論文 / 著書情報
Article / Book Information

論題(和文)	文字認識技術を利用した講義動画のスライド同定
Title(English)	Slide Identification in the Lecture Movie Using Character Recongition
著者(和文)	武部 浩明, 小澤 憲秋, 勝山 裕, 横田治夫, 直井聡
Authors(English)	Hiroaki Takebe, Noriaki Ozawa, Yutaka Katsuyama, Haruo Yokota, satoshi naoi
出典(和文)	電子情報通信学会論文誌（D）, Vol. J91-D, No. 9, pp. 2280-2292
Citation(English)	IEICE Transactions on Information and Systems, Vol. J91-D, No. 9, pp. 2280-2292
発行日 / Pub. date	2008, 9
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) 2008 Institute of Electronics, Information and Communication Engineers.

論　文

文字認識技術を利用した講義動画のスライド同定

武部　浩明[†a)]　小澤　憲秋[†]　勝山　裕[†]　横田　治夫[††]
直井　聡[†]

Slide Identification in the Lecture Movie Using Character Recognition

Hiroaki TAKEBE[†a)], Noriaki OZAWA[†], Yutaka KATSUYAMA[†], Haruo YOKOTA[††], and Satoshi NAOI[†]

あらまし　e-Learning が普及しつつあり，複数のメディアを有機的に統合した e-Learning 向け教育コンテンツのニーズが大きくなってきた．講義動画とプレゼンテーション資料のスライドを同期させて再生する講義動画同期コンテンツを作成するためには，講義動画中のスライドが切り換わるタイミングを検出する作業が必要であり，大きなコストがかかっていた．そこで，この問題を解決するために，文字認識技術を用いて，講義動画の各フレームとプレゼンテーション資料の各スライドをオフラインで自動的に同期づける手法を提案する．具体的には，動画の各フレームを文字認識した結果とプレゼンテーション資料のスライドに含まれる文字について，同じ 2 文字間の関係を同時に満たす文字がどれくらいあるかを表す文字配置に基づく類似度を計算することによりマッチングを行い，スライドを同定する．実際の講義動画に対して本手法を適用し，スライドの同定性能を計測するとともに，本手法によって動画中のスライドが切り換わるタイミングをメタデータとしてもつ動画コンテンツを作成する作業コストを算出した．提案手法を用いることで，高精度にスライドを同定でき，作業コストを大幅に削減できることが分かった．

キーワード　e-Learning，教育コンテンツ，動画，メタデータ，文字認識

1. まえがき

情報技術の発達に伴い，教育分野でもパソコンやインターネットを利用した e-Learning が普及しつつある．e-Learning には，通常の教室で行われる授業と比較すると，学習者が，いつでも，どこでも，マイペースで学習できるというメリットがある．特に，Web ブラウザなどのインターネット・WWW 技術を使う「WBT」（Web Based Training）と呼ばれる学習形態は，社内教育などで広く利用されている．WBT では動画を利用したコンテンツがあるが，VOD（Video on Demand）のように単なる講義や講演のビデオを流すだけでは，学習者が飽きやすく学習意欲を保つことが難しいという問題がある．よって，WBT による教育効果を上げるためには，講義や講演のビデオ，講義に使用したプレゼンテーション資料，関連する論文などの資料等の多様な教育コンテンツの相互に関連する部分を統合した，学習者に興味を持たれ学習しやすい魅力的な教育コンテンツを提供することがかぎとなっている．

複数のメディアを有機的に統合した例として，講義中の講師を撮影した動画（図 1）と説明に用いたスライドを同期させて画面に表示するコンテンツの一例を図 2 に示す．このコンテンツには，動画中においてプレゼンテーション資料の各スライドが表示されていた時間を XML 形式でファイルに保持しており，この情報をもとに，画面左上に表示されている動画中におけるスライドの進行に同期して，右側のスライドページが変化する．また，左下にあるプレゼンテーション資料の各スライドへのリンクによって，任意のスライドを説明する動画を検索することが可能である．これにより，復習時にキーワードでスライドを検索し，それに対応する場面の動画を再生するなど，動画とスライ

† （株）富士通研究所，川崎市
FUJITSU LABORATORIES LTD., 1–1 Kamikodanaka, 4-chome, Nakahara-ku, Kawasaki-shi, 211–8588 Japan
†† 東京工業大学，東京都
Global Scientific Information and Computing Center, Tokyo Institue of Technology, 2–12–1 Ookayama, Meguro-ku, Tokyo, 152–8552 Japan
a) E-mail: takebe.hiroaki@jp.fujitsu.com

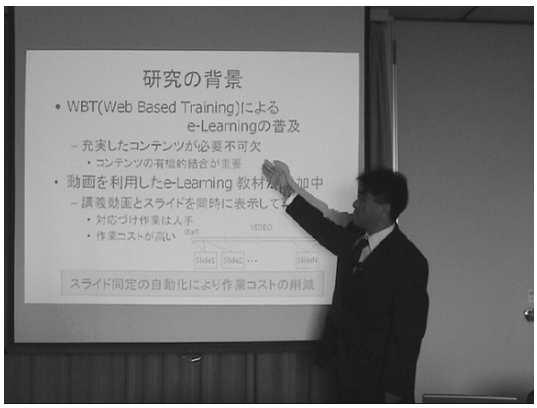

図 1　講義動画の例
Fig. 1　Example of lecture movie.

図 2　同期再生コンテンツ
Fig. 2　e-Learning content with movie and slides.

ドなどの有機的統合がなされている．

このような学習コンテンツを作成するためには，動画中でプレゼンテーション資料のスライドが切り換わるフレームを探し出しておき，メタデータとして記述し管理する必要がある．しかし，現状は，オーサリングツールを用いた手作業で，スライドが切り換わるフレームを検出し，メタデータを作成している．このスライドが切り換わるフレームを検出する作業は，動画全体をトレースし，絶え間なく見届ける必要があり，大変な作業コストがかかる．また，オンラインの同期検出では，マウスをクリックした時間を記憶しておき，その情報を利用する方法もあるが，記憶するビデオを接続した特別な計算機環境が必要なことと，限定したマウス操作によりプレゼンテーションの仕方が限られること，既存のコンテンツを扱えないこと，ビデオ撮影後の編集ができないことや統合するコンテンツの対象の拡大が難しいことなどの問題がある．

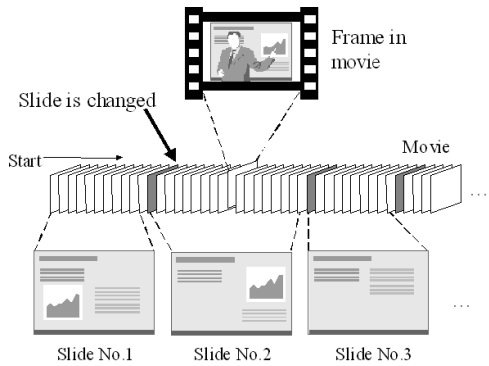

図 3　動画フレームとスライドの同定
Fig. 3　Identification of frame in movie with slides.

本論文では，この問題を解決するために，文字認識技術を用いて，講義動画の各フレームとプレゼンテーション資料の各スライドをオフラインで自動的に同期づける手法を提案する．具体的には動画の各フレームを文字認識した結果とプレゼンテーション資料のスライドに含まれる文字を比較して類似度を求めることによって，フレームがプレゼンテーション資料のどのスライドに対応するかを決定する．

本研究では，講義をビデオカメラで撮影した動画を対象とする．動画中のフレームに含まれているプレゼンテーション資料のスライドを同定することによって，動画で表示されているスライドが変化するフレーム（講師がプレゼンテーション資料のスライドを切り換える瞬間）を検出し，各スライドが動画で表示されている時間区間を自動的に特定する（図 3）．本研究に関連する研究としては，映像メディア処理の分野におけるシーンチェンジの検出 [2]〜[5] が挙げられる．動画中の連続したフレームの意味のあるまとまりは，シーン，ショット，あるいはカットと呼ばれ，シーンチェンジは動画中でフレームの内容が大きく変化することをいう．動画像処理等で用いられるシーンチェンジの検出手法は，フレーム全体の内容が大きく変化する時点をとらえるため，本研究の対象のような，フレーム中の部分領域であるスライドの領域が変化する時点をシーンチェンジとしてとらえることが難しいという問題がある．また，うまくスライドの変化をとらえることができたとしても，講義中にはスライドの順番が前後することがある．そのため，単にスライドが変化したことを検出できただけでは不十分であり，スライドの内容を確認して，どのスライドに変化したのかまでを把握する必要がある．

以下 **2.** で本論文で提案する手法を説明し，**3.** で評価実験とその結果を述べ，**4.** で考察を行い，最後に **5.** で結論と今後の課題を示す．

2. 文字ペアマッチング

講義を撮影するときは臨場感を出すために，映し出される資料のスライドだけでなく，それを説明する講師も一緒に撮影することが多い．動画中で説明しているスライドを同定するのに，プレゼンテーション資料のスライド画像をテンプレートとして画像マッチングする方法が考えられるが，動画のフレームから，映し出されているスライドの領域を特定して抽出しなければならない．しかし，フレーム中のスライドは，映し出され方によって輪郭がひずんだり，講師の体の一部が映り込んで輪郭が途切れたりする．更に，スライドの背景が黒などの濃い色で作成されているときは輪郭が出にくいなど，スライドの領域を正しく抽出することは難しい課題である．また，たいていのプレゼンテーション資料はすべてのスライドにわたって背景が同一であり，ページによって変化する部分は文字や図の内容である．しかし，異なるスライド間で画像上変化する部分は変化しない背景部分と比較すると面積は圧倒的に少ないことが多い．よって，画像情報だけからスライドを同定することが難しい場合もある．そこで，スライド内の文字情報に着目し，文字の配置がスライドの特徴を表すことに基づいて動画フレームとプレゼンテーション資料スライドをマッチングする方式を提案する．具体的には，フレームを文字認識した後，フレーム中のスライド領域の位置と大きさを推論しながら，フレームとスライドに関して同じ 2 文字間の関係を同時に満たす文字がどれくらいあるかを表す文字配置に基づく類似度を計算しマッチングを行う．

提案手法では，図 4 に示すように，動画フレームの文字認識結果とプレゼンテーション資料のスライドに含まれる文字情報を比較してマッチングを行うことによりスライド同定を行う．処理全体のフローを図 5 に示す．まず，講義の際に用いたプレゼンテーション資料のファイル内部を解析する処理によって各スライドに含まれる文字のコードと座標を抽出する．これをスライド文字情報と呼ぶことにする．次に，フレーム画像に対して文字認識処理 [7] を施し文字認識結果を得て，これと各スライドのスライド文字情報とを比較して類似度を計算しマッチングを行う．ビデオカメラなどによって撮影された画像を対象にして文字認識を

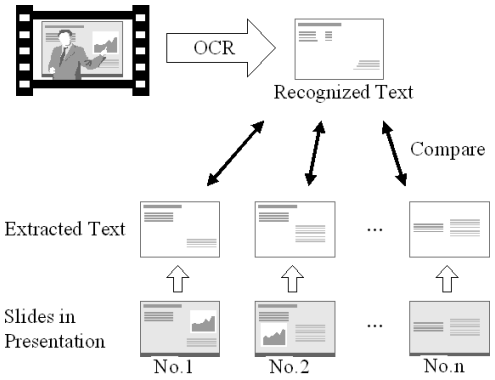

図 4　文字認識結果を用いたスライド同定
Fig. 4　Slide identification by character-based matching.

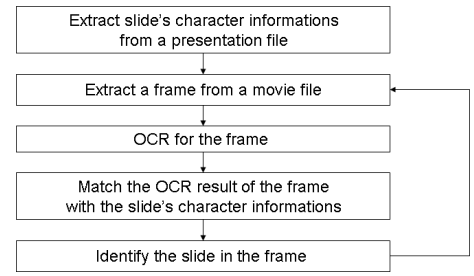

図 5　文字認識結果を用いたスライド同定の処理フロー
Fig. 5　Processing flow of slide identification by character based matching.

行う場合，解像度の低さの問題や，ノイズによる画質劣化の問題があるため，認識結果の信頼性は非常に低くなる．比較的認識しやすい環境である，テレビ映像を対象としたテロップ文字の認識でも，文字認識率は 75%程度の精度しかない [6]．更に，講義を撮影した動画の場合には，講師がスクリーンの前に出たり，影が写り込んだりすることがあるので，スライドが部分的に見えない状況が発生し，文字認識結果の一部が得られないことになる．例えば，図 6 にフレーム画像の例を示す．講師がスクリーンの前に立っているため，スライドの文字の一部が見えない状況にある．よって，このフレーム画像を文字認識したとしても，文字認識結果は図 7 のようになり，文字情報は一部欠落してしまう．したがって，認識結果がスライドの文字とすべて一致すると仮定して文字列の比較を行うことはできない．このような状況のもとでも，精度良くスライドを同定できる必要がある．そこで，認識結果とスライドのそれぞれから得られるあらゆる 2 文字間の関係の中から，互いに両立する文字配置を求めることにより

図 6　フレーム画像の例
Fig. 6　Example of frame.

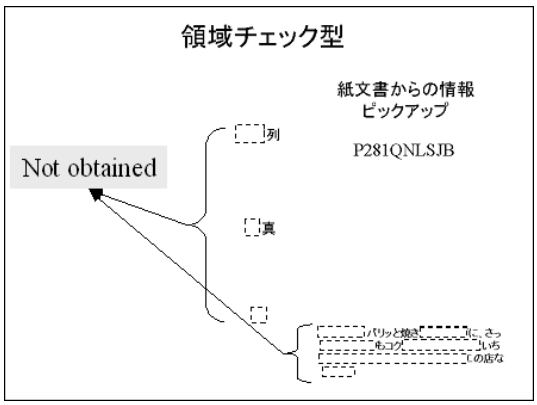

図 7　文字認識結果
Fig. 7　OCR result.

類似度を計算しマッチングを行う．具体的には，まず，認識結果とスライドの両方に含まれる文字の中で，文字コードが一致する組合せを取り出す．そして，それらの組合せの中から，2 文字間の配置関係が両立するものを抽出する．図 8 に例を示す．上がスライド文字情報を表し，下が認識結果を表している．ここで，スライド文字情報における「領域チェック型」の「領」と「ピックアップ」の「ク」の 2 文字に着目する．認識結果でも，「領域チェック型」と「ピックアップ」は正しく認識されており，「領」と「ク」の 2 文字の関係にあるのは，「領域チェック型」の「領」と「ピックアップ」の「ク」の 2 文字のほかに，「領域チェック型」の「領」と「ク」の 2 文字がある．このとき，スライド文字情報における「領域チェック型」の「領」と「ピックアップ」の「ク」の 2 文字と，認識結果における「領域チェック型」の「領」と「ピックアップ」の「ク」の

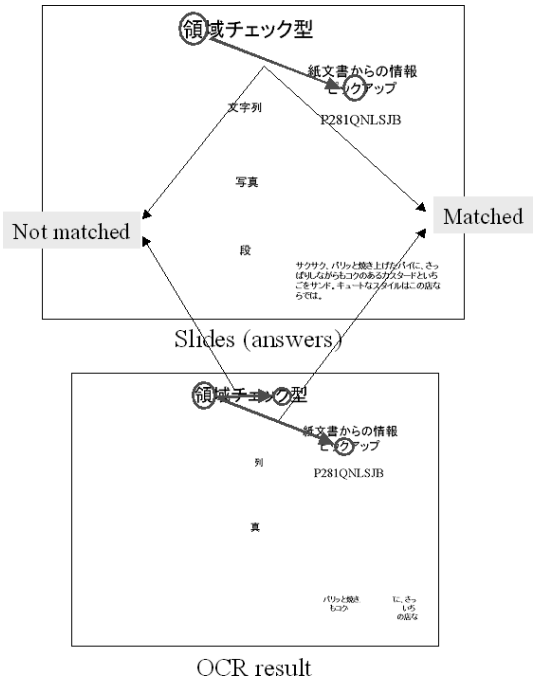

図 8　2 文字間の配置関係の両立性
Fig. 8　Positional compatibility of character pairs.

2 文字の間の関係は，平行移動と拡縮によって重なり合う，互いに相似の関係にあるため，2 文字間の配置関係が両立すると考える．一方，スライド文字情報における「領域チェック型」の「領」と「ピックアップ」の「ク」の 2 文字と，認識結果における「領域チェック型」の「領」と「ク」の 2 文字の間の関係は，互いに相似の関係にないため，2 文字間の配置関係が両立しないと考える．両立する組合せからスライド画面の拡大縮小量と平行移動量を算出し，拡大縮小量と平行移動量の最頻値を求めることにより，同じ 2 文字間の関係を同時に満たす文字の割合を求め，認識結果とスライドとの間における類似度を算出する．あらゆる 2 文字間の関係を用いてマッチングするため，認識結果が欠落していたり，誤っていたりしても，他の文字で補完することができ，スライド画面の拡大縮小量と平行移動量を頑健に推定できる．そして，フレームの認識結果に対して最も類似度が高いスライドを求めて，そのフレームに対応するスライドと同定する．以下，認識結果とスライド文字情報の文字配置に基づく類似度を求める方法を詳細に述べる．

（1）　文字集合

文字に対して，(x, y, g, p) の数値の四つ組を考える．

ここで，x，y は文字の外接矩形の中心座標，g は文字コード，p は認識結果の信頼度とする．認識結果の信頼度には，[8] において提案されている正読確率を用いる．よって，認識結果の信頼度は 0 から 1 の値をとる．正読確率の概要を以下に簡単に述べる．まず，あるパターンを文字認識した結果，第 1 位候補文字の認識距離値が f_1 であり，第 2 位候補文字の認識距離値が f_2 であるとする．このとき，認識の確信度を $f_2/(f_1+f_2)$ により定義する．次に，確信度から正読確率へ変換する．変換の仕方は多数の評価用パターンを用いて統計的に決めておく．具体的には，0.5 から 1 までを離散的に等間隔に区切っておき，評価用パターンを認識したときの認識の確信度の値が最も近い値を求め計数する．離散的に区切った値 q について，評価用パターンの中で認識の確信度が最も近いものの個数を $N(q)$ とし，また，認識の確信度が q に最も近く，かつ第 1 候補が正解であるものの個数を $N_{OK}(q)$ とする．そして，$N_{OK}(q)/N(q)$ を q に対する正読確率とし，変換テーブルとして保持しておく．認識実行時には，確信度とそれが最も近い q を計算し，変換テーブルに基づいて正読確率へ変換する．プレゼンテーション資料のスライドに含まれる文字（スライド文字情報）から得られる四つ組の集合を集合 A，認識結果の文字に対して得られる四つ組の集合を集合 B とする．ここで，プレゼンテーション資料のスライドに含まれる文字の個数を m とし，集合 A の各要素は右上の添字を用いて区別する．また，同様に，認識結果から得られる文字の個数を n とし，集合 B の各要素は右上の添字を用いて区別する．更に，集合 A に関しては，文字の信頼度 p は常に最大値をとるものとする．

$$A = \left\{ \boldsymbol{a^i} = (x, y, g, p) \;\middle|\; i = 1, 2, \ldots, m \right\} \quad (1)$$

$$B = \left\{ \boldsymbol{b^j} = (x, y, g, p) \;\middle|\; j = 1, 2, \ldots, n \right\} \quad (2)$$

（2） 文字の組合せ

集合 A と集合 B の間で文字コードが同じ組合せをすべて取り出し，それらの中で，信頼度 p があるしきい値 th_p 以上のものの集合を，集合 C とする．しきい値 th_p は，どの程度の信頼性がある文字を対象とするかによって決めるものとする．ここでは，ベクトルの各成分の値をベクトルの記号の右下に添字を付加させることによって表すことにする．

$$C = \left\{ \boldsymbol{c} = (\boldsymbol{a}, \boldsymbol{b}) \in A \times B \;\middle|\; \begin{array}{l} a_g = b_g \\ b_p \geq th_p \end{array} \right\} \quad (3)$$

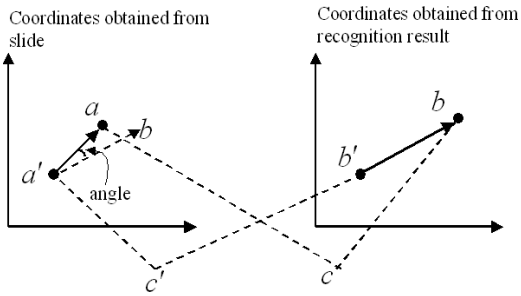

図 9　2 文字を結ぶベクトルのなす角
Fig. 9　Angle between vectors of character pairs.

（3） 2 文字間の両立性

集合 C に属する任意の二つの要素に対し，二つの要素間の関係が「両立」する組合せの集合を集合 D とする．ここで，集合 C に属する二つの要素が「両立」するとは，集合 C に属する二つの要素に対応する集合 A の 2 文字と集合 B の 2 文字の配置関係が互いに相似の関係にある，すなわち，平行移動と拡縮によって重なり合う，よって 2 文字を結ぶベクトルの画像全体に対する角度（図 9）が同じことをいう．具体的には，集合 C の要素 $c = (a, b)$ と $c' = (a', b')$ が以下の式を満たす．th_α は画像のひずみをどの程度許容するかによって決めるものとする．

$$D = \left\{ \boldsymbol{d} = (\boldsymbol{c}, \boldsymbol{c'}) \in C \times C \;\middle|\; angle(\boldsymbol{d}) \leq th_\alpha \right\} \quad (4)$$

ただし，

$$\begin{aligned} angle(\boldsymbol{d}) &= angle((\boldsymbol{c}, \boldsymbol{c'})) \\ &= angle((\boldsymbol{a}, \boldsymbol{b}), (\boldsymbol{a'}, \boldsymbol{b'})) \\ &= \left| \arctan\left(\frac{a_y - a'_y}{a_x - a'_x}\right) - \arctan\left(\frac{b_y - b'_y}{b_x - b'_x}\right) \right| \end{aligned} \quad (5)$$

（4） 拡大縮小量と平行移動量の計算

更に，集合 D に属する各要素 $d = (c, c')$（$c = (a, b)$，$c' = (a', b')$ とする．）に対し，拡大縮小量 r と平行移動量 O_x，O_y を，以下のように計算する．そして，d に対して得られる拡大縮小量と平行移動量 (r, O_x, O_y) を $\tilde{d}$ とおき，それらの集合を集合 $\tilde{D}$ とする．

$$\tilde{D} = \{\tilde{d} = (r, O_x, O_y) \mid d \in D\} \quad (6)$$

$$r = \frac{b_x - b'_x}{a_x - a'_x} \quad (7)$$

$$O_x = b_x - r a_x \quad (8)$$

$$O_y = b_y - ra_y \tag{9}$$

（5） スライドと認識結果の類似度の計算

集合 $\tilde{D}$ の各要素に対し，r，O_x，O_y に関するヒストグラムをそれぞれ計算し，最頻値 max_r，max_x，max_y を求める．拡大縮小量 r は 0 から 5 倍までを想定し，ヒストグラムでは正規化によって 0 から 100 までの整数値に換算する．次に，集合 $\tilde{D}$ の要素の中で，最頻値からある幅 th_r，th_x，th_y 以内に含まれる要素を求め，それらの集合を集合 E とする．しきい値 th_r，th_x，th_y は画像のひずみをどの程度許容するかによって決めるものとする．そして，スライドと認識結果において対応のとれた一致する文字を，集合 E の各要素に対応する集合 A と集合 B の要素とする．これらの個数を s とするとき，それぞれの文字数で割った $s/n + s/m$ をスライドと認識結果の類似度とする．

$$E = \left\{ \tilde{d} \in \tilde{D} \;\middle|\; \begin{array}{l} |\tilde{d_r} - m_r| \le th_r \\ |\tilde{d}_{O_x} - m_x| \le th_x \\ |\tilde{d}_{O_y} - m_y| \le th_y \end{array} \right\} \tag{10}$$

3. 実験と結果

3.1 スライド同定に文字情報を用いることの有効性

プレゼンテーション資料のスライドは，通常全ページにわたり背景が同一である．背景部分の面積は，文字など内容を示す部分の面積よりも大きいため，スライド同定には画像情報よりも文字情報が有効なことが多い．そこで，画像情報と文字情報がスライド同定にどの程度有効であるかを評価した．プレゼンテーション資料のスライド画像に対し，画像情報による類似度と文字情報による類似度のそれぞれによって，異なるスライド画像同士の類似度がどのような分布になるかを調べた．プレゼンテーション資料のスライドを直接 720×540 のサイズの画像にし，画像情報の類似度には，画像間の類似性を表す評価尺度として正規化相関 [9] を用いる．プレゼンテーション資料のスライドから直接文字情報を取得し，文字情報の類似度には **2.** で説明した文字配置に基づく類似度を用いる．文字の信頼度はすべて 1 とし，しきい値 th_α は 3 度，th_r は 1，th_x と th_y はスライド画像サイズの 1.2%を使用した．19 種類の講義等で用いられたプレゼンテーション資料のセット K に対し，各資料について，異なるページ同士の類似度を算出し，正規化と値の切上げによって類似度を 1 から 100 の整数の値に変換して頻度分布を求めた．結果を図 10 に示す．文字情報による類似度は類似度 1 に鋭いピークがあるが，画像情報による類似度は全体でゆるやかなカーブを描いており，文字情報による類似度が異なるスライド同士を良好に分離できることが分かる．

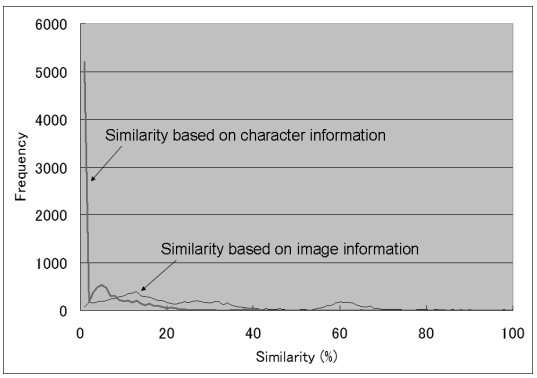

図 10 異なるスライド間の類似度の頻度分布
Fig. 10 Distribution of similarities between different slides.

3.2 提案手法の評価実験

文字ペアマッチングによる手法について，スライド同定精度を求める実験を行った．まず，文字情報が部分的に得られないときにどの程度安定的に同定が可能であるかを評価し，次に，実際の講義動画に対する評価実験を行った．しきい値 th_p は 0.1，しきい値 th_α は 15 度，th_r は 5，th_x と th_y はスライド画像サイズの 6%を用いた．

3.2.1 文字情報の不完全性に対する頑健性

文字情報が部分的に得られない状況を人工的に発生させることによって，文字ペアマッチングによる手法に関する文字情報の不完全性に対する頑健性を評価した．プレゼンテーション資料のスライドから直接文字情報を取得した後，ある割合 e の文字をランダムに決定し除去する．残った文字情報がスライドから得られる文字情報となる．これは，スライドが映っているフレームを認識率 $1 - e$ で得られる文字認識情報に近い状況である．ただし，フレームを文字認識した結果得られる文字情報には誤った文字情報も含まれるためランダムに除去して得られる文字情報よりも更に質は劣ると考えられる．ランダムに除去した文字情報を各スライドと文字ペアマッチングし，スライド同定精度を評価した．19 種類の講義等で用いられたプレゼンテーション資料セット K を対象とした．スライド数は合計 547 であるが，このうち図のみからなるスライド 21

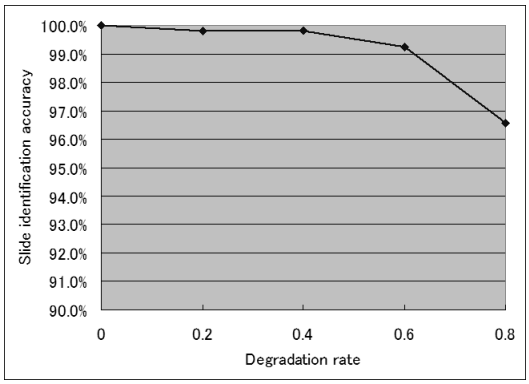

図 11　不完全な文字情報に対する文字ペアマッチングの精度
Fig. 11　Slide identification accuracy for incomplete character informations.

ははじめから除外した．また，目次のスライドで色だけが異なる複数のスライドのようにスライドは異なるが文字情報が全く同じものは，同定結果としては同一視した．r を 0，0.2，0.4，0.6，0.8 の五つの値に対して精度の変化を求めた．結果を図 11 に示す．e が 0.4 まではほぼ 100%，e が 0.6 でも 99%以上の正確さで同定できており，文字情報が不完全であっても非常に頑健に同定ができている．

3. 2. 2　講義動画に対する評価実験

文字ペアマッチングによる手法について，講義動画に対する評価実験を行った．実験には 20 種類の講義動画を用いた．講義としては 19 種類のプレゼンテーション資料セット K を用いたものであり，このうち一つの講義は前半と後半に分かれているため計 20 種類の講義動画となる．動画は，室内におけるプロジェクタを用いての発表を DV カメラで撮影したもので，24 ビットカラー，VGA（640 × 480），圧縮なしでキャプチャしたものである．撮影した動画には，カメラのズームや移動が一部含まれているものもあるが，基本的にはカメラは固定されている．また動画の構図は，講演者とスライドの両方が入る構図が多い．マッチングに用いるフレームは，動画中から 1 フレーム/秒で取り出した．プレゼンテーション資料のファイルの中には，アニメーションなどが含まれることがあるし，実際の発表には用いられなかった補足説明用のスライドが含まれていることがある．また，使用したマシン環境は，OS が Windows 2000，CPU は Xeon 2.8 GHz，メモリは 2 GB である．

動画から取り出したフレームの順番でフレームごとにスライドを同定する処理を行う．正しいスライドを選択できた場合を正解として，全フレーム数に対する正解フレーム数の割合を正解率とし，動画のフレームがどの程度正しくスライドと同定されたかを表す指標にする．実験の結果を表 1 にまとめた．表には，サンプル番号，発表資料に含まれるスライド数，正解率，処理時間を記した．動画のフレームは 1 秒当り 1 枚抽出するようにしたので，動画のフレーム数が動画の時間（秒）になる．対象となる全フレームに対して，動画時間の 2/3 の処理時間で，94.0%の正解率でフレーム中のスライドの同定に成功し，実用的に十分なレベルに到達したといえる．同定に成功したフレームの例

表 1　正解率評価結果
Table 1　Experimental results of an accuracy rate.

番号	スライド数（枚）	フレーム数（枚）	正解率（%）	処理時間（秒）
1	6	667	99.7	399
2	9	659	99.8	399
3	9	831	99.9	378
4	11	677	100.0	343
5	10	743	99.9	677
6	14	725	83.3	430
7	31	1369	63.3	761
8	25	1232	98.6	717
9	58	2337	94.2	1446
10	49	2027	95.7	1067
11	40	1305	88.0	950
12	35	1380	98.0	832
13	25	1567	99.9	1085
14	28	1204	89.5	573
15	45	1287	76.9	869
16	51	1113	98.2	546
17	22	1707	95.7	921
18	39	5528	97.2	5166
19	39	1571	99.6	1021
20	40	2845	96.2	1871
計	586	30774	94.0	20452

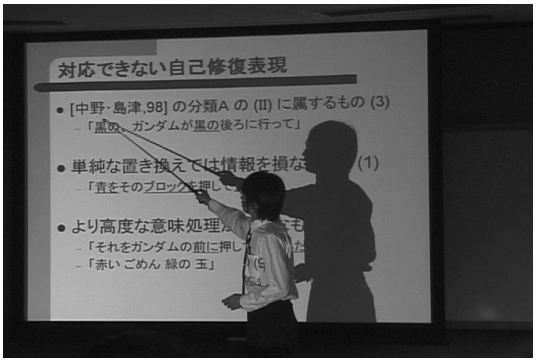

図 12　スライド同定の成功例
Fig. 12　Example of frame to be successfully idetified.

を図 12 に示す．講師が映し出されているスライドの前に立ち，講師の影がスライドを大きく遮っているが，正しく同定できた．

同定に失敗している原因を調査した結果，主に以下の四つの理由によるものであった．

（1） 誤認識による失敗

文字認識を失敗するため，一致するスライドがないか，スライド内容が類似しているものなどに誤って同定される．図 13 から図 16 に失敗例を示す．図 13 と図 15 にフレームを示し，図 14 と図 16 に文字認識用の 2 値画像とその文字認識結果を示す．文字認識用の 2 値画像は，カラー画像から図や背景模様を除外してなるべく文字のみを抽出し 2 値化したものであるが，図や背景模様の一部が残ったり，文字が一部抽出されないこともある．図 13 と図 14 は，2 値化処理がうまくいかずに文字の周辺にノイズが多数発生し，文字認識にほとんど失敗している．図 15 と図 16 は，文字認識のための 2 値化は成功し文字はほぼ正しく抽出されているが，背景の一部が文字として抽出され，文字認識処理において，中央の複数に分かれている行を一つの行と誤って認識して正しい文字認識結果がほとんど得られなかった．このように，文字認識の失敗とし

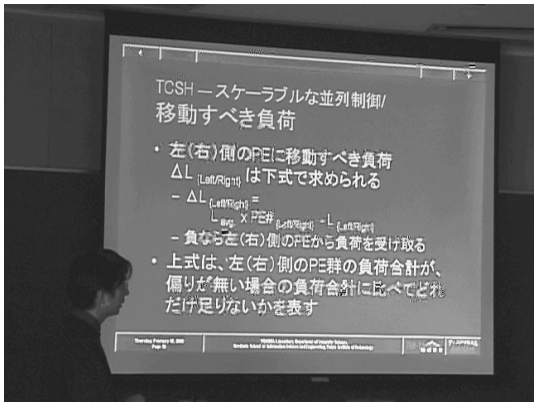

図 13　スライド同定の失敗例 1
Fig. 13　Example 1 of frame to be not idetified.

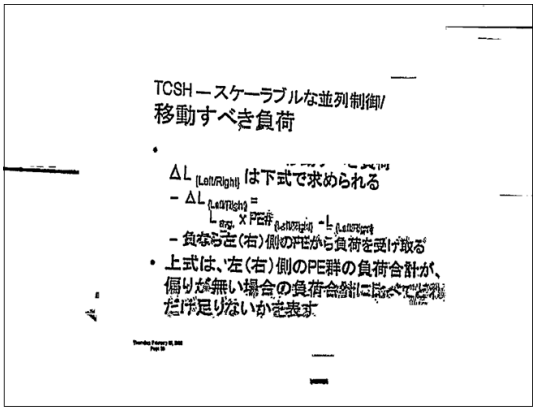

図 14　スライド同定失敗例 1 の文字認識用 2 値画像と認識結果
Fig. 14　Binarized image and recognition result of example 1.

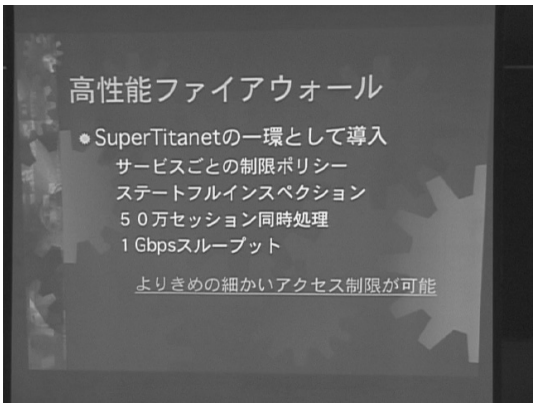

図 15　スライド同定の失敗例 2
Fig. 15　Example 2 of frame to be not idetified.

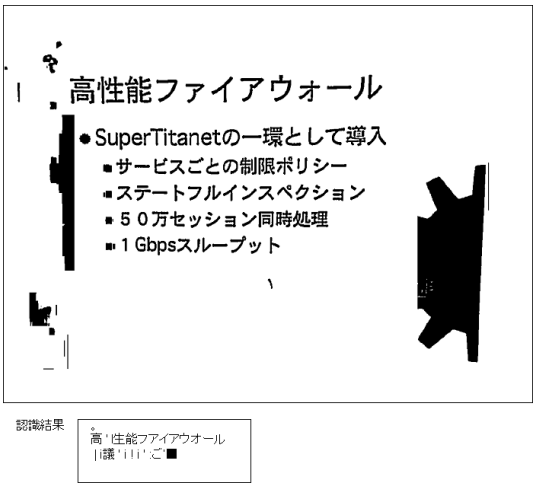

図 16　スライド同定失敗例 2 の文字認識用 2 値画像と認識結果
Fig. 16　Binarized image and recognition result of example 2.

ては，文字認識処理の入力となる 2 値画像において，文字部分と背景部分が逆転してしまったり，文字の一部が図と接触して文字認識の対象とならないなど，2 値化処理がうまくいかないことや，背景の一部がノイズとして作用することが主要な原因であった．

（2） アニメーションが使用されている場合

スライド文字情報はすべての文字が表示された最終状態での情報となる．そのため，最初に見出しだけが表示されていて，徐々に本文が現れるようなスライドの場合，同じような見出しをもつ他のスライドと間違える．

（3） テキストの内容は同じだが，色が異なる場合

テキスト部分は同じ内容であるが，発表内容の目次的な役割をするために，これから説明する部分だけが別の色で強調されているようなスライドが用いられることがある．実際には別のスライドなのだが，それらの区別を誤る．

（4） スライド中の大部分が画像や数式の場合

今回はスライドのファイルを解析してスライド文字情報を取り出しているため，画像だけが貼り付けられて作成されたスライドの場合，スライド文字情報に文字コードの情報が含まれないため，対応するスライドを見つけることができない．表示上数式などの文字が描かれているスライドであっても，実は数式部分が画像であるため，スライド文字情報には数式部分の文字情報がなく，スライドを同定できないケースがあった．

特にアニメーションによる影響が大きく，スライドの途中から正しく同定できている場合や，逆に途中から間違う場合が見られた．理由 (3) によるものは，含まれる文字を利用した同定という点では正解であると考えられる．サンプル中には，それぞれ異なる背景やスタイルが使用されていたが，今回のサンプルにおいてはそれらによる影響はあまり見られなかった．

4. 考　　察

4.1 二分探索

動画から順番に取り出したフレームごとに同定処理を行うと，動画全体の同定に処理時間がかかる．そこで，同定するフレームを動画から取り出すやり方を動画のフレーム集合に対して二分探索する方式を適用した．

二分探索方式では，異なる時点の 2 枚のフレーム F_t，F_{t+L} に対し，それらのスライドを同定させ，この異なる時点の 2 枚のフレームについて同定されたスライド番号が同じなら，その間のフレームはすべて同じ番号のスライドとみなす．スライドの番号が異なっていたら，F_t と F_{t+L} の中間時点のフレーム $F_{t+L/2}$ とフレーム F_t，及び，フレーム $F_{t+L/2}$ とフレーム F_{t+L} について，同様の処理を再帰的に繰り返していく．F_t と F_{t+L} は一定の間隔をおいて選択し，先頭フレームから最後尾フレームへ一定間隔の二分探索を繰り返し行う．

結果を表 2 にまとめた．順番に取り出したフレームごとに同定処理を行う方式（連続方式）は 94.0%の正解率であったが，二分探索方式の正解率は 92.8%であり，1.2%低下した．これは，講師が講義において表示するスライドを前のものに戻して説明したときに，二分探索方式ではずっとそのスライドが説明されていると誤ってみなしてしまうからである．しかし，処理時間は，もともとの動画の時間に対する割合で表すと，連続方式が 66.5%である一方，二分探索方式は 16.2%であり，大幅に効率化されることが分かった．

4.2 修正コスト

スライドを自動で同定させた後，スライドの切り換わりが正しいかどうかを検証し修正する作業が必要である．この修正まで含めた作業コストと，最初からスライドの切り換わりを手作業で入力したときの作業コストとの比較を考察した．

表 2　正解率評価結果
Table 2　Experimental results of an accuracy rate.

番号	スライド数（枚）	フレーム数（枚）	正解率（%）	処理時間（秒）
1	6	667	99.7	56
2	9	659	99.8	93
3	9	831	99.9	63
4	11	677	100.0	91
5	10	743	99.9	107
6	14	725	83.0	164
7	31	1369	55.8	165
8	25	1232	94.7	373
9	58	2337	95.1	624
10	49	2027	95.3	417
11	40	1305	90.1	393
12	35	1380	97.5	274
13	25	1567	98.0	287
14	28	1204	87.0	165
15	45	1287	75.8	302
16	51	1113	98.2	184
17	22	1707	95.1	201
18	39	5528	97.4	506
19	39	1571	99.6	104
20	40	2845	94.8	413
計	586	30774	92.8	4979

スライド同定結果の修正作業は以下のようにモデル化できる．スライド同定結果のスライド切り換わり点において，その前後のスライドが正しいかどうかを判定し，動画の時間順にスライド同定結果を修正していく．具体的には，検出されたスライドの切り換わり点において，以下の検証とその結果に基づく修正を行う．

(a) 1) 切り換わり点 T_k の前のスライドが正しいかどうかを判定する．次に，2) T_k が正しい切り換わり点であるか，すなわち，T_k の前後のスライドが異なるかどうかを判定する．そして，3) T_k のスライドが正しいかどうかを判定する．

(b) a-1) の判断で正しくないと判断された場合，その前の切り換わり点 T_{k-1} と T_k の間に切り換わり点が存在することになり，T_k から T_{k-1} へさかのぼりながら切り換わり点を探索し生成する．

(c) a-2) の判断で正しくないと判断された場合，T_k を削除する．

(d) a-3) の判断で正しくないと判断された場合，T_k のスライド番号を修正する．

(a) は，切り換わり点の確認動作，すなわち，切り換わり点の直前のスライド番号を確認し，切り換わり点に移動して現在のスライド番号を確認するという動作になる．(b) は，逆再生ボタンを押し，スライドの切り換わりを観察してから再生を止め，切り換わり点に移動し，現在のスライド番号を確認して，そのスライド番号を入力し，切り換わり点を生成するボタンを押すという動作になる．このコストは，切り換わり点生成コスト+動画部分の再生時間となる．(c) は，切り換わり点を削除するボタンを押すという動作になる．(d) は，現在のスライド番号を入力し，切り換わり点を更新するボタンを押すという動作になる．

コストはすべて時間（秒）に換算する．ボタンクリックコストを b-cost，番号入力コストを n-cost，切り換わり点生成コストを g-cost，切り換わり確認コストを c-cost とすると，動画全体における各コストは以下の式で計算できる．

(a) 検出された切り換わりの個数 × c-cost

(b) 再現されなかった切り換わりの個数 × g-cost + 対応付け結果が誤っている動画部分の再生時間

(c) 適合しなかった切り換わりの個数 × b-cost

(d) スライド番号を誤った切り換わりの個数 × (n-cost+b-cost)

スライド同定結果から得られるスライドの切り換わり点に関する再現率と適合率の各コストとの関係は次のようになる．(b) のコストは切り換わり点の非再現率（1-再現率）に比例し，(c) のコストは切り換わり点の非適合率（1-適合率）に比例する．(d) のコストは，切り換わり点の位置は合っているがスライド番号を誤った率に比例する．

以上のように計算される評価実験結果に対する修正コストとすべて手入力で同期対応付けを行う場合のコストの比較を行った．すべて手作業でスライドの切り換わりを入力するコストは，全再生時間+切り換わりの個数 × g-cost になる．結果を表に示す．ここで，再現率とは，時刻が正しくかつスライド番号が正しい切り換わりの再現率を指し，適合率は，時刻が正しい切り換わりの適合率を指す．b-cost は 1，n-cost は 1，g-cost は 8，c-cost は 5 として計算を行った．g-cost，c-cost に関しては，これらに含まれる，切り換わり点への移動，スライド番号の確認と入力，ボタンクリックの回数から設定した．

すべて手入力で同期対応付けを行う場合のコストを表 3 に，連続方式と二分探索方式の二つの方式によるスライド自動同定の結果を修正して同期対応付けを行う場合のコストを表 4 に示す．連続方式と二分探索方式の二つの方式を比較すると，正解率は連続方式が二分探索方式よりも高かったが，切り換わりの再現率はほぼ互角になることが分かる．また，適合率は二分

表 3　手入力による同期対応付け作業コスト
Table 3　Cost for syncronizing manually movie frames with slides.

番号	フレーム数（枚）	切り換わり回数	コスト（秒）
1	667	4	699
2	659	9	731
3	831	9	903
4	677	11	765
5	743	8	807
6	725	13	829
7	1369	29	1601
8	1232	25	1432
9	2337	45	2697
10	2027	61	2515
11	1305	34	1577
12	1380	34	1652
13	1567	26	1775
14	1204	28	1428
15	1287	24	1479
16	1113	35	1393
17	1707	22	1883
18	5528	29	5760
19	1571	8	1635
20	2845	45	3205
計	30774	499	34766

表 4　自動同期対応付けによる作業コスト
Table 4　Cost for syncronizing movie frames with slides automatically.

番号	連続方式					二分探索方式				
	再現率 (%)	適合率 (%)	処理時間 (秒)	コスト (秒)	総コスト (秒)	再現率 (%)	適合率 (%)	処理時間 (秒)	コスト (秒)	総コスト (秒)
1	100	100	399	22	421	100	100	56	22	78
2	100	100	399	46	445	100	100	93	46	139
3	100	100	378	46	424	100	100	63	46	109
4	100	100	343	55	398	100	100	91	55	146
5	100	100	677	41	718	100	100	107	41	148
6	69.2	39.1	430	282	712	76.9	61.1	164	238	402
7	41.4	43.8	761	804	1565	31.0	55.0	165	862	1027
8	92.0	95.8	717	154	871	92.0	100	373	196	568
9	68.9	62.1	1446	530	1976	71.1	73.1	624	456	1080
10	77.0	87.3	1067	476	1543	78.7	92.9	417	460	877
11	70.6	69.2	950	426	1376	67.6	72.2	393	389	782
12	91.2	91.2	832	224	1056	94.1	91.4	274	229	504
13	100	100	1085	131	1216	96.2	100	287	165	452
14	75.0	72.7	573	339	912	57.0	92.0	165	328	493
15	58.3	43.9	869	581	1450	58.3	58.1	302	536	837
16	97.1	100	546	197	743	97.1	100	184	197	381
17	90.9	95.2	921	196	1117	90.9	100	201	200	401
18	82.8	81.3	5166	349	5515	79.3	89.3	506	321	827
19	75.0	87.5	1021	58	1079	75.0	87.5	104	58	162
20	88.9	76.9	1871	419	2290	84.4	97.4	413	400	814
計	80.8	76.7	20452	5376	25828	80.0	86.4	4979	5245	10224

探索方式が大きく上回っている．その結果，コストも二分探索方式の方が小さくなる．更に，処理時間を含めた総コストでは，二分探索方式が連続方式の半分以下となっている．また，すべて手作業で行う作業コストと比較すると，連続方式は約 74%の作業コストとなり，二分探索は約 29%の作業コストとなることが分かる．これにより，二分探索方式による自動スライド同定によって同期対応付けを行えば，約 70%の作業コストを削減できることになる．

4.3　文字の誤認識による影響

誤認識した文字が文字ペアマッチング全体に与える影響を考察した．あるフレームについて，認識結果の文字数を v 個とし，このうち，認識に正解した文字数を w 個とすると，誤認識した文字数は $v-w$ 個となる．このとき，誤認識した文字が文字ペアマッチングに与える影響は，認識結果における任意の 2 文字の組のうち，誤認識した文字が含まれる組の全体に対する割合によって測られる．ここで，全体の文字の組合せに対する誤認識した文字が含まれる割合を u とすると，以下の式で計算される．

$$u = 1 - \frac{w(w-1)}{v(v-1)} \tag{11}$$

認識結果の信頼度は 0 から 1 までの値をとる．この実験では，認識結果の信頼度のしきい値には，ほとんどすべての正解を含むことになる 0.1 を用いたが，更に高いしきい値を用いて，マッチングに用いられる文字を限定していけば，誤認識した文字を減少できるため，誤読の文字が含まれる割合 u は減少していく．しかしながら，マッチングに用いられる全体の個数も減少することになるので，スライド内容が類似している他のスライドに誤って同定する危険性も高まる．マッチングで用いられる全体の個数を減少させずに，u を低める方法として，マッチングにおいて生成される文字の組に対して，2 文字に関する認識に信頼度のうち大きい方をその 2 文字の組合せの信頼度として付加し，重みづけることが考えられる．この方法では，認識の信頼度の低い文字であってもマッチングの対象とでき，認識信頼度によって重み付けされるため，誤認識した文字の影響を下げることができる．あるフレームのサンプル（文字数約 150）に対して，認識の信頼度のしきい値を，0.1 とした場合，正解文字の 100%が含まれるが，u は 0.47 であった．また，認識信頼度のしきい値を 0.7 とした場合，正解文字は 64%が含まれ，u は 0.36 であった．一方，マッチングにおいて生成される文字の組を信頼度によって重みづける方法の場合は，u が 0.34 であった．この方法は，しきい値を 0.1 とした場合と比較すると，ともにすべての正解文字を対象としているが，誤読した文字が与える影響を減少

させていることが分かる．また，しきい値を 0.7 とした場合と比較すると，誤読した文字が与える影響はほぼ同等であるが，マッチングの対象となる正解文字をより多く含む．誤認識による失敗を防ぐため，文字ペアマッチングにおいて生成される文字の組を信頼度によって重みづける方法の導入を今後検討する．

5. むすび

本論文では，e-Learning 向け講義動画同期コンテンツの作成支援を目的として，講義動画中のフレームとプレゼンテーション資料のスライドを対応づける手法を述べた．講義動画中のフレームに対する文字認識結果と，プレゼンテーション資料のスライドに含まれる文字を利用した文字ペアマッチング手法を提案した．文字ペアマッチングでは，同じ 2 文字間の関係を同時に満たす文字がどれくらいあるかを表す文字配置に基づく類似度を計算しマッチングを行うことによって，動画フレーム中から得られる文字情報が不完全な場合でも正しくマッチングすることが可能である．実際の講義動画とプレゼンテーション資料を用いた実験によって同定性能を測定した結果，良好な精度でスライドの同定を行えることが分かった．

今後の課題としては，スライドの中に文字情報が含まれていない場合への対処として，スライドのファイルに対して文字認識を行うことにより文字情報を追加して取得することと，画像マッチングの導入を検討する．画像マッチングでは，講義を撮影した動画が照明などの撮影環境によって画質が変動するし，プロジェクタや OHP を用いる発表では，機材の設置場所によって，投影されたスライドの輪郭がひずんだり画面の明るさが変化したりする．このような変動要因がある場合でも，正しくスライドを同定しなければならない．

また，アニメーションが含まれるスライドの同定精度向上に取り組む．提案した手法では，同定誤りを生じる一番の原因はアニメーションが含まれるスライドであった．プレゼンテーションではアニメーションが利用されることは少なくないため，これらに対応できる手法が必要となる．しかし，アニメーションの動きにはたくさんのバリエーションがあることに加え，変化する領域が画像に対して局所的であることが多く，その変化を正確にとらえることが難しい．また，アニメーション情報の抽出方法やスライドをアニメーションの出現場面に応じて複数枚に分けて階層的に管理するなどの対策が必要である．更に，今回の実験では 1 秒間に複数回のスライド変化が発生しないという仮定のもとに 1 フレーム/秒で画像を取り出して処理を行っていた．アニメーションの変化に対応するためには，より短い間隔で映像を取り出す必要があるので，もっと多くのフレームを処理する必要があり，処理時間の増大という問題についても対応する必要がある．

文　献

[1] 小澤憲秋，武部浩明，勝山　裕，直井　聡，横田治夫，“文字認識を利用した講義動画中のスライド同定，” FIT2002 情報技術レターズ，LI-5, Sept. 2002.

[2] L. He, E. Sanocki, A. Gupta, and J. Grudin, “Auto-summarization of audio-video presentations,” Proc. ACM Multimedia’99, pp.489–498, 1999.

[3] M.A. Smith and T. Kanade, “Video skimming and characterization through the combination of image and language understanding techniques,” CMU Tech. Rep. CMU-CS-97-111, Feb. 1997.

[4] Y. Nakamura and T. Kanade, “Semantic analysis for video contents extraction–spotting by association in news video,” ACM Multimedia, pp.393–401, 1997.

[5] 有木康雄，“DCT 特徴のクラスタリングに基づくニュース映像のカット検出と記事切り出し，” 信学論（D-II），vol.J80-D-II, no.9, pp.2421–2427, Sept. 1997.

[6] 森　稔，倉掛正治，杉村利明，塩　昭夫，鈴木　章，“背景・文字の形状特徴と動的修正識別関数を用いた映像中テロップ文字認識，” 信学論（D-II），vol.J83-D-II, no.7, pp.1658–1666, July 2000.

[7] H. Kamada and K. Fujimoto, “High-speed, high-accuracy binarization method for recognizing text in images of low spatial resolutions,” Proc. ICDAR’99, pp.139–142, 1999.

[8] 藤本克仁，鎌田　洋，“正読確率を用いた高速高精度な文字，” 1996 信学ソ大（情報・システム），D-361, Sept. 1996.

[9] 高木幹雄，下田陽久（監修），新編 画像解析ハンドブック，pp.1671–1672, 東京大学出版会，2004.

（平成 19 年 4 月 18 日受付，11 月 2 日再受付）

武部　浩明　（正員）

1992 東大・理・数学卒．1995 同大大学院数理科学研究科数理科学専攻修士課程了．同年（株）富士通研究所入社．以来，画像処理やパターン認識の研究に従事．

小澤　憲秋　（正員）

1996 東北大・工・情報卒．2001 同大大学院情報科学研究科博士後期課程了．博士（情報科学）．同年（株）富士通研究所入社．画像処理や画像認識等の研究に従事．

勝山　　裕　（正員）

1984 東北大・工・通信卒．1986 同大大学院工学研究科情報工学専攻修士課程了．同年富士通（株）入社．以来，画像処理やパターン認識の研究に従事．

横田　治夫　（正員）

1980 東工大・工・電物卒．1982 同大大学院・情報・修士課程了．同年富士通（株）入社．同年 6 月（財）新世代コンピュータ技術開発機構研究所．1986（株）富士通研究所勤務．1992 北陸先端大・情報・助教授．1998 東工大・情報理工・助教授．2001 東工大・学術国際情報センター教授．工博．主として分散インデキシング，データ工学向けアーキテクチャ，高機能ストレージシステム，ディペンダブルシステム等に関する研究に従事．日本データベース学会理事，ACM SIGMOD 日本支部評議委員，2003〜2005 本会データ工学研究専門委員会委員長，情報処理学会，人工知能学会，IEEE，ACM 各会員．

直井　　聡　（正員）

1983 慶大・工・電気卒．1985 同大大学院工学研究科電気工学専攻修士課程了．同年，（株）富士通研究所入社．以来，文字パターン処理，画像処理，文字認識，e-learning の研究に従事．現在，言語・メディア研究部部長．工博．2001 年 9 月より東京工業大学学術国際情報センター客員助教授．2005 年 6 月同大学術国際情報センター客員教授．1998/1999 年度 ISS 財務幹事．2001/2002 年度 PRMU 研究会幹事．現在，PRMU 委員．情報処理学会会員．2003 年度より JEITA 認識形入力方式標準化委員会委員．2003 年度より映像情報メディア学会メディア工学研究会幹事．