information Integration

To evaluate the integration method, we calculate the MRR of $I_{c+skc/p}$, $I_{c+skc/c}$, $I_{c[p+skc/p]}$, $I_{c[p+skc/c]}$ varying ψ per 1 point. We fixed the parameters: $\theta = 0.5$, $\delta = 4$, $\varepsilon_1 = 5.0$, $\varepsilon_2 = 0.5$, $\lambda = course$.

Figure 2 shows the comparison of the four indicators varying ψ. When $\psi = 0$, each indicator equals I_c and MRR is 0.560 which is not considering speech information. The peak MRR of each method is: $I_{c+skc/p}$=0.595, $I_{c+skc/c}$=0.592, $I_{c[p+skc/p]}$=0.619 $I_{c[p+skc/c]}$=0.610. This result shows that the consideration of speech information is positive effect on scene

[1] http://julius.sourceforge.jp/

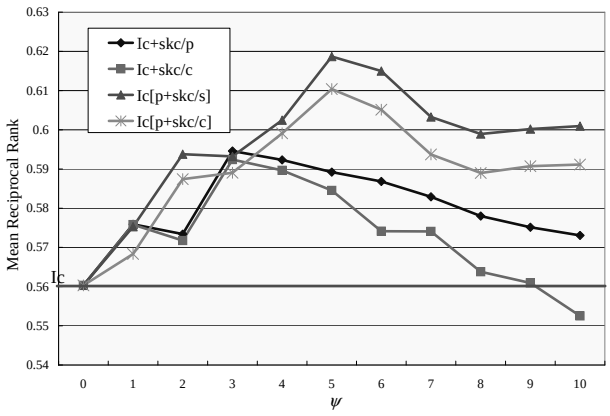

Figure 2. Comparison of four indicators varying ψ per 1 point

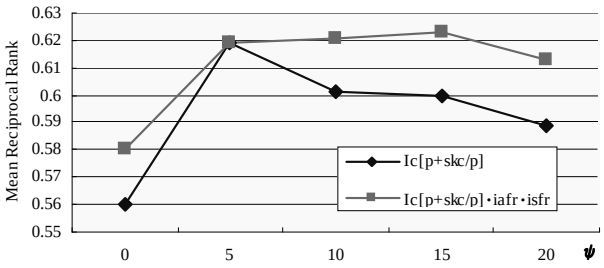

Figure 3. Comparison of impression indicators considering the rareness of a term

retrieval and the influence of speech information of neighbor scenes is useful.

To evaluate the rareness factor of speech information, we compare $I_{c[p+skc/p]}$ to $I_{c[p+skc/p]}\cdot iafr\cdot isfr$. Figure 3 shows MRR of $I_{c[p+skc/p]}$ and $I_{c[p+skc/p]}\cdot iafr\cdot isfr$ varying ψ.

An indicator considering rareness factor, "$I_{c[p+skc/p]}\cdot iafr\cdot isfr$" improves the MRR and its peak value is 0.623($\psi = 15$). This result shows the consideration rareness factor of speech information is effective for scene retrieval.

5.3 Evaluate of Pointing Information Integration

We evaluate our proposed methods to filter the laser pointer information using actual lecture materials. First we describe the setting and data used in our experiments, then we show the experimental results and discuss characteristics of the lectures.

We fix the parameters: $\theta = 0.4$, $\delta = 4$, $\varepsilon_1 = 5.0$, $\varepsilon_2 = 0.5$ and $\psi = 1$. We also distributed the probabilities of a hit by the pointer $H(l,q)$ to five neighborhood lines as 0.4, 0.3, 0.15, 0.1, and 0.05.

We compare our proposed weighting schemata, $I_{c[d+phd_{/p}]}$, $I_{c[d+phd_{/s},p+skc_{/p}]}$, $I_{c[d+phd_{/ps},p+skc_{/p}]}$, to the existing methods, I_c and $I_{c[d+phd]}$.

Figure 4 shows the MRR of the two lectures. The experimental results in Fig. 4 indicate that the proposed weighting schemata are more precise than I_c and $I_{c[d+phd]}$. The MRR decreases in $I_{c[d+phd]}$ when ω_d increases excessively, but not in our proposed methods. This means that the influence

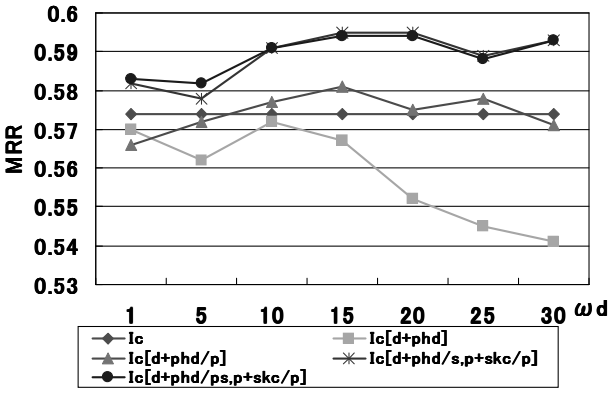

Figure 4. Comparison of MRR of the two Lectures

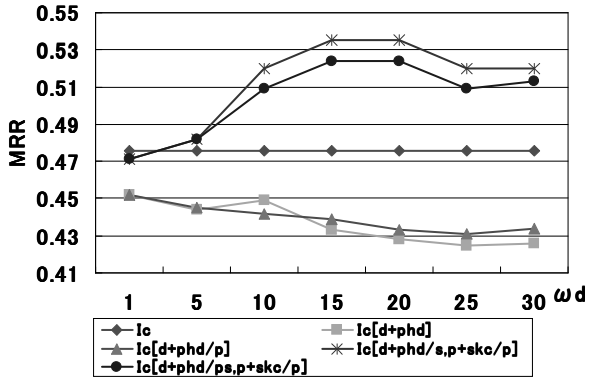

Figure 5. Comparison of MRR of the Lecture about Database

of irrelevant pointing grows larger as ω_d steadily enlarging, whereas the proposed methods moderate the influence of the laser pointer information.

Figure 5 and 6 illustrate the MRR on each lecture. According to Fig. 5, the laser pointer information does not affect positively on the scene retrieval by the existing methods. This is because the lecture has construction of slides and topics that is not suitable for ranking based on the information of the text on the slides. However, this graph shows that the laser pointer information filtered by speech information improve the MRR when parameter ω_d increases. As a result, we can moderate the effect of the laser pointer for purposes expect emphasis keywords.

In contrast, Fig. 6 shows that the laser pointer information in the existing methods favorably influence the outcome of scene retrieval, and the weighting schema filtered by text in the slide gets a best deal. This means that this lecture's characteristics differ to the former's, and is suitable for methods based on slide construction.

5.4 Analysis of Lecture Characteristics

As characteristics of the lectures are various, and effective information for ranking scenes differ by lectures, we analyze some lecture characteristics of slide materials so as to examine the influence of the characteristics on scene retrieval. First, we suppose that the structural differences between lectures af-

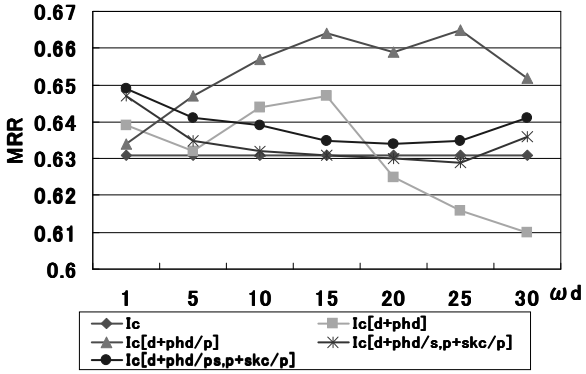

Figure 6. Comparison of MRR of the Lecture about Computer Architecture

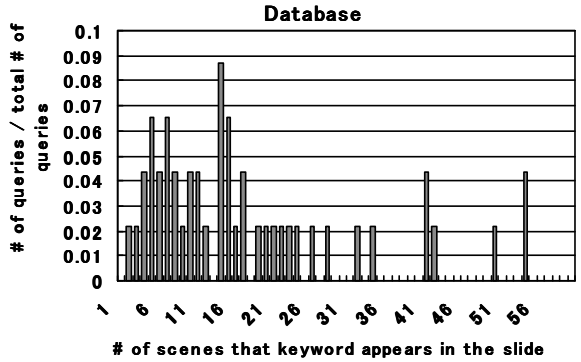

Figure 7. Distribution of the Search Keywords of the Database Lecture

fect the ranking scenes. For example, the lecture of database includes some exercise scenes. Exercises scenes usually have long duration because students need some time to solve problems, and lecturers sometimes use a slide for exercises that a particular keyword appears many times. These distinctions make scene retrievals difficult.

We also suppose that the difference of keyword distribution influences the difficulty of ranking scenes. Figure 7 and 8 show the distribution of the search keywords using this experiments. Most keywords of the computer architecture lecture appear in a few particular scenes in contrast to some keywords of the database lecture which appear in more than 20 scenes.

Table 1 shows the average number of scenes that a search keyword appears in the slide for each lecture, and standard deviation of it. This result means that the keywords of the database lecture distribute twice more widely than the computer architecture lecture. That is to say, the scene retrieval on the database lecture is more difficult than on the computer architecture lecture.

Table 1. Search Keywords Distribution on each Lecture

	average # of scenes	standard deviation
Database	17.9	13.8
Architecture	8.2	6.6

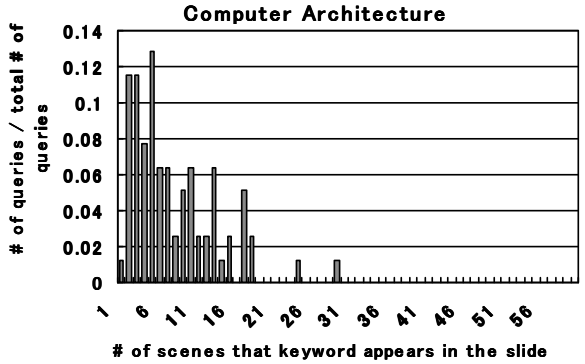

Figure 8. Distribution of the Search Keywords of the Computer Architecture Lecture

6 Conclusions and Future Work

In this paper, we discussed how to exploit various features in lecture videos such as the laser pointer and speech information for the efficient retrieval method. We introduced our information integration approaches. We explained our weighting schemata with speech information considering the criteria of appearance of the keyword in each slide and how influence of speech information spreads to neighbor scenes. We also explained our weighting schemata by using the filtered laser pointer based on keyword occurrences in slides and speech. Moreover, we evaluated our approach by using actual lecture videos and presentation slides.

Our experimental results using actual lecture materials indicate that the consideration of speech information and its rareness factor can improve the precision of retrieval and the dedicated weighting schemata for filtered laser pointer information are also effective for scene retrieval. We also evaluated on each lecture, confirmed that the influence of irrelevant laser pointer information was moderated on scene retrieval, and analyzed difference of each lecture's characteristics for each information.

As future direction to this work, we should build up methods that apply the laser pointer information and the speech information comprehensively and more effectively. Moreover, consideration of the influence of the speech recognition rate is also an important issue for the integration. We have to discuss about several problems when we use the speech information, for example, recognition error, and notation difference between dictionaries used in speech recognition and in text in slides.

Acknowledgment

This work is partially supported by a Grant-in-Aid for Scientific Research of MEXT Japan(#15017233, 16016232, 18049026), Tokyo Institute of Technology 21COE Program "Framework for Systematization and Application of Large-Scale Knowledge Resources", and CREST of JST(Japan Science and Technology Agency).

References

[1] G. D. Abowd. Classroom 2000: an experiment with the instrumentation of a living edu cational environment. *IBM Syst. J.*, 38(4):508–530, 1999.

[2] A. Fujii, K. Itou, and T. Ishikawa. Lodem: A system for on-demand video lectures. *Speech Communication*, 48(5):516–531, May 2006.

[3] A. G. Hauptmann and M. J. Witbrock. Informedia: news-on-demand multimedia information acquisition and retrieval. In M. T. Maybury, editor, *Intelligent multimedia information retrieval*, pages 215–239. MIT Press, 1997.

[4] G. J. Jones and R. J. Edens. Automated alignment and annotation of audio-visual presentations. In *Proc. Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries(ECDL'02)*, pages 276–291, Rome, Italy, Sep. 2002.

[5] Y. Kambayashi, K. Katayama, Y. Kamiya, and O. Kagawa. Index generation and advanced search functions for multimedia presentation material. In *Proc. of ER97 Workshop on Conceptual Modeling in Multimedia Information Seeking*, 1997.

[6] Y. Katsuyama, N. Ozawa, J. Sun, H. Takebe, T. Kobayashi, H. Yokota, and S. Naoi. A new solution for extracting laser pointer information from lecture videos. In *Proc. of E-learn2004*, pages 2713–2718, Washinton DC, USA, Nov. 2004.

[7] T. Kobayashi, T. Muraki, S. Naoi, and H. Yokota. An implementation of experimental system for storing and retrieving unified presentation contents (in japanese). *IEICE Transactions on Information and Systems*, J88-D-I(3):715–726, Mar. 2005.

[8] K. Maekawa, H. Koiso, S. Furui, and H. Isahara. Spontaneous speech corpus of japanese. In *Proc. LREC2000*, volume 2, pages 947–952, Athens, Greece, May 2000.

[9] O. Marques and B. Furht. *Content-Based Image and Video Retrieval*. Kluwer, 2000.

[10] R. Müller and T. Ottmann. The "Authoring on the Fly" system for automated recording and replay of (tele)presentations. *Multimedia Syst.*, 8(3):158–176, 2000.

[11] W. Nakano, T. Kobayashi, Y. Katsuyama, S. Naoi, and H. Yokota. Treatment of laser pointer and speech information in lecture scene retrieval. In *Proc. Eighth IEEE Intl Symp. on Multimedia*, pages 927–932, San Diego, USA, Dec. 2006.

[12] W. Nakano, Y. Ochi, T. Kobayashi, Y. Katsuyama, S. Naoi, and H. Yokota. Unified presentation contents retrieval using laser pointer information. In *Proc. of SWOD2005*, pages 170–173, Apr. 2005.

[13] H. Okamoto, W. Nakano, T. Kobayashi, S. Naoi, H. Yokota, K. Iwano, and S. Furui. Presentation-content retrieval integrated with the speech information. *IEICE Transactions on Information and Systems*, 2007. (to appear).

[14] E. M. Voorhees and D. M. Tice. The trec-8 question answering track evaluation. In *Proc. of TREC-8*, pages 83–105, 1999.

[15] H. Yamazaki, K. Iwano, K. Shinoda, S. Furui, and H. Yokota. Using presentation slide information for lecture speech recognition. Technical Report SP2006-122, IEICE, Dec. 2006.

[16] H. Yokota, T. Kobayashi, T. Muraki, and S. Naoi. Uprise: Unified presentation slide retrieval by impression search engine. *IEICE Transactions on Information and Systems*, E87-D(2):397–406, Feb 2004.

[17] H. Yokota, T. Kobayashi, H. Okamoto, and W. Nakano. Unified contents retrieval from an academic repository. In *Proc. of International Symposium on Large-scale Knowledge Resources LKR2006*, pages 41–46, Mar. 2006.