論文 / 著書情報
Article / Book Information

Title	Treatment of Laser Pointer and Speech Informationin Lecture Scene Retrieval
Author	Wataru NAKANO, Takashi Kobayashi, Yutaka KATSUYAMA, Satoshi NAOI, Haruo YOKOTA
Journal/Book name	Proc. of IEEE International Symposium on Multimedia (ISM2006), Vol. , No. , pp. 927-932
Issue date	2006, 12
DOI	10.1109/ISM.2006.153
URL	http://www.ieee.org/index.html
Copyright	(c)2006 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.
Note	このファイルは著者（最終）版です。 This file is author (final) version.

Treatment of Laser Pointer and Speech Information in Lecture Scene Retrieval

Wataru NAKANO[1] Takashi KOBAYASHI[2] Yutaka KATSUYAMA[3] Satoshi NAOI[3,2]
Haruo YOKOTA[2,1]

[1]Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
[2]Global Scientific Information and Computing Center, Tokyo Institute of Technology
[3]FUJITSU LABORATORIES Ltd.
{wnakano@de.cs, tkobaya@gsic}.titech.ac.jp, {katsuyama, naoi.satoshi}@jp.fujitsu.com, yokota@cs.titech.ac.jp

Abstract

We have previously 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 information in lecture scene retrieval. In this paper, we discuss the treatment of the laser pointer and speech information, and propose two methods to filter the laser pointer information using keyword occurrence in slides and speech. We also propose weighting schemata with filtered laser pointer information using slide text and speech information. We evaluate our approach by using actual lecture videos and presentation slides.

1. Introduction

Recently, systems to store and retrieve integrated multimedia contents, such as video and documents, have been proposed[1, 2, 3] and widely used in a variety of contexts, such as web-based training and e-learning. In particular, it is important for e-learning systems that users can both retrieve suitable content and find effectively the particular scenes that they wish to study.

We have proposed the UPRISE (Unified Presentation Slide Retrieval by Impression Search Engine) system[4, 5, 6]. It combines slides in a presentation and a video recording of the presentation, and retrieves a sequence of desired presentation scenes from archives of the combined content. UPRISE retrieves a relevant scene by weighting schemata that consider keyword positions in slides, duration of a scene, and context in a presentation. In our previous work[4], we showed that UPRISE's precision is much better than those of ordinary *tf-idf*-based approaches.

We have also proposed a method to reflect the influence of laser pointer activity on the weighting schemata, and we evaluated the proposed method with actual presentations[7]. The experimental results showed that the UPRISE's precision was improved by using laser pointer information. However, by analyzing the result, we found that various aspects of laser pointer pointing needed to be destinguished because several ones, such as to illustrate relationships between multiple concepts, and ambiguous pointing, had no positive effect on the scene retrieval.

In this paper, we discuss the treatment of laser pointer activity and speech information in lecture scene retrieval, and we address the problem of the influence of irrelevant pointing on retrieval. We propose two methods to filter the laser pointer information using keyword occurrence in slides and speech to moderate the influence. We also propose weighting schemata to combine filtered laser pointer information with slide text and speech information.

Some previous studies have investigated crossover retrievals for lecture contents[8, 9, 10, 11]. However, some of them do not result in actual retrieval methods and systems. [11] provided a lecture passage retrieval system using transcription by speech recognition. However, they did not consider the case of backtracking or reuse of the slide materials.

The remainder of this paper is organized as follows. In Section 2, we consider a number of weighting schemata to retrieve the unified contents. Then, we discuss the treatment of laser pointer information in lecture scene retrieval in Section 3. In Section 4, we propose a method to filter the laser pointer information, and Section 5 reports experiments and results using actual lecture materials. We summarize the paper's main points in the final section.

2. Weighting Schemata in UPRISE

In UPRISE, we modeled a lecture's contents as a sequence of scenes divided by slide changes and proposed a slide identification technique[12] for automatic contents creation.

Because our definition of a scene is the duration between slide changes, there are many scenes in which the same slide appears because of backtracking or reuse by the lecturer, as shown in Fig 1. In this case, we cannot distinguish the scenes in which the same slide appears by using only text information in the slide.

To distinguish these scenes, we have incorporated other information such as the context and duration of scenes, speech information and laser pointer information. We have proposed impression indicators as the weighting schemata in UPRISE[4].

2.1. Adding Structure Information

First, we consider the structure of a lecture slide. If a given word appears in the title of the slide or in lines less indented, the value of the position impression is high, whereas, if the keyword only appears in lines indented deep, 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 in s, $P(s.l)$ is a function of the assigned point in the line l in the slide for s, and $C(s,k,l)$ is a function of counting keywords k in the line l of the slide for scene s.

2.2. Adding with Duration Information

Duration information is useful for distinguishing multiple appearances of the same slide caused by backtracking or reuse by the lecturer. To add the duration information to the weighting schema, we propose the duration-impression indicator; its value 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. Adding Context Information

We add information about the 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 with its duration defined as a window-size parameter δ:

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

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.

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

This means that we can alter the effect of the information for its initial context and later contexts. For example, when we want to emphasize the starting point of an explanation related to the keyword, we set $\varepsilon_1 = 5$ and $\varepsilon_2 = 0.5$.

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

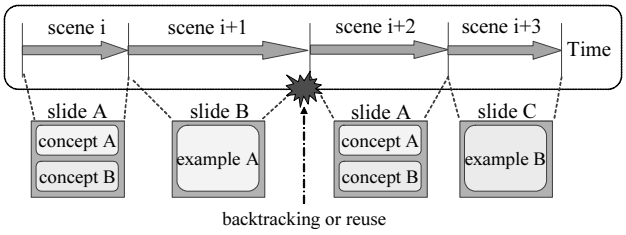

Figure 1. Example of scenes in which the same slide appears

2.4. Adding with Speech Information

We proposed a method for using the speech information in a video[6]. If a target keyword not only appears in the slide for a scene but is also frequently uttered in the scene, the scene should be ranked highly because the keyword is explained we in detail in the scene.

In [6], we extracted the speech information from a lecture video using a technique of speech recognition, and then we proposed $skc(s,k)$, the number of utterances of target keyword k in scene s. We also proposed weighting schemata to reflect the influence of speech by the combination of skc and I_c.

3. Weighting Schemata Combined with Laser Pointer Information

Lecturers usually use a laser pointer to emphasize part of the text in a slide. In other words, when a laser pointer is used in a presentation, the information about points selected

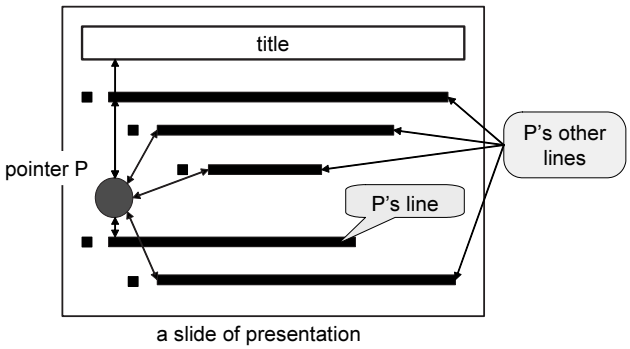

Figure 2. Extracting laser pointer information

by the laser pointer can be used to improve the precision of retrievals. We have already proposed a method of reflecting the laser pointer information on the weighting schemata[7].

In [7], we first extracted the laser pointer information by using the method of extracting radiant information from the laser pointer as coordinates in the slide by the technique of image analysis[13]. We extracted subscenes from a scene to make each subscene contain continuous pointer information. Because there is ambiguity regarding target keywords caused by shaking or habits of the lecturer, we distribute the possibilities of a hit by the pointer in a subscene to the neighborhood lines and make their sum unity:

$$\sum_{l=1}^{L(s)} H(l,q) = 1,$$

where $H(l,q)$ denotes the hit probability of line l in subscene q of scene s. We make $H(l,q)$ a function of the distance between the line l and the point, as Fig. 2.

We then create 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 in the weighting schemata, we modify Ic by adding $phd(s,k)$ to the term for the scene duration, and we denote it as $I_{c[d+phd]}$:

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

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

We evaluated the weighting schemata combined with the laser pointer information using actual lecture material in [7], and the experimental results indicate that the laser pointer information was effective in improving the precision of retrieval.

4. Filtering the Laser Pointer Information

As mentioned before, we need to consider difference in various aspects of laser pointer pointing. To moderate the effect of several ones, we filter out laser pointer information related to these aspects. In this section, we propose two methods to filter the laser pointer information based on keyword occurrence in slides and in speech. In addition, we propose a method considering both conditions together.

4.1. Filtering Based on Keyword Occurrence in Slides

When we retrieve scenes, we usually use multiple keywords in a query. However, in our previous work, we considered the influence of the laser pointer for each keyword independently. Therefore, the laser pointer information affected the ranking of scenes even if not all the keywords in a query existed in a slide.

To correct this problem, we propose a method in which the laser pointer is ignored except when all keywords in a query exist in a slide, because we suggest that lines highlighted by the laser pointer that do not contain all keywords should not add emphasis for the query. We define $phd_{/p}(s,k)$ 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 the 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:

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

4.2. Filtering Based on Keyword Occurrence in Speech

When a laser pointer hits target keywords in a slide but the keywords are not spoken by the lecturer, we can assume that the lecturer did not use the laser pointer to emphasize the keywords. To moderate the effects of such laser pointer information, we propose that the laser pointer is ignored except when the keywords are spoken by the lecturer in the

scene. We define $phd_{/s}(s,k)$ 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 $I_{c[d+phd/s]}$:

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

where ψ is the spoken-keyword-count parameter to change the effect of speech on the rating, and $skc/p(s,k)$ is the function that calculates $skc(s,k)$ only when the keyword k exists in the slide using in the scene s:

$$skc_{/p}(s,k) = \begin{cases} skc(s,k) & I_p(s,k) \neq 0 \\ 0 & I_p(s,k) = 0 \end{cases}$$

4.3. Filtering Based on Both Conditions

Because keyword occurrences in slides and in speech are independent, we propose a weighting factor that considers both. We define an indicator $phd_{/ps}(s,k)$, as follows:

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

We also propose a weighting schema $I_{c[d+phd_{/ps},p+skc_{/p}]}$ to replace $phd_{/s}(s,k)$ by $phd_{/ps}(s,k)$ in $I_{c[d+phd_{/s},p+skc_{/p}]}$.

5. Experimental Evaluation

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

5.1. Experiment Setting and Data

We evaluated our proposed methods to filter laser pointer information using a series of videos from two actual lectures: one about databases and one about computer architecture.

For the experiments, we ran 124 retrievals using different sets of keywords, 78 about computer architecture and 46 about databases. Testers selected one relevant scene as the best scene corresponding to the keywords.

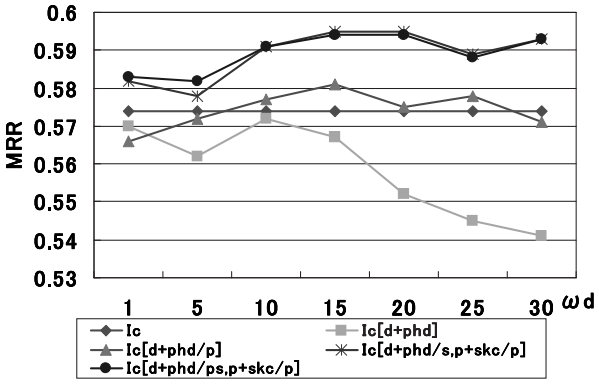

Figure 3. MRRs for the two lectures

We used the open-source speech recognition software Julius [1] to derive the speech information from the lecture videos. We added some words from the lecture slides to the dictionary for speech recognition, which were not originally included in it.

We fixed 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.

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

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

where N is the number of retrieval.

5.2. Experimental Results

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 3 shows the MRRs for the two lectures. The results 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 is made too large, but the proposed methods do not. This means that the influence of irrelevant pointing grows larger with a steadily enlarging ω_d, whereas the proposed methods moderate the influence of this laser pointer information.

Figures 4 and 5 illustrate the MRRs for each lecture. According to Fig. 4, the laser pointer information had no positive effect on the scene retrieval with the existing methods in the databases lecture. This is because the lecture's construction of slides and topics is not suitable for ranking based on

1 http://julius.sourceforge.jp/

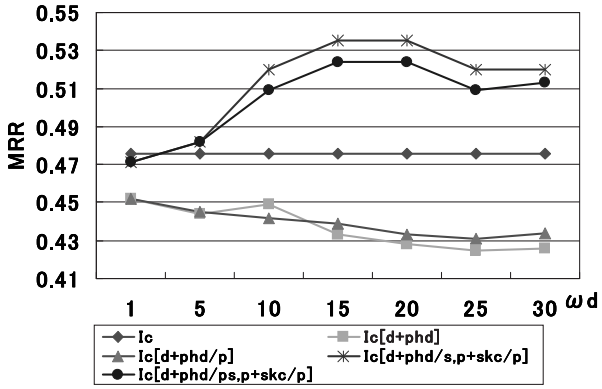

Figure 4. MRRs for the lecture about databases

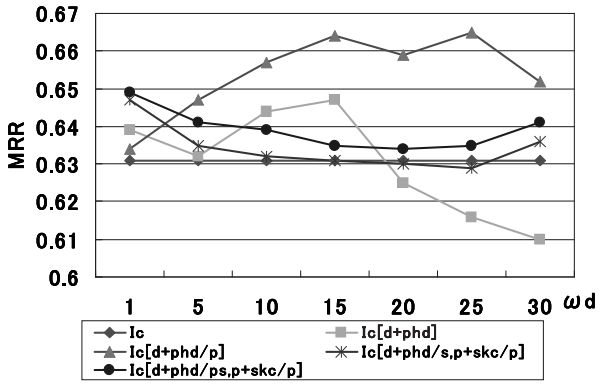

Figure 5. MRRs for the lecture about computer architecture

information about the text on the slides. However, this graph shows that the laser pointer information filtered by speech information improves the MRR when the parameter ω_d increases. As a result, we can moderate the effect of the emphasis the laser pointer adds to keywords.

In contrast, Fig. 5 shows that, with the existing methods, the laser pointer information favorably influenced the outcome of scene retrieval in the computer architecture lecture, and the weighting schema filtered by text in the slide achieves the best score. This means that this lecture's characteristics differ from the those of the first lecture, and is suitable for retrieval methods based on slide construction.

5.3. Analysis of Lecture Characteristics

As the above discussion shows, the characteristics of the two lectures differ. We analyze some lecture characteristics of slide materials to understand the influence of the characteristics on scene retrieval. First, we assume that structural differences between lectures affect the ranking of scenes.

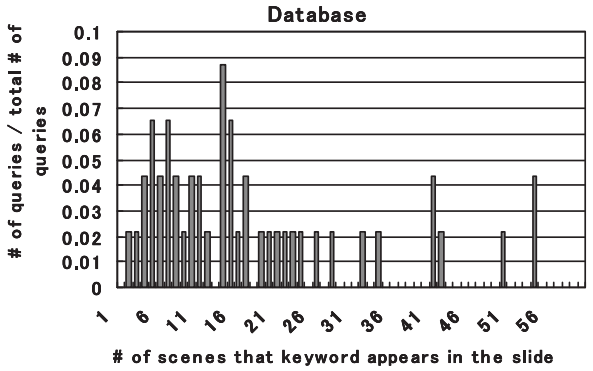

Figure 6. Distribution of the search keywords of the database lecture

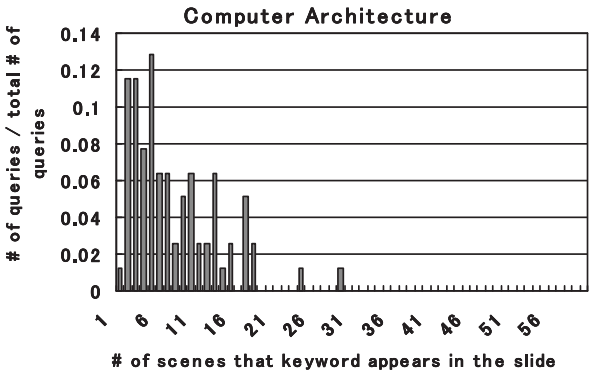

Figure 7. Distribution of the search keywords of the computer architecture lecture

For example, the lecture on database includes some exercise scenes. Exercise scenes usually have long durations because the students require some time to solve problems, and lecturers sometimes use a slide for exercises in which a particular keyword appears many times. These characteristics make scene retrievals more difficult.

We also assume that differences in keyword distribution influence the difficulty of ranking scenes. Figures 6 and 7 show the distribution of the search keywords used in these experiments. Most keywords of the computer architecture lecture appear in a few scenes in contrast to some keywords in the databases lecture, which appear in more than 20 scenes.

Table 1 shows the average number of scenes containing

Table 1. Search keywords distribution on each lecture

	Average # of scenes	Standard deviation
Database	17.9	13.8
Architecture	8.2	6.6

a search keyword in the slide for each lecture, and the standard deviation. These results show that the keywords in the database lecture are distributed twice as widely as those in the computer architecture lecture. That is, scene retrieval for the database lecture is more difficult than for the computer architecture lecture.

6. Conclusions and Future Work

In this paper, we have discussed the treatment of laser pointer and speech information in lecture scene retrieval, and propose two methods to filter the laser pointer data $phd(s,k)$ based on keyword occurrences in slides and speech. We then combine the filtered laser pointer information with the weighting schemata in UPRISE.

Our experimental results using actual lecture materials indicate that the dedicated weighting schemata for filtered laser pointer information are effective in improving the precision of retrieval. We also evaluated each lecture, confirming that the influence of irrelevant laser pointer information is moderated on scene retrieval, and we analyzed differences in each lecture's characteristics for each information type.

In this paper, we use the speech information only to filter the laser pointer information, but we have proposed that speech information is effective in retrieving scenes. Therefore, we plan to develop a method in future work that applies laser pointer and speech information comprehensively and more effectively.

We also plan to evaluate the influence of the speech recognition rate. We must consider some problems when we apply speech information; for example, recognition error and notation difference between the dictionary used in speech recognition and the text in slides. We can simulate the situation where we can correct recognized speech by using manual transcription of speech text.

The rareness factor is important for the weighting schemata in information retrieval. For *tf-idf*, the inverse document frequency, *idf*, is the rareness factor. We have reported on the effectivity of the rareness factor for scene retrieval [6].

We will consider the rareness factor of keywords in slides and speech in our proposed methods.

Moreover, we will focus on the case of users submitting synonymous keywords with text on slides and resolve the problem by, for example, using thesaurus expansion of the keywords.

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] R. Müller and T. Ottmann. The "Authoring on the Fly" system for automated recording and replay of (tele)presentations. *Multimedia Syst.*, Vol. 8, No. 3, pp. 158–176, 2000.

[2] 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*, pp. 215–239. MIT Press, 1997.

[3] G. D. Abowd. Classroom 2000: an experiment with the instrumentation of a living edu cational environment. *IBM Syst. J.*, Vol. 38, No. 4, pp. 508–530, 1999.

[4] H. Yokota, T. Kobayashi, T. Muraki, and S. Naoi. UPRISE: Unified Presentation Slide Retrieval by Impression Search Engine. *IEICE Trans. on Info. and Syst.*, Vol. E87-D, No. 2, pp. 397–406, 2 2004.

[5] T. Kobayashi, T. Muraki, S. Naoi, and H. Yokota. A Searching System on Unified Presentation Contents (in Japanese). *IEICE Trans. on Info. and Syst.*, Vol. J88-D-I, No. 3, pp. 715–726, 3 2005.

[6] 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*, pp. 41–46, 3 2006.

[7] 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*, pp. 170–173, 4 2005.

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

[9] 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.

[10] G. J. F. Jones and R. J. Edens. Automated Alignment and Annotation of Audio-Visual Presentations. In *Proc. of ECDL2002*, pp. 276–291, Sep. 2002.

[11] A. Fujii, K. Itou, and T. Ishikawa. LODEM: A system for on-demand video lectures. *Speech Communication*, Vol. 48, No. 5, pp. 516–531, 2006.

[12] N. Ozawa, H. Takebe, Y. Katsuyama, S. Naoi, and H. Yokota. Slide Identification for Lecture Movies by matching Characters and Images. In *Document Recognition and Retrieval XI*, Vol. 5296-10 of *Proc. of SPIE*, pp. 74–81, Jan 2004.

[13] 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*, pp. 2713–2718, 10 2004.

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