論文 / 著書情報
Article / Book Information

論題(和文)	講義講演シーン検索手法におけるレーザーポインタ情報と音声情報の粒度を考慮した統合
Title(English)	Considering Granularity of Laser Pointer and Speech in Information Integration for Lecture Scene Retrieval
著者(和文)	仲野亘, 小林隆志, 直井聡, 横田治夫, 古井貞熙
Authors(English)	Wataru NAKANO, Takashi KOBAYASHI, Satoshi NAOI, Haruo YOKOTA, Sadaoki
掲載誌(和文)	DEWS2007論文集
Citation(English)	Proceedings of DEWS2007
Vol, no, pages	Vol. , No. , pp. E1-3
発行日 / Pub. date	2007, 3
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2007 Institute of Electronics, Information and Communication Engineers.

講義講演シーン検索手法における レーザーポインタ情報と音声情報の粒度を考慮した統合

仲野　亘[†]　小林　隆志[††]　直井　聡[†††,††]　横田　治夫[††,†]　古井　貞熙[†]

† 東京工業大学大学院情報理工学研究科計算工学専攻　〒 152–8552 東京都目黒区大岡山 2–12–1
†† 東京工業大学学術国際情報センター　〒 152–8550 東京都目黒区大岡山 2–12–1
††† 株式会社 富士通研究所　〒 211–8588 神奈川県川崎市中原区上小田中 4–1–1
E-mail: †wnakano@de.cs.titech.ac.jp, ††tkobaya@gsic.titech.ac.jp, †††, ††naoi.satoshi@jp.fujitsu.com,
†{yokota,furui}@cs.titech.ac.jp

あらまし　我々はこれまで，講義・講演における撮影動画と発表資料をメタデータにより疎結合した統合コンテンツとして蓄積し，その特性を利用した高度なシーン検索を提供する UPRISE を提案してきた．さらに講師が用いたレーザーポインタの照射情報や講師が発話した音声情報を利用することで，シーン検索をより効果的に行う手法も提案してきた．本稿では，シーン検索精度を改善するために，レーザーポインタ情報と音声情報の粒度を考慮し，両者を高精度に対応させることで統合する手法を提案する．また，撮影動画を用い，各レーザーポインタ照射の解析を行うことで，提案手法の特性を分析し，それに基づいた改良手法の提案も行う．提案手法およびその改良手法を実際の講義コンテンツに適用して検索実験を行い，手法の有効性を確認する．

キーワード　情報統合，e-learning，情報検索，コンテンツ処理

Considering Granularity of Laser Pointer and Speech in Information Integration for Lecture Scene Retrieval

Wataru NAKANO[†], Takashi KOBAYASHI[††], Satoshi NAOI[†††,††], Haruo YOKOTA[††,†], and Sadaoki FURUI[†]

† Department of Computer Science, Graduate School of Information Science and Engineering,
Tokyo Institute of Technology
Ookayama 2–12–1, Meguro-ku, Tokyo 152–8552, Japan
†† Global Scientific Information and Computing Center, Tokyo Institute of Technology
Ookayama 2–12–1, Meguro-ku, Tokyo 152–8550, Japan
††† FUJITSU LABORATORIES LTD.
Kamikodanaka 4–1–1, Nakahara-ku, Kanagawa, 211–8588 Japan
E-mail: †wnakano@de.cs.titech.ac.jp, ††tkobaya@gsic.titech.ac.jp, †††, ††naoi.satoshi@jp.fujitsu.com,
†{yokota,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 propose a method to integrate between laser pointer and speech information by considering the granulity of both two information. We also analyze the relationship between each laser pointer shot and query word for several cases, and improve the proposed method using the result of analysis. We evalutate our approach using actual lecture videos and presentation slides.

Key words　Information Integration, e-learning, Information Retrieval, Contents Processing

1. はじめに

近年，動画や文書，音声ストリームなどの複数のメディアによるコンテンツを統合し，それらを蓄積，検索するシステムが数多く研究，および提案されており [1]〜[4]，e-Learning や講演のアーカイブ化など，様々な用途に用いられている．特に e-Learning 用のコンテンツに対しては，利用者が必要とするコンテンツを検索できるだけでなく，コンテンツのどの箇所から視聴するべきかを効果的に発見できることが重要である．

そのような検索を実現するために，我々は教育コンテンツの統合，蓄積，および統合コンテンツに対する高度な検索機能を実現するシステムである UPRISE(Unified Presentation Slide Retrieval by Impression Search Engine) を提案してきた [5].

UPRISE では，動画ストリームを資料スライドの切り替えタイミングによってシーンという単位に分割し，各シーンとそこで使用された資料スライドを対応付けることでそれらを統合する．また，各シーンに対して，対応する資料スライドの文字・構造情報，シーンの時間長の情報などから検索用インデクスを作成することで，高度な検索を可能としている．従来用いられてきたスライド中の文書検索ではなく，スライドの切り替えタイミングによってシーンを分割し，それらを検索の単位とすることで，動画中で講師がバックトラックをしたり，巻き戻りがあったりすることなどにより複数のシーンで同一のスライドが使用されている場合でも，それらを異なるシーンとして区別することができるという利点がある．

我々はこの UPRISE の検索手法において，講師が用いるレーザーポインタなどのポインティング情報に着目し，ポインティング情報を統合することで検索精度を向上させる検索手法を提案してきた [6]．また，講師の発話した音声情報を統合することで検索精度を向上させる検索手法の提案も行ってきた [7]．さらに，実際には検索キーワードに関連していないレーザーポインタ照射の情報が，シーン検索に与える影響を緩和することを目的とし，レーザーポインタ照射情報のフィルタリング手法の提案も行った [8]．これらの提案では実際の講義コンテンツを用いて検索実験を行い，それぞれの手法の有効性を示してきた．

しかし, [8] で行った提案では，レーザーポインタ情報と音声情報を異なる粒度で対応させており，2 つの情報を適切に利用することができていなかった．その結果，検索キーワードに関連していないレーザーポインタ照射の除去についても不十分であったと考える．

そこで，本稿では，検索キーワードに関連していないレーザーポインタ照射の情報をより完全に排除するために，レーザーポインタ情報と音声情報の粒度を統一することで, [8] で提案したフィルタリング手法の改良を行う．

次に，講義動画を用いて実際のレーザーポインタ照射を分析することで，レーザーポインタ照射を検索キーワードへの関連性に基づいて分類し，提案手法であるフィルタリング手法の特性を考察する．さらに，分析に基づき，提案手法の問題点を解決するために，各シーンのスライドタイトルを利用した例外処理を付与する手法を提案する．

以下では，まず 2. 節において本研究の関連研究を述べる．次に，3. 節で UPRISE の概要を示し，4. 節で動画中のレーザーポインタ情報を検索に利用するための従来手法について述べる．5. 節では，レーザーポインタ情報をフィルタリングする手法の改良方法について提案し，従来の適合度との統合を行う．また，5.2 節では，実際の講義を用い，検索実験を行って提案手法の評価を行う．その後，6. 節においてそれぞれのレーザーポインタ照射を分析し，その結果に基づいて提案手法の問題点を解決する手法の提案を行う．この手法についても同様に実験を行い，その有効性を確認する．最後に 7. 節においてまとめと今後の課題を述べる．

2. 関連研究

講義や講演などのコンテンツを扱うシステムの研究において，音声情報を利用する試みは多く行われている. [9], [10] では音声認識による音声情報を利用することを目的とした試みがなされているが，音声情報を利用した検索の提案は行っていない．

[4] は講師の発話内容を用いて教育用の動画像を検索する SEMP を提案しているが，発話内容を手作業で原稿起こししたテキストがあることを前提としており，また，検索結果である動画像の順位付けは行っていない．オンデマンド講演システム LODEM [11] は，資料テキストの情報と音声認識による講義音声を用いて，パッセージと呼ばれる発話のまとまり単位での検索を行う手法を提供している．しかし，資料スライドのバックトラックや巻き戻りへの対応や，音声の誤認識への対応は行っていない．また, [12] では対象ビデオ中のシーン情報に加え，音素インデクスファイルを生成して検索に利用し，製品化を行っているが，性能評価などの報告はなされていない．

講師が用いたレーザーポインタの情報を検索に利用した研究はほとんど行われていない. [13] では，遠隔講義システムにおける講義検索手法として，検索キーワードを含むスライド検索などの他に，講師がマウスポインタを用いてスライドのある位置を指している場面を検索する手法を提案している．しかし，これらの検索手法はそれぞれが別個のものであり，各シーンの重要度を総合的に算出するような検索は行っていない．

3. UPRISE の概要

本節では，UPRISE の概要について述べる．まず，UPRISE によるコンテンツ統合とその検索の概要を示し，次に検索に用いる基本的な適合度について述べる．そして，音声情報を統合した適合度について説明する．

3.1 UPRISE のシステム

UPRISE における，メタデータを用いたコンテンツ統合の概念図を図 1 に示す．メタデータには，動画のどの時刻にスライドの切り替えが起こったかというシーン情報と，その際にどのスライドを用いていたかという同期情報，スライドに含まれる文字列情報とその構造に対するインデクスを含める．これらの情報を保持するメタデータによってコンテンツを緩く結合することにより，個々のコンテンツが持つ情報に修正を加えることなくコンテンツの同期表示を実現し，柔軟な統合を可能にし

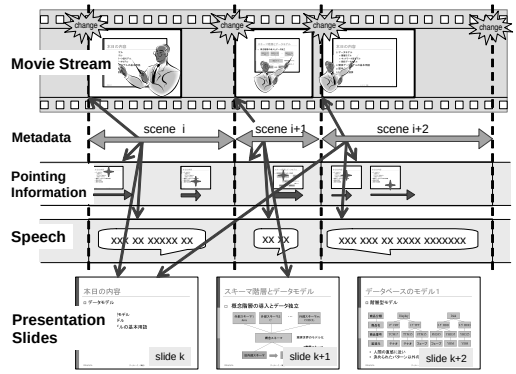

図 1　プレゼンテーション資料と動画の統合

ている．UPRISE のシステムの詳細については [14] を参照されたい．

UPRISE では，動画中に同じスライドが複数回出現する場合にそれらを異なるシーンとして区別し，個別に適合度を算出する．これにより，それぞれのプレゼンテーションは対応する動画のシーンの集合として抽象化され，プレゼンテーション中の任意のシーンが検索可能になる．

3.2　基本的な適合度

UPRISE の検索機能は，検索キーワードに対する適合度をシーンごとに算出し，上位のシーンから表示する．この適合度のうち，最も基本的なものは適合度 I_c である．I_c はスライドの文書構造，シーンの時間の長さ，前後シーンの文脈の 3 種類の情報を元に，以下の式により算出される．

以下では，まず I_c を構成する適合度 I_p および I_d について説明した後，I_c について述べる．

3.2.1　適 合 度 I_p

適合度 I_p はスライドの文書構造を考慮した適合度であり，以下の式によって定義される．

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

ここで，s はシーン，k はキーワード，l は行数であり，$P(s,l)$ はシーン s で用いられたスライドの行 l に与えられるポイント，$C(s,k,l)$ はシーン s で用いられたスライドの行 l にキーワード k が含まれる個数を表している．さらに $P(s,l)$ において行のインデントや文字の大きさに応じて重み付けをすることにより，キーワードの出現回数だけでなく出現位置も考慮することができる．

3.2.2　適 合 度 I_d

適合度 I_d は I_p にシーンの時間情報を付加した適合度であり，以下の式によって定義される．

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

ここで，$T(s)$ はシーン s の時間であり，θ は時間の影響の強弱を定めるパラメタである．これによって，説明を長く行っているシーンを重要視することができる．

3.2.3　適 合 度 I_c

適合度 I_c は I_d にシーンの前後関係を付加した適合度であり，以下の式によって定義される．

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

ここで，δ は考慮する前後シーンの範囲を定めるパラメタであり，$E(x,\varepsilon_1,\varepsilon_2)$ は前後関係の強弱を定める関数である．$E(x,\varepsilon_1,\varepsilon_2)$ は以下のように定義される．

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

適合度 I_c では，シーンの適合度はその前後 δ の範囲から影響を受け，ε が小さいほど影響を受けやすくなる．例えば $\delta = 4$, $\varepsilon_1 = 5.0$, $\varepsilon_2 = 0.5$ の時，そのシーンの適合度は前後 4 シーンの適合度に影響を受け，後に続くシーンのほうにより強い影響を受ける．

なお，この I_c のパラメータ群 $(s,k,\theta,\delta,\varepsilon_1,\varepsilon_2)$ は UPRISE における適合度関数の基本パラメタ群であるため，本稿では以降これを Φ として簡略化表現する．

3.3　音声情報を考慮した適合度

講義における講師の発話内容は，スライド中の文字列情報と同様に直接的にそのシーンの内容を表している．さらに，同一スライドを用いたシーンにおいても，発話内容によってそのシーンの差別化を行うことができる．そこで我々は音声情報を利用した適合度の提案を行ってきた [7].

[7] では，まず音声認識によって講義中の音声情報を抽出する．そして，あるシーン s 中でキーワード k が発話された回数を $skc(s,k)$ (Spoken Keyword Count) とした．この $skc(s,k)$ を適合度 I_c と統合することにより，音声情報を利用した適合度を提案した．

4.　レーザーポインタ情報を考慮した適合度

我々はこれまで，検索キーワードに対してレーザーポインタが多く当たっているシーンはそのキーワードに対してより適切であると考え，それらのレーザーポインタ情報を考慮した適合度を提案してきた [6]．本節では，レーザーポインタ情報の抽出手法と，レーザーポインタ情報を統合した適合度について説明する．また，検索キーワードに関連しないレーザーポインタ照射を排除するために，レーザーポインタ情報をフィルタリングする手法 [8] についても説明する．

4.1　レーザーポインタ情報の抽出

撮影動画から高精度でレーザーポインタの光点座標を抽出する手法 [15] を用いて，1 秒ごとにスライド中の光点座標を抽出し，スライド上で最も近い行の文字列を取得する．次に，同じ行の文字列を取得した連続の光点を 1 回のレーザーポインタ照射と定義し，一つのレーザーポインタ情報として統合する．また，この 1 回のレーザーポインタに相当する部分をサブシーンと定義する．

ここで，レーザーポインタはある 1 行に対して正確に当て続けることが容易でないため，一回のレーザーポインタに対し，近傍の数行を次候補として取得しておく．最も近い行とその付

近のいくつかの行を組にしてレーザーポインタ情報とすることで，講師の意図と光点とのぶれをある程度解消することができる．

このように抽出したレーザーポインタ情報に対し，まず，レーザーポインタの照射回数はその情報の信頼度を考慮し，キーワードが全候補行に含まれていたときに 1 とするような，回数の期待値 $H(l,q)$ として数値化する．$H(l,q)$ はサブシーン q のレーザーポインタが，行 l に照射された回数の期待値であり，全ての行の $H(l,q)$ を合計すると 1 となる．

このレーザーポインタごとの照射回数期待値に対し，各レーザーポインタが当たっていた時間を掛け合わせることで，レーザーポインタの照射時間の期待値が得られる．この照射時間の期待値をシーンごとに合計したものを，$phd(s,k)$ (Pointer Hit Duration) とする．ただし，s はシーン，k はキーワードである．シーン s，キーワード k における $phd(s,k)$ の式を以下のように定義する．

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

ここで，$T(q_i)$ はサブシーン q_i の時間を表す．すなわち，サブシーン q_i に対応するレーザーポインタの照射時間を表す．

4.2 レーザーポインタ情報と I_c の統合

4.1 節で得られた phd を適合度 I_c と統合することにより，レーザーポインタ情報を利用した適合度を提案してきた．これまでに，シーンごとの時間情報である $T(s)$ に対し，レーザーポインタの時間の期待値である $phd(s,k)$ を足し合わせ，レーザーポインタが当たっていたときにそのシーンの時間に加点するという統合手法 $I_{c[d+phd]}$ が最も有効であるということがわかっている [6]．$I_{c[d+phd]}$ は以下の式で定義される．

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

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

ここで，ω_d は phd の適合度計算における影響度合いを調節するパラメタである．例えば，$\omega_d = 10$ の場合，レーザーポインタが対象キーワードに 1 秒間当たることはそのシーンが 10 秒伸びることに相当する．

4.3 レーザーポインタ情報のフィルタリング

講義や講演で行われるレーザーポインタ照射は，あるキーワードやトピックを強調する目的以外にも，様々な目的で行われる．例えば，図や表を説明する際に補助的に用いるものや，複数概念間の関連を示す軌跡を描くもの，または照射意図がはっきりしないあいまいなものなどがある．そこで，このような様々な照射の中から，検索キーワードに対して関連する照射だけをシーン検索に利用するために，我々はレーザーポインタ情報をフィルタリングする手法の提案を行った [8].

[8] では，2 種類の条件を用いてレーザーポインタ照射のフィルタリングを行った．まず，スライドテキスト中でのキーワードの出現の有無に着目し，検索クエリが複数キーワードの場合に，その全てがスライド中に出現しない場合はそのシーンでの検索キーワードへのレーザーポインタ照射を無視するという手法を提案した．また，このフィルタリング手法を用いて計算した $phd(s,k)$ を $phd_{/p}(s,k)$ とし，以下の式で定義した．

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

ここで，K は検索語を形態素解析して得られた単語の集合である．

次に，講師の発話した音声中でのキーワードの出現の有無に着目し，検索キーワードがそのシーンで 1 度も発話されていない場合はそのシーンでの検索キーワードへのレーザーポインタ照射を無視するという手法を提案した．また，このフィルタリング手法を用いて計算した $phd(s,k)$ を $phd_{/s}(s,k)$ とし，以下の式で定義した．

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

以下，本稿では，これらのフィルタリング手法をそれぞれ P-フィルタリング，および S-フィルタリングと呼称する．

なお, [8] では，P-フィルタリングと S-フィルタリングの両方を同時に用いて計算した $phd(s,k)$ を $phd_{/ps}(s,k)$ とし，これらのフィルタリング手法を適合度 $I_{c[d+phd]}$ に統合して評価した結果，$phd_{/ps}(s,k)$ を統合した適合度 $I_{c[d+phd_{/ps},p+skc_{/p}]}$ が最も優れているという結果を得た．

5. S-フィルタリングの改良

4.3 節で述べた S-フィルタリング手法は，レーザーポインタ情報と音声情報を異なる粒度で対応させている．そのため，粒度を統一し，高精度に対応させることで，より一層の効果が望めると考える．以下では，検索キーワードに関連しないレーザーポインタ照射をより多く排除するために，S-フィルタリング手法を改良した NS-フィルタリング手法を提案する．そして，実際の講義コンテンツを用いた実験により，NS-フィルタリング手法を用いた適合度計算手法と従来のフィルタリングを用いた手法を比較する．

5.1 NS-フィルタリング

4.3 節で述べた S-フィルタリング手法では，各照射をフィルタリングするかどうかはそのシーンの発話中におけるキーワードの有無により決定されていた．しかしこの手法では，例えば，レーザーポインタ照射の時刻とキーワードの発話の時刻が数分離れていたとしても，それらが同じシーンに属するものであれば，音声中出現条件を満たし，有効な照射としてシーン検索に利用される．これは，レーザーポインタ情報をサブシーンという粒度で扱うことに対し，音声情報はシーンという粒度で扱っていたことが原因であると考える．

そこで，本研究では，音声情報をサブシーンの粒度で扱うことで，2 つの情報の粒度を統一し，より高精度に対応させることを考える．そのために，レーザーポインタ照射中，およびその前後 r_1, r_2 秒においてキーワードが発話されていない場合，その照射をフィルタする，というフィルタリング手法を提案す

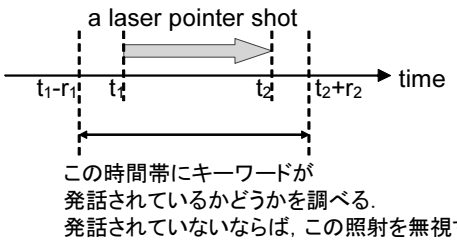

図 2　改良した S-フィルタリング手法

る．この手法により，実際に発話と同時に行われたレーザーポインタ照射と，それ以外の照射を区別することができ，前者のレーザーポインタ照射のみをシーン検索に利用することができる．図 2 はこの手法の処理を示した図である．

このフィルタリング手法を NS-フィルタリングとし，NS-フィルタリングを用いて計算した $phd(s,k)$ を，$phd_{/ns}(s,k)$ とする．$phd_{/ns}(s,k)$ を以下の式で定義する．

$$phd_{/ns}(s,k) = \sum_{q_i \in s} \sum_{l=1}^{L(s)} H(l,q_i) \cdot T(q_i) \cdot exist(q_i,k,r_1,r_2)$$

ただし，$exist(q_i,k,r_1,r_2)$ はサブシーン q_i 中，およびその前後 r_1, r_2 秒においてキーワード k が発話されていれば 1，されていなければ 0 の値をとる関数である．

また，NS-フィルタリングと同時に，4.3 節で述べた P-フィルタリングを用いた $phd(s,k)$ を，$phd_{/pns}(s,k)$ とする．$phd_{/pns}(s,k)$ は，K を検索語を形態素解析して得られた単語の集合とすると，以下の式で定義される．

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

このように定義した $phd_{/ns}$，$phd_{/pns}$ をそれぞれ適合度 $I_{c[d+phd]}$ に統合する．さらに, [8] で行った手法と同様に，音声情報 skc による適合度の加算分も考慮する．この適合度をそれぞれ $I_{c[d+phd_{/ns},p+skc_{/p}]}$，$I_{c[d+phd_{/pns},p+skc_{/p}]}$ とする．

5.2 実　　験

5.1 で提案した NS-フィルタリング手法の効果を検証するため，実際の講義のコンテンツを UPRISE に登録し，登録したコンテンツに対して各適合度ごとの検索実験を行った．以下ではその実験に関して説明し，実験結果に対して考察を行う．

実験では，データベースについての講義 (全 11 回)，計算機アーキテクチャについての講義 (全 12 回) をコンテンツ化し，検索対象とした．ただし，録音に問題があり，音声情報を得られなかった回は実験から除外した．

講義の音声情報は，連続音声認識ソフトウェア Julius (注1) を用い，言語モデルと音響モデルとして，山崎らが [16] で作成したもののうち，話者適応を行っていないものを用いた．

また，単語辞書に登録されていない用語は音声認識の結果に出現しない．そこで，山崎らが [16] において作成した辞書に，資料スライド中から辞書に含まれていない単語を追加したものを講義ごとに作成し，音声認識に使用した．

実験ではこれらのデータを用い，以下の条件の下で行った．

- 基本となるパラメタ Φ は $\theta = 0.4$, $\delta = 4$, $\varepsilon_1 = 5.0$, $\varepsilon_2 = 0.5$ とし，音声の影響の強弱を表すパラメタ ψ は 1 に固定した．
- レーザーポインタの光点に対し 5 つの候補行を取得し，照射回数期待値 $H(l,q)$ を第 1 候補から順に 0.4，0.3，0.15，0.10，0.05 という値に設定した．
- 各適合度ごとに 92 種類のキーワードを検索した．なお，計算機アーキテクチャ関連のキーワードが 59 種類，データベース関連のキーワードが 33 種類である．
- 各適合度に対して，phd の影響の強弱を表すパラメタ ω_d を 1 から 30 まで 5 刻みに変更し，計 7 回の計測を行った．
- 検索対象範囲はキーワードの正解シーンの含まれる講義ごととした．
- キーワードに対して最もよく解説していると判断したシーンをそのキーワードの正解シーンとし，適合度ごとに，正解シーンが何番目に順序付けされたかを記録した．

また，音声認識用辞書と資料スライド間の表記揺れや，音声認識と UPRISE のコンテンツ登録時に異なる形態素解析エンジンを用いていることによる影響を緩和するため，以下の処理を行った．

- 英語表記された検索キーワードを除外
- 全角英数字を半角に置換
- 検索キーワードとして用いた専門語の表記を資料スライドのものに統一

評価に際しては，正解シーンを各キーワードに対して 1 つとしていることから，平均逆数順位 (Mean reciprocal rank:MRR) を用いた．MRR は質問応答システムの評価に用いられることが多く [17]，質問ごとに最初に出現した正解の順位の逆数を求め，それらを全質問にわたって平均することで定義される．

本実験の MRR は，N を検索回数とすると以下の式で求めることができる．

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{i 番目の検索での表示順位}}$$

5.3 実 験 結 果

提案手法である $I_{c[d+phd_{/ns},p+skc_{/p}]}$，$I_{c[d+phd_{/pns},p+skc_{/p}]}$ の適合度と，従来のフィルタリング手法を用いた適合度 $I_{c[d+phd_{/s},p+skc_{/p}]}$，$I_{c[d+phd_{/ps},p+skc_{/p}]}$，およびレーザーポインタ情報と音声情報を考慮しない適合度 I_c の比較を行った．なお，NS-フィルタリングのパラメタ r_1, r_2 の値は $\{r_1, r_2\} = \{5, 5\}$ とした．

図 3 は，2 講義の検索キーワードを用いた検索による，各適合度の MRR の変化を示したグラフである．グラフより，提案手法である NS-フィルタリングを適用した適合度は，パラメタ ω_d の値によっては従来の適合度よりも良い結果を得るが，全体として従来手法とほぼ変わらない検索精度であるということがわかる．また，S-フィルタリング，NS-フィルタリングどちらにおいても，P-フィルタリングと組み合わせた方が MRR が向

(注1)：http://julius.sourceforge.jp/

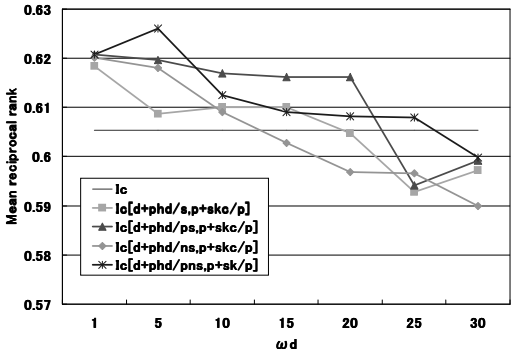

図 3　ω_d を変化させたときの適合度ごとの適合率の推移

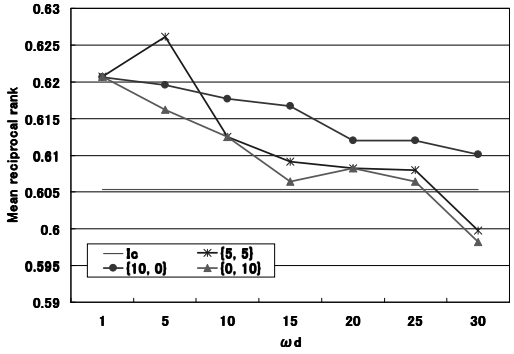

図 4　$I_{c[d+phd_{/pns},p+skc_{/p}]}$ における r_1, r_2 の適合率への影響

上することがわかる．

次に，適合度 $I_{c[d+phd_{/pns},p+skc_{/p}]}$ において，NS-フィルタリングのパラメタである r_1, r_2 の値を $\{r_1, r_2\} = \{10, 0\}$ に変更し，再度検索実験を行った．図 4 は，$\{r_1, r_2\} = \{5, 5\}, \{10, 0\}, \{0, 10\}$ の 3 種類の設定による MRR の変化を示したグラフである．

図 4 より，$\{r_1, r_2\} = \{0, 10\}$ では他の 2 種類の設定よりも MRR が低下している．$\{r_1, r_2\} = \{0, 10\}$ という設定は，レーザーポインタ照射と同時か，照射後の 10 秒間に検索キーワードが発話されているような照射のみをシーン検索に利用する設定である．このことから，今回の実験に用いた講義においては，キーワードの発話の前に行われたレーザーポインタ照射よりも，発話の後に行われたレーザーポインタ照射の方が，より多く検索キーワードに関連している照射であったということがわかる．

6.　レーザーポインタ照射の解析

5.3 節で示した実験結果より，提案手法である NS-フィルタリングの有意な効果を確認することができなかった．そこで，S-フィルタリングと NS-フィルタリングによって実際のレーザーポインタ照射情報がどのように除去されているのかを分析することで，NS-フィルタリングの特性を調査する．以下では，まず，レーザーポインタ照射情報をその意図する行を分析することで分類し，どれだけの照射が検索キーワードに関連しているのかを調べる．そして，その分析結果を元に，NS-フィルタリングを改良する手法を提案する．

6.1　レーザーポインタ照射情報の分類

まず，P-フィルタリングと S-フィルタリングを同時に用いた $phd_{/ps}$ において，2 つのフィルタリングで除去されずに $phd_{/ps}$ の計算に利用されたレーザーポインタ照射情報の分類を行い，どれだけのレーザーポインタ照射が検索キーワードに関連しているのかを調査した．分類では，5.2 節の実験に用いた検索クエリのうち，各講義 10 クエリずつの合計 20 クエリに対し，フィルタリングで除去されなかった照射全てを講義動画で確認し，講師が照射しようとしている行を発話や軌跡から判断した．その結果を用い，各レーザーポインタ照射情報を以下の 5 種類に分類した．

表 1　レーザーポインタ照射情報の分類

DB	ps	pns(5,5)	pns(10,0)
◎	51	16	17
○	186	9	13
×	97	6	8
××	30	2	1
?	17	1	4

ARCH	ps	pns(5,5)	pns(10,0)
◎	80	48	50
○	269	58	51
×	206	50	39
××	16	1	0
?	60	19	22

- 検索キーワードを含む行に当てようとした照射 (◎)
- 検索キーワードを含まない行に当てようとしているが，検索キーワードと関連があると判断した照射 (○)
- 検索キーワードとは関連がないと判断した照射 (×)
- 実際の動画では照射は行われていない，誤認識による照射情報 (××)
- 上記 4 種類の分類をするのに判断が難しい照射 (?)

検索キーワードと関連がないと判断した照射 (×) には，複数のトピックを持つようなスライドにおいて，検索キーワードとは別のトピックに属する行を意図した照射や，テキストがアニメーションによって順次出現するようなスライドにおいて，まだキーワードに関する行が出現していない状態で行われた照射などが含まれる．

また，判断が難しいとした照射 (?) には，曖昧に行われたため，動画からは意図する行がわからない照射や，関連の有無の判断が難しい行や図への照射などが含まれる．

さらに，S-フィルタリングの代わりに，本稿の提案手法である NS-フィルタリングを行うとこれらの照射情報がどれだけ除去されるのかを調査した．なお，NS-フィルタリングのパラメタ r_1, r_2 の値として，$\{r_1, r_2\} = \{5, 5\}, \{10, 0\}$ の 2 種類の設定について調査を行った．

表 1 は各講義ごとの P-フィルタリング+ S-フィルタリングで残ったレーザーポインタ照射の分類結果，および，それらが P-フィルタリング+ NS-フィルタリングではいくつ残ったかを示す表である．

まず，P-フィルタリングと S-フィルタリングを行っただけでは，検索キーワードと関連のない多くの照射情報が除去されないままであることがわかる．ここで S-フィルタリングの代わりに NS-フィルタリングを行った際の結果を見ると，×，××に分類される検索キーワードと関連のない照射情報の大部分が除去されているが，同時に◎および○に分類される検索キーワードに関連した照射情報も多くのものが除去されてしまっていることがわかる．特に○については，非常に多くの照射が除去されてしまっている．◎に比べて○の除去数が多い理由として，レーザーポインタ照射は対象となる行の内容，またはそれに類

表 2　レーザーポインタ照射情報の分類 (例外処理含む)

DB	ps	pns(5,5)	pns(10,0)	pns(5,5) +title	pns(10,0) +title
◎	51	16	17	22	23
○	186	9	13	153	157
×	97	6	8	12	14
××	30	2	1	19	19
?	17	1	4	1	4

ARCH	ps	pns(5,5)	pns(10,0)	pns(5,5) +title	pns(10,0) +title
◎	80	48	50	54	55
○	269	58	51	202	204
×	206	50	39	58	47
××	16	1	0	5	5
?	60	19	22	32	39

似する内容を発話しながら行うことが多く，○に分類される照射では，対象となる行に検索キーワードが含まれていないために近傍の音声にキーワードが出現しないことが多い，という傾向があるためと考える．

6.2　NS-フィルタリングの改良

6.1 節の分析により，NS-フィルタリングはレーザーポインタ照射情報を過剰に除去していたことがわかった．そこで，キーワードに関連する照射をできるだけ除去しないために，特に○に分類される照射について考える．

○に分類された照射の中で多く見られた例として，各照射を分類する過程において以下の 3 種類の事例を確認した．

- 照射が意図した行はインデント構造において下位であり，その上位レベルの行にはキーワードが含まれる．意図した行は，キーワードが含まれる行を補足説明している．
- 照射が意図した行と同じ段落にキーワードを含む行があり，段落全体の主題がそのキーワードを含む行になっている．
- スライドタイトルには検索キーワードが含まれるが，本文にはキーワードは含まれない．ただし，本文全体がタイトルに含まれている検索キーワードについて説明を行っている．

本稿ではこれらのうち，スライドタイトルにのみ検索キーワードが含まれる事例に着目し，タイトルに検索キーワードが含まれないシーンは P-フィルタリング+ NS-フィルタリングを行い，検索キーワードが含まれるシーンでは例外として P-フィルタリング+ S-フィルタリングを行うという手法を提案する．この，例外処理を含めたフィルタリング手法を，S/NS-フィルタリングとする．

表 2 は，例外処理を含めた手法である S/NS-フィルタリングを行った後の照射情報の分類結果を表 1 に追加したものである．スライドタイトルの例外処理を含めることによって，○に分類されたレーザーポインタ照射を多く残すことに成功していることがわかる．特に講義 ARCH においては，NS-フィルタリングによって約 80%の○に分類される照射が除去されてしまっていたのが，S/NS-フィルタリングでは約 25%の除去に抑えることができている．

6.3　再実験とその結果

6.2 節において提案した，スライドタイトルにおける検索キーワードの出現の有無に基づいた例外処理がシーン検索に与える効果を検証するため，再度検索実験を行った．なお，実験に用いたデータおよび，各パラメタの設定は 5.2 節で行った実験と同一のものを用いた．

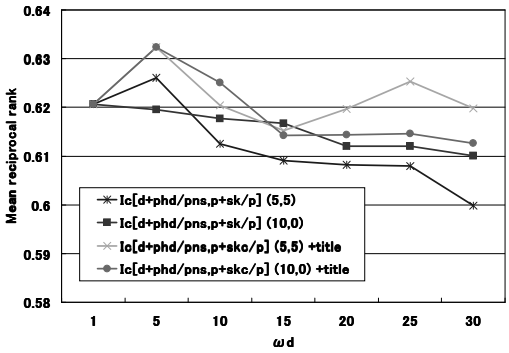

図 5　例外処理が MRR に与える影響

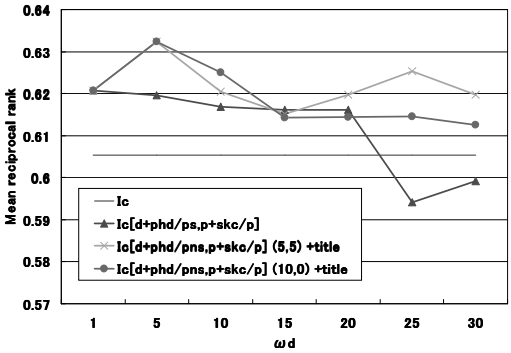

図 6　従来の適合度との比較

図 5 は適合度 $I_{c[d+phd_{/pns},p+skc_{/p}]}$ に対してスライドタイトルによる例外処理を適用したときの MRR の推移を示したグラフである．グラフより，適合度 $I_{c[d+phd_{/pns},p+skc_{/p}]}$ において，スライドタイトルによる例外処理を適用することで全体的に MRR が向上している．このことから，6.2 節の分析結果で示した，S/NS-フィルタリングによって多くの○に分類されるレーザーポインタ照射が残されるということが，実際のシーン検索に対しても良い影響を与えていることがわかる．

また，図 6 は，S/NS-フィルタリングを適用した適合度 $I_{c[d+phd_{/pns},p+skc_{/p}]}$ と，従来の適合度である I_c や $I_{c[d+phd_{/ps},p+skc_{/p}]}$ を比較したグラフである．図 6 より，従来の適合度と比較しても，MRR が改善している．これにより，提案手法である NS-フィルタリングは，その特性を補う処理を行うことによって，シーン検索精度の改善に有益な効果を与えることがわかる．

したがって，検索キーワードに関連するレーザーポインタ照射の情報のみを選別し，それ以外の照射情報を排除する精度を向上させていくことが，実際の講義コンテンツにおけるシーン検索の検索精度を向上させることにつながったと考える．

7.　まとめと今後の課題

7.1　ま　と　め

本稿では，UPRISE のシーン検索で利用するレーザーポインタ情報において，検索キーワードと関連の無いレーザーポインタ照射の情報をより多く排除するために，それぞれの照射時刻の近傍でのキーワード発話の有無を条件とした NS-フィルタリ

ング手法を提案した．

次に，従来手法によるフィルタリング結果と NS-フィルタリングによるフィルタリング結果を分析し，除去されなかったレーザーポインタ照射情報の分類を行った．その結果，NS-フィルタリングでは検索キーワードと関連のない照射の除去には成功しているものの，同時に多くの関連する照射も除去してしまっていたことがわかった．

そこで，それらの照射を除去しないための例外処理として，スライドタイトルに検索キーワードが含まれている場合は S-フィルタリングを，含まれていない場合は NS-フィルタリングを行うという，S/NS-フィルタリングを提案した．再実験の結果より，この手法は UPRISE のシーン検索精度を従来の手法より向上させることを確認した．このことより，検索キーワードに関連するレーザーポインタ照射の情報を選別する精度を向上させることが，シーン検索の検索精度向上につながることが確認できた．

7.2 今後の課題

本研究の今後の課題として，まず，レーザーポインタ照射情報の分析を本実験に用いた全ての検索クエリで行うことが必要である．さらに，より高精度に各レーザーポインタが検索キーワードに関連しているか否かを選別する手法を考案することが必要である．特に，6.1 節において○と分類されたレーザーポインタ照射を除去されないようにすることが重要であり，6.2 節で示したような○タイプの特徴を利用することで良い結果が得られると考える．

また，NS-フィルタリングのパラメタである $\{r_1, r_2\}$ の適切な値は，講義の種類，特に異なる講師の講義において変化する可能性がある．そのため，異なる講師の講義コンテンツを増やして実験を行うことは非常に有効であると考える．

その他の課題としては，音声認識精度がレーザーポインタ情報のフィルタリングに与える影響について調べる必要があると考える．音声認識精度が 100%になった条件における実験として，講義音声を人手により書き起こしたテキストを用いての実験は有効である．

さらに，提案した適合度では音声中，およびスライド文字列中での複数の検索キーワードにおける特定性を考慮していない．これらの特定性をシーン検索において考慮することは有効であり [7], [18]，今回の提案手法においてもこれらの特定性を考慮することでさらに精度が向上すると考える．

謝　　辞

本研究で用いた Julius と音響，言語モデルの使用にあたりご協力頂いた，東京工業大学大学院情報理工学研究科計算工学専攻の篠田浩一助教授，岩野公司助手，山崎裕紀氏に感謝致します．また，本研究で用いたスライド同定技術およびレーザーポインタ照射の情報抽出技術についてご協力頂いた，株式会社富士通研究所の勝山裕氏，小澤憲秋氏に感謝致します．

なお，本研究の一部は，文部科学省科学研究費補助金特定領域研究 (15017233,16016232,18049026)，独立行政法人科学技術振興機構 CREST，および東京工業大学 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の助成により行われた．

文　　献

[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] Alexander G. Hauptmann and Michael 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] Maria da Graca Pimentel, Yoshihide Ishiguro, Gregory D. Abowd, Bolot Kerimbaev, and Mark Guzdial. Supporting educational activities through dynamic web interfaces. *Interacting with Computers, special issue on interacting with the active Web*, Vol. 13, No. 3, pp. 353–374, February 2001.

[4] 森本容介, 室田真男, 清水康敬. 教育用動画像検索システムと時間情報同期方法の開発. 電子情報通信学会論文誌 D-I, Vol. J88-D-I, No. 10, pp. 1515–1524, 2005.

[5] Haruo Yokota, Takashi Kobayashi, Taichi Muraki, and Satoshi 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.

[6] 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.

[7] 岡本拓明, 仲野亘, 小林隆志, 直井聡, 横田治夫, 岩野公司, 古井貞熙. 音声情報を統合したプレゼンテーションコンテンツ検索. 電子情報通信学会論文誌, Vol. J90-D, No. 2, pp. 209–222, 2 2007.

[8] Wataru Nakano, Takashi Kobayashi, Yutaka Katsuyama, Satoshi Naoi, and Haruo Yokota. Treatment of laser pointer and speech information in lecture scene retrieval. In *Proc. of Eighth IEEE Intl. Symp. on Multimedia*, pp. 927–932, 12 2006.

[9] 中澤聡, 佐藤研治, 奥村明俊. 講演音声とプレゼンテーション資料の対応付けによる講演検索. Technical Report 情処研報 2005-SLP-55-12, 情報処理学会, 2 2005.

[10] Alessandro Vinciarelli and Jean-Marc Odobez. Application of information retrieval technologies to presentation slides. *IEEE Trans. on Multimedia*, Vol. 8, No. 5, pp. 397–406, 10 2006.

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

[12] 株式会社富士ゼロックス：MediaDEPO. http://ubiquitous-media.com/technology/index.html.

[13] 片山薫, 香川修見, 神谷泰宏, 對馬英樹, 吉廣卓哉, 上林彌彦. 遠隔教育のための柔軟な講義検索手法. 情報処理学会論文誌, Vol. 39, No. 10, pp. 2837–2845, Oct. 1998.

[14] 小林隆志, 村木太一, 直井聡, 横田治夫. 統合プレゼンテーションコンテンツ蓄積検索システムの試作. 電子情報通信学会論文誌, Vol. J88-D-I, No. 3, pp. 715–726, 3 2005.

[15] Yutaka Katsuyama, Noriaki Ozawa, Jun Sun, Hiroaki Takebe, Takashi Kobayashi, Haruo Yokota, and Satoshi Naoi. A New Solution for Extracting Laser Pointer Information from Lecture Videos. In *Proc. of E-learn2004*, pp. 2713–2718, 10 2004.

[16] 山崎裕紀, 岩野公司, 篠田浩一, 古井貞熙, 横田治夫. 講義音声認識における講義スライド情報の利用. 情処学研報, 2006-SLP-64-39, pp. 221–226, 12 2006.

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

[18] Haruo Yokota, Takashi Kobayashi, Hiroaki Okamoto, and Wataru Nakano. Unified contents retrieval from an academic repository. In *Proc. of International Symposium on Large-scale Knowledge Resources LKR2006*, pp. 41–46, 2006.