LI-5

文字認識を利用した講義動画中のスライド同定

Slide Identification in Lecture Video by Using Character Recognition

小澤憲秋† N.Ozawa†　**武部浩明**† H.Takebe†　**勝山裕**† Y.Katsuyama†　**直井聡**†‡ S.Naoi†‡　**横田治夫**‡ H.Yokota‡

1. はじめに

e-learning による学習形態は、WBT (Web Based Training)システムによる、動画を用いた学習が拡がりつつある。例えば、講義中の講師を撮影した動画と説明に用いたスライドを同時に画面に表示して、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と再生するなど、スライドや動画などのデータの有機的統合が重要である。これを実現するためには、動画中でスライドが切り替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、このような学習用コンテンツの作成は、オーサリングツールを用いての手作業に頼っているのが現状である。この作業は映像全体をトレースする必要があり、大変なコストがかかる。

本稿では、コンテンツ作成時のコスト削減を目的として、講義などを撮影した動画中における各スライドの開始時刻と終了時刻を自動的に検出する手法を提案する。具体的には動画中の各フレームを文字認識した結果とスライドのテキストを比較することによって、フレーム中にあるスライドを同定する。

2. 文字認識を用いたスライドの同定

2.1 課題と問題点

動画中の変化を検出する手法としてシーンチェンジの検出[1]が考えられるが、映像中に含まれるスライド領域の文字だけが部分的に変化した場合にはシーンチェンジとは捉えられないことがある。また、講義中にはスライドの順番が前後することがある。従って、スライドが変化したことを検出するだけではなく、どのスライドであるかを同定する必要がある。

スライド画像をテンプレートとして、画像マッチング[2]などによってフレーム中のスライドを判断する手法も考えられるが、発表で用いられるスライドは同じようなレイアウトであることが多いために、画像間の特徴の比較を行うだけでは正確な判断ができない。

従って、スライド中の内容まで判断する必要がある。そのために文字認識を用いて文字列を抽出する。認識結果が完全であれば、文字列の比較を行うことでスライドを判断できるが、映像の解像度が低くノイズも含まれるため、認識結果には誤りを生じる。テレビのニュース映像を対象としたテロップ文字認識率は 75%程度の精度しかない[3]。また、講師がスライドの前を横切るなどすると、完全な文字列が得られないなどの問題点がある。

† 株式会社 富士通研究所, Fujitsu Laboratories Ltd.
‡ 東京工業大学 学術国際情報センター,
Global Scientific Information & Computing Center,
Tokyo Institute of Technology

図 1　フレーム内のスライドの同定

2.2 提案手法

以上の課題を解決するため、文字ベースで比較を行う手法を提案する。各フレームを文字認識し、使用されている文字とその座標を抽出する。スライドの情報と比較することにより、どちらにも含まれる二文字の組の位置関係をすべて調査して、それらの関係が一致する頻度を用いてスライドを同定する。以下、その方法を述べる。

(1) 各文字に対して(*code, x, y, certainty*) の数値の組を考える。ここで、*code* は文字コード、*x, y* は文字の外接矩形の中心座標、*certainty* は認識結果の信頼度とする。スライドから得られる文字の集合を *A*、認識結果から得られる文字の集合を *B* とする。ここで、スライドから得られる情報は既知であるので、集合 *A* の *certainty* は常に最大値をとる。

$$A = \{a_i = (code, x, y, certainty) \mid i = 1,2,..., m\} \quad \textbf{(1)}$$

$$B = \{b_i = (code, x, y, certainty) \mid i = 1,2,..., n\} \quad \textbf{(2)}$$

(2) 集合 *A* と集合 *B* の中で文字コードが同じ組合せをすべて取り出し、その集合を *C* とする。この時、集合 *B* からは *certainty* がある閾値 *th* 以上の文字のみを採用する。

$$C = \{c_k = (a_i, b_j) \in A \times B \mid a_i(code) = b_j(code) \ and \ b_j(certainty) \geq th, \ i = 1,2,..., m, \ j = 1,2,..., n\} \quad \textbf{(3)}$$

(3) 集合 *C* に属する文字が「両立」する組合せの集合 *D* とする。ここで「両立」とは、*C* に属する 2 つの要素を取り出したときに、認識結果の二文字とスライド中の二文字の位置関係が同じ状態にあることをいう。具体的には以下の式を満たす。

$$D = \{d_k = (c_i, c_j) \in C \times C \mid angle(d_k) \leq th, \ i < j, \ i, j = 1,2,..., N\} \quad \textbf{(4)}$$

ただし

$$angle(d(c_1(a_1, b_1), c_2(a_2, b_2))) = \left| \tan^{-1} \frac{a_1(y) - a_2(y)}{a_1(x) - a_2(x)} - \tan^{-1} \frac{b_1(y) - b_2(y)}{b_1(x) - b_2(x)} \right| \quad \textbf{(5)}$$

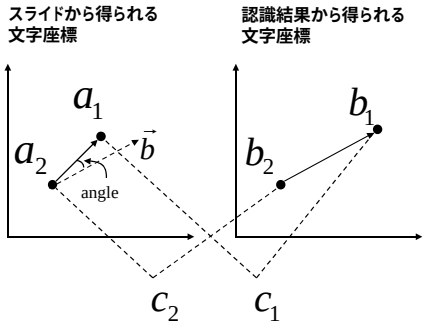

図 2　文字の「両立」(位置関係)を比較

(4)　スライド領域の拡大縮小や並行に対応するために、

$$d(c_1(a_1,b_1),c_2(a_2,b_2)) \in D \quad \textbf{(6)}$$

に対して

$$ratio(d)=\frac{b_1(x)-b_2(x)}{a_1(x)-a_2(x)},$$
$$O_x(d)=b_1(x)-ratio(d)a_1(x), \quad \textbf{(7)}$$
$$O_y(d)=b_1(y)-ratio(d)a_1(y)$$

を計算し

$$\tilde{d}=(ratio,O_x,O_y,c_1,c_2),$$
$$\tilde{D}=\{\tilde{d}\mid d\in D\} \quad \textbf{(8)}$$

とする。

(5)　$\tilde{D}$ の要素に対し、$ratio,O_x,O_y$ に関するヒストグラム h_r,h_x,h_y を作成し、それぞれの最頻値 m_r,m_x,m_y を求める。

(6)　$\tilde{D}$ の要素の中で、最頻値から $th_{\tilde{D}}$ の幅で近接する要素に属する集合 E の個数を求め、最も値の大きいスライドを選択する。

$$E=\{\tilde{d}\in\tilde{D}\mid \left|\tilde{d}(ratio)-m_r\right|\le th_{\tilde{D}} \ and \ \left|\tilde{d}(Ox)-m_x\right|\le th_{\tilde{D}} \ and \ \left|\tilde{d}(Oy)-m_y\right|\le th_{\tilde{D}}\} \quad \textbf{(9)}$$

3.　実験結果と考察

プレゼンテーションをビデオで撮影し、スライドの対応付けを行った。動画は秒間 1 フレームでサンプリングし、各フレームの解像度は 640x480 画素である。スライドの情報は、PowerPoint のファイルから文字コードと座標を抽出した。約 10 分～30 分の動画 8 本を用いて、各フレームがどのスライドを含んでいるかを判断する。カメラアングルはほぼ固定であるが、多少の移動や話者がスライドをさえぎることなどがある(図 3)。

結果を表 1 に示す。正しいスライドと対応付けることのできたフレームを正解として、8 ファイルの平均で 96.7%という正解率が得られた。誤りの原因は、スライド中の文字数が少なく他のスライド中に同じ文字列が含まれている場合、文字のほとんどが数式である場合、スライドの変わり目の場合などであった。

図 3　実際の認識対象例、右は文字の拡大図

表 1　スライド同定の実験結果

No.	スライド枚数	フレーム数	正解フレーム数	正解率(%)
1	9	659	659	100
2	9	830	786	94.7
3	11	678	678	100
4	10	743	742	99.9
5	35	1379	1240	89.9
6	25	1566	1564	99.9
7	51	1112	1037	93.3
8	22	1705	1636	96
平均	-	-	-	96.7

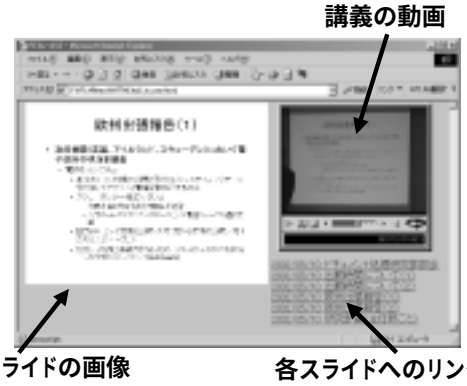

図 4　ブラウザによる学習画面例

図 4 に本手法を用いて作成した学習教材例を示した。AVI と PowerPoint ファイルを用意して処理すると、自動的に各スライドのタイトルや表示範囲を抽出し、動画と同期再生するために必要なファイル群を出力する。再生には、Web ブラウザと Plug-in があれば特別なソフトウェアは必要ない。動画はスライドバーで任意の時刻から再生でき、各スライドの先頭へのリンクも示されている。

4.　まとめ

本稿では、講義動画中のスライドの開始・終了時刻を自動的に検出するため、文字認識を用いてスライドを同定する手法を提案した。提案方式を用いることによって、動画を用いた e-learning コンテンツ作成時の作業コストを削減することができる。また、コンテンツの作成時および再生時に、特別な環境を必要とせず、従来の紙資料との対応づけにも応用できるのが特徴である。

今後の課題としては、文字情報が含まれないスライドへの対応が必要である。そのためには、画像特徴との併用などが考えられる。

[参考文献]

[1]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, (1997).

[2] 斉藤文彦:「遺伝的アルゴリズムを用いた画素選択テンプレートによる画像マッチング」, 信学論 D-II, Vol.J84-D-II, No.3, pp.488-499, (2001)

[3]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-DII, No.7, pp.1658-1666,(2000).