G-006 動的計画法を用いた音声による講義ビデオシーン自動分割
Lecture Video Segmentation Derived from Speech by Dynamic Programming

隅田 飛鳥 † 金寺 登 † 池端 孝夫 †
Asuka Sumida Noboru Kanedera Takao Ikehata

1. はじめに

近年高速なネットワーク環境が整備され、ビデオ教材を用いて自宅で手軽に予習・復習することが可能となってきた。しかし利用できるビデオ教材はまだ少ない。この原因の一つはビデオを編集するために非常に多くの労力と時間を要するためだと考えられる。ビデオを編集するためにはビデオの取り込み、シーン分割位置の検索、シーンの削除・移動・マージなどの作業を行う必要がある。特にシーン分割位置を検索するにはビデオを始めから最後まで繰り返し見る必要があり時間・労力ともに大きな負担となる。そこでビデオ教材作成を支援する方法として、ビデオシーンを自動分割するシステムの開発を検討している (図 1)。

自動的にシーン分割位置推定を行うために、ビデオ中の映像あるいは音声を用いる方法が研究されている。ビデオ中の映像を用いて自動シーン分割位置推定を行う研究に関して数多くの報告がある。これらの報告によれば、シーンの切り替わり位置で映像が大きく変化する場合においては高精度に分割を行うことができる [1]。しかし講義ビデオの内容が映像と密接に関連して変化することは少ないため、映像情報のみによるシーン分割は困難と考えられる。一方、講義ビデオの内容は音声と非常に密接に関連して変化する。そこで本研究では編集前の講義ビデオ中から抽出された音声情報よりシーンを自動分割する。講演 [2, 3] や編集後の講義 [4, 5] を対象としたシーン分割に関する研究はあるが、本研究のように編集前の講義を対象としたものはほとんどない。編集前の講義をシーン分割する場合には不要部分が数多く存在するためトピックの切り替わりの検出がさらに困難であると予想される。

2. シーン分割位置推定のための指標

一般的にビデオは複数のシーンからなっている。ビデオ中の隣接するシーン間が似ていれば一続きのシーン、似ていなければシーン間に分割点が存在すると考えられる。シーン間をコンピュータ上で比較するためには、各シーンの話題情報を何らかの指標に変換しなければならない。指標には TF-IDF[2, 4] や TF-IDF を考慮した相互情報量 [5]、χ^2 値 [6] などがよく用いられている。一方、独立成分分析 (Independent Component Analysis;ICA) を用いて次元数が語彙数に依存しない指標に文書を変換する方法 [7, 8] が提案されている。ICA では独立成分分析を用いて、語-文行列（各文における語の頻度）を、話題と語の関係、話題と文の関係に変換する [7, 8]。本研究では独立成分分析を用いて得られる関係のうち、シーン分割のための指標として話題と文の関係を利用する。

†石川高専

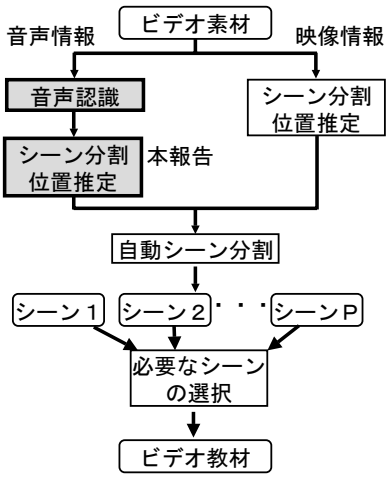

図 1: ビデオ教材作成支援システム例

3. シーン分割方法

講義や講演のビデオ素材とは別に配付資料などのテキストが用意できる場合には、ビデオ中の音声とテキストを対応付けることでシーンを効率的に分割できる [3, 4]。一方、講義や講演のテキストを利用できない場合には、話題の転換点を表す談話標識を用いる方法 [2] や隣接シーン間の音声を比較する方法 [5] が提案されている。本研究では編集前の講義ビデオを対象にテキスト情報が与えられない場合を想定する。

図 2 にシーン分割方法の概要を示す。一般的に隣接するシーン間が似ていれば一続きのシーン、似ていなければシーン間で分割できると考えられる。

このような考え方を元に隣接するシーンの類似度が極小となる谷の位置をシーン分割位置とする手法 (Hearst 法) が Hearst らによって提案されている [9, 10]。Hearst 法は類似度が極小となる谷の位置をシーン分割位置とする手法であるため、分割数をコントロールすることが難しい。ビデオセグメンテーションでは分割数をコントロールできる方が望ましい。分割数を指定できれば、分割し

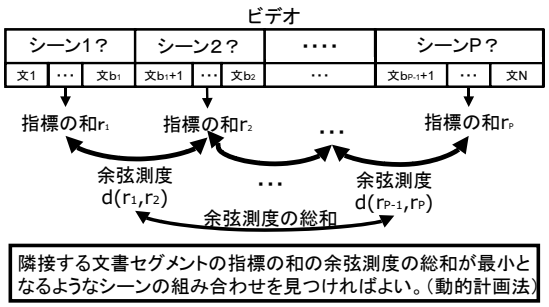

図 2: シーン分割方法概要

たい位置にシーン分割点が挿入されていなくても、分割数を増やすことでユーザが逐次的にシーン分割点を検索する手間を省くことができる。ユーザにとって分割数が多く余分に付加されたシーン分割点を無視する負担に比べ、分割数が少なくシーン分割点を検索する負担の方が大きいと考えられる。本研究では分割数をコントロールでき、かつ音声認識誤りにも頑健なシーン分割推定方法を検討する。

文書をある分割数で分割した時、互いに隣接するシーンが文書全体にわたって似ていなければ、文書を適切に分割できると考えられる。つまり隣接するシーン間における類似度の総和が最小となるようなシーンの組合せを見つければ良い。そこで本研究ではシーン分割位置推定を類似度の総和が最小となるようなシーンの組合せを探す問題とみなし、動的計画法 (Dynamic Programming;DP) を用い解く方法を提案する。

以下に比較に用いる Hearst 法と、DP を用いて解く提案手法について説明する。

3.1 Hearst 法によるシーン分割方法 [9, 10]

Hearst 法は文書中の隣り合うブロック間における類似度の極大値と極小値との差が大きい極小値の位置を分割位置とする手法である。まず文書中の単語の前方、後方に段落程度の大きさの窓をかけ、それぞれブロックとみなす。前後のブロックの単語出現頻度から余弦測度を求め、ブロック位置 i におけるブロック間の類似度 $score(i)$ を次式で定義する。

$$score(i) = \frac{\sum_t w_{t,b_1} w_{t,b_2}}{\sqrt{\sum_t w_{t,b_1}^2 \sum_t w_{t,b_2}^2}}$$

ここで t は各語、b_1、b_2 は隣接する各ブロックを、w_{t,b_n} は単語 t がブロック b_n に出現する頻度を表す。類似度 $score(i)$ を w 語分ずつ基準点をずらしながら同様に求める。

類似度の極大点は話題が盛り上がっている部分、極小点はシーン境界と考えられる。つまり次式で与えられる類似度の極大値と極小値との差が大きいほどシーン境界らしいと考えられる。

$$depth\ score(j) = (y_l - y_j) + (y_r - y_j)$$

ここで j は注目している極小点のブロックの位置、l は j に隣接する左の極大点のブロックの位置、r は j に隣接する右の極大点のブロックの位置、y_l、y_j、y_r はそれぞれ l、j、r における類似度を表す。$depth\ score$ はシーン境界らしさを表すので小さい値だとシーン境界とは考えにくい。よってある閾値より大きい $depth\ score$ における j の位置を境界候補とみなす。閾値として $depth\ score$ の 平均 − 標準偏差 $\times 0.5$(the conservative measure;HC) と 平均 − 標準偏差(the liberal measure;LC) がよく用いられる。なお類似度の微弱な振動を除去するために $depth\ score$ を求める前に類似度に対してスムージングを行った。

3.2 DP によるシーン分割方法

まずビデオから得られた文書 (自立語のみ) を仮に複数の文書セグメントに分割する。次に各文書セグメントを指標に変換し隣接する文書セグメントが似ているかどうかを調べる。隣接する文書セグメントが似ているかどうか調べるには指標の余弦測度を用いる。余弦測度が小さい程文書セグメントは似ておらず、大きい程文書セグメントは似ていると考えられる。つまり指標の余弦測度の総和が最小であれば全ての文書セグメント間が似ていないことになり、文書全体を適切にシーン分割できると考えられる [5]。そこで本研究ではシーン分割位置推定を余弦測度の総和が最小となるようなシーンの組合せを探す問題とみなし、動的計画法 (Dynamic Programming;DP) を用い解く方法を提案する。

まず指標 I を用いて文 1〜N を P 分割し 1〜b_1, b_1+1〜b_2, ..., $b_{P-1}+1$〜N の文書セグメントにする。指標 I の各列は文 1〜N に対応する。p 番目の文書セグメント $b_{p-1}+1$〜b_p に対応する指標 I の和 $\boldsymbol{r}_p$ を次式で定義する。

$$\boldsymbol{r}_p = \sum_{m=b_{p-1}+1}^{b_p} I_m$$

但し I_m は I の第 m 列とする。隣接する $\boldsymbol{r}_p$ と $\boldsymbol{r}_{p+1}$ の余弦測度の和が最小になるように推定シーン境界 $\hat{B}_p = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_{P-1})$ を次式で決定する。

$$\hat{B}_p = \arg\min_{B_p} \sum_{p=1}^{P-1} d(\boldsymbol{r}_p, \boldsymbol{r}_{p+1}) \qquad (1)$$

但し B_p は $B_p = (b_1, b_2, \ldots, b_{p-1})$ を表す。また $d(\boldsymbol{a}, \boldsymbol{b})$ はベクトル$\boldsymbol{a}$, $\boldsymbol{b}$間の余弦測度を表す。余弦測度を求める際、ベクトル$\boldsymbol{a}$が示す文書に含まれる語数が一定の語数に満たない場合、前方に三角スムージングを行った。同様にベクトル$\boldsymbol{b}$が示す文書に含まれる語数が一定の語数に満たない場合、後方に三角スムージングを行った。

(1) 式を解くために動的計画法を用いる。まず、文 1〜i を j 分割したときの隣接文書セグメント間の累積距離 $g(i,j)$ を次式で定義する。

$$g(i,j) = \min_{B_j} \sum_{p=1}^{j-1} d(\boldsymbol{r}_p, \boldsymbol{r}_{p+1}) \qquad (2)$$

$\boldsymbol{s}(i,j)$ を第 j 番目の文書セグメントが文 i で終了するときの $\boldsymbol{r}_j$ とすると、以下のようにシーン境界を求めることができる。

1. j=1 のとき i=1,2,... ,N について

$$g(i,1) = 0$$

$$\boldsymbol{s}(i,1) = \sum_{m=1}^{i} I_m$$

2. $j \geq 2$ のとき i=j,j+1,... ,N について

$$\hat{k}(i,j) = \arg\min_{k=j-1,\cdots,i-1} \left\{ g(k,j-1) + d\left(\boldsymbol{s}(k,j-1), \sum_{m=k+1}^{i} I_m \right) \right\}$$

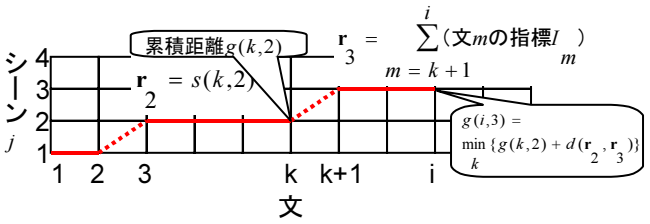

図 3: DP によるシーン分割点探索例

$$\boldsymbol{s}(i,j) = \sum_{m=\hat{k}(i,j)+1}^{i} I_m$$

$$g(i,j) = g(\hat{k}(i,j), j-1) + d\left(\boldsymbol{s}(\hat{k}(i,j), j-1), \boldsymbol{s}(i,j)\right)$$

3. p=P-1,P-2,... ,1 について

$$\hat{b}_p = \hat{k}(b_{p+1}, p+1)$$

図 3 に文 1〜i を 3 分割したときの隣接文書セグメントの累積距離 $g(i,3)$ を求める例を示す。3 番目の文書セグメントが文 $k+1$〜i であると仮定したとき、3 番目の文書セグメントの指標は $\boldsymbol{r}_3 = \sum_{m=k+1}^{i} I_m$ で与えられる。2 番目の文書セグメントが文 k で終了するとすれば、2 番目の文書セグメントの開始点は $\hat{k}(k,2)+1$ で与えられ、2 番目の文書セグメントの指標は $\boldsymbol{r}_2 = \boldsymbol{s}(k,2) = \sum_{m=\hat{k}(k,2)+1}^{k} I_m$ で与えられる。したがって $g(i,3)$ は、文 1 から k を 2 分割したときの累積距離 $g(k,2)$ に、2 番目の文書セグメントの指標 $\boldsymbol{r}_2$ と 3 番目の文書セグメントの指標 $\boldsymbol{r}_3$ の余弦測度を加えたものになる。

4. シーン分割結果

4.1 実験条件

実験対象として表 1 に示すビデオ素材を用意した。これらのビデオ素材は、5 名の男性教員による約 90 分の講義 5 回分である。表 1 における文数は 1 sec 以上の無音区間が継続するかどうかで区切られた境界候補（音声区間）数である。収録には接話型ヘッドセットを用いたため、雑音等の影響は少ない。対象となるビデオ素材から音声情報のみを抽出し、16 kHz にダウンサンプリングを行った。次に音声区間ごとに日本語ディクテーション基本ソフトウェア（98 年度版）[11] を用いて音声認識を行った。音響モデルは 2000 状態 16 混合の tri-phone とし、各種学習・評価条件は文献 [11] と同様である。ただし、学習・評価データには男性話者のみを用いた。言語モデルは講演の書き起しテキストにより学習された言語モデル [12] を用いて認識を行った。認識結果から得られた文書の単語正解率 §・単語正解精度 ¶・未知語率を表 1 に示す。

§

$$\text{単語正解率} = \frac{\text{総単語数} - \text{置換誤り単語数}}{\text{総単語数}}$$

¶

$$\text{単語正解精度} = \frac{\text{総単語数} - \text{置換誤り単語数} - \text{付加単語数}}{\text{総単語数}}$$

表 1: ビデオ素材

ビデオ素材	文数	共通正解境界数	単語正解率 [%]	単語正解精度 [%]	未知語率 [%]
1	539	21	50.9	33.5	7.1
2	592	23	46.7	31.2	2.9
3	544	14	40.3	22.7	8.2
4	468	18	32.0	12.4	5.5
5	430	24	45.8	26.3	8.6
平均	515	20	43.1	25.2	6.5

音声認識によって得られたテキストから自立語のみを抽出後、提案手法・Hearst 法を用いて分割を行い比較した。提案手法では、独立成分分析 (ICA) による指標を用いシーン分割を行った。独立成分数は予備実験 [13] より表 1 に示されている文数の約 0.15 倍とした。なお、余弦測度を求める際それぞれ 100 語で三角スムージングを行った。Hearst 法では窓の幅を 80 語とし、8 語ずつシフトしながら類似度を求めた。また前後 16 語でスムージングを行った。

正解データとして、5 名の評価者に対象としたビデオ素材を提示しシーン境界の許容範囲を求めてもらった。許容範囲が 3 名以上一致する範囲を OR 合成し、正解とした。各データの共通正解境界数を表 1 に示す。また必要以上に長い無音区間は不要部分として削除される可能性が高いため、5 sec 以上の無音区間の両端もシーン境界に追加した。

以下に評価尺度として用いた再現率、適合率の式を示す。

$$\text{再現率 } (recall) = \frac{\text{回答中の正解数}}{\text{正解数}}$$

$$\text{適合率 } (precision) = \frac{\text{回答中の正解数}}{\text{回答数}}$$

本研究では再現率を優先した。シーン分割位置推定において再現率が低い場合、ユーザが逐次的にシーン分割点を検索しなければならず、労力は大きいと考えられる。一方、余分に付加されたシーン分割点は無視すればよい。よって、シーン分割位置推定による誤りが少ない（再現率が高い）ことが望ましい。

4.2 実験結果

表 2 に音声認識結果による Hearst 法の結果を示す。境界候補は表 2 における閾値以上の *depth score* の位置とし、HC、LC、None はそれぞれ *depth score* の 平均 − 標準偏差 × 0.5、平均 − 標準偏差、0 を表す。また表 2 における分割率は分割数を各ビデオ素材の文数で割った値の平均を表し、再現率、適合率は各ビデオ素材の再現率、適合率の平均を表す。この結果より Hearst 法では閾値によって分割率を自由に決められないことが分かる。

図 4 に音声認識結果によるシーン分割結果を示す。横軸は分割数を各ビデオ素材の文数で割った分割率で、縦軸は各ビデオ素材の再現率・適合率の平均を表す。この図と表 2 を比較すると提案手法は Hearst 法より分割率

表 2: Hearst 法による結果

閾値	分割率	再現率 [%]	適合率 [%]
HC	0.09	28.4	12.4
LC	0.09	29.2	12.7
None	0.12	34.3	11.5

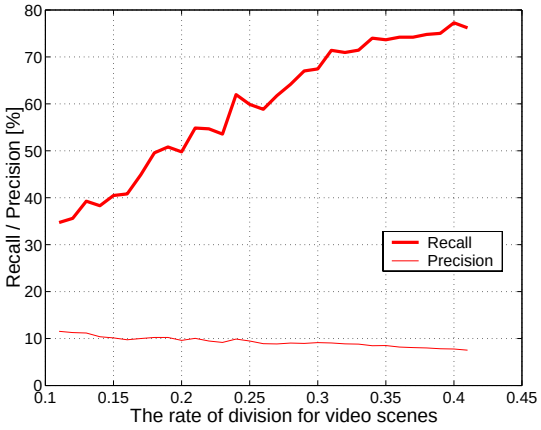

図 4: 音声認識結果によるシーン分割結果

を自由に設定でき、より高い再現率が得られることが分かる。

提案手法を用い、音声認識結果と書き起しテキストによるシーン分割を行った結果を図 5 に示す。横軸、縦軸は図 4 と同様である。この結果より音声認識結果を用いたシーン分割性能は書き起しテキストと同程度であることが確認された。これは音声認識性能がある程度低くても複数箇所において同じ誤りであればシーン分割には影響を与えないためと考えられる。

5. まとめ

ビデオ教材作成支援を目的として、編集前の講義ビデオから抽出した音声情報により、ビデオを自動分割した。シーン分割には独立成分分析を用いたトピック表現（指標）とポーズ情報を利用した。シーンの対応付けには DP を用い、隣接するシーンの余弦測度の総和が最小になるように最適化した。実験の結果、提案手法を用いることで、Hearst 法より高い分割性能を持ちながら、分割数を自由に設定できることがわかった。また、音声認識結果を用いても書き起しテキストと同程度のシーン分割性能が得られることが確認された。

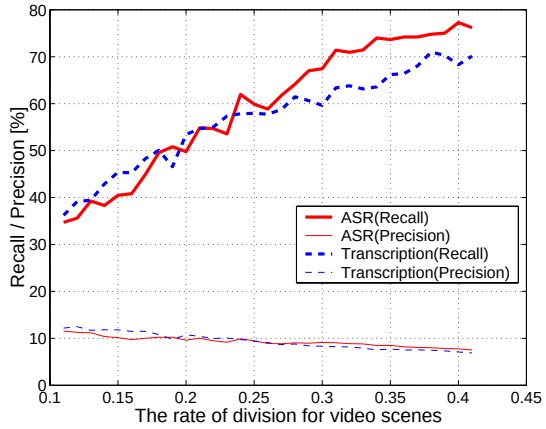

図 5: 音声認識結果と書き起しテキストによるシーン分割結果

謝辞

本研究の一部は文部科学省科学研究費補助金 (課題番号 14580246) を受けて行われた。

参考文献

[1] 中村裕一，外村佳伸，“見たい部分を簡単に短時間で，” 電子情報通信学会誌，pp.346–353，1999.

[2] 長谷川将宏，秋田祐哉，河原達也，“談話標識の抽出に基づいた講演音声の自動インデキシング，” 情報処理学会研究報告，pp.35–42，2001.

[3] 伊藤克亘，藤井敦，石川徹也，“音声文書検索を用いたオンデマンド講義システム，” 信学技報，pp.55–60，Dec. 2001.

[4] 山本夏夫，緒方淳，有木康雄，“トピックセグメンテーションに基づく講義ビデオの構造化の検討，” 情報処理学会研究報告，pp.59–64，2002.

[5] 緒方淳，山本夏夫，鷹尾誠一，有木康雄，“講義データを対象とした音声認識と構造化の検討，” 情報処理学会研究報告，pp.79–84，2001.

[6] K.Ohtsuki, T.Matsuoka, S.Matsunaga, and S.Furui, “Topic extraction based on continuous speech recognition in broadcast news speech,” IEICE Trans. Inf. & Syst., vol.E85-D, no.7, pp.1138–1144, 2002.

[7] A. Kabán, Latent Variable Models With Application to Text Based Document Representation, Ph.D thesis, The University of Paisley, 2001, http://cis.paisley.ac.uk/kaba-ci0/ata_thesis.zip.

[8] A.Kabán, and M.A.Girolami, “Fast extraction of semantic features from a latent semantic indexed corpus,” Neural Processing Letters, vol.15, no.1, pp.31-43, 2002.

[9] M. Hearst, “Multi-paragraph segmentation of expository text,” 32nd. Annual Meeting of the Association for Computational Linguistics, pp.9–16, 1994.

[10] M. Hearst, “Texttiling: Segmenting text into multi-paragraph subtopic passages,” Association for Computational Linguistics, pp.33–64, 1997.

[11] 河原達也，李晃伸，小林哲則，武田一哉，峯松信明，伊藤克亘，山本幹雄，山田篤，宇津呂武仁，鹿野清宏，“日本語ディクテーション基本ソフトウェア 98 年度版，” 音響学会誌，vol.56, no.4，pp.255–259，2000.

[12] 南條浩輝，加藤一臣，李晃伸，河原達也，“大規模な日本語話し言葉データベースを用いた講演音声認識，” 電子情報通信学会論文誌，vol.J86-DII, no.4，pp.450–459，2003.

[13] 隅田飛鳥，金寺登，寺家谷純，池端孝夫，船田哲男，“独立成分分析を用いた音声による講義ビデオシーン分割，” 電子情報通信学会技術研究報告 SP2003-61，vol.103, no.220, pp.7–12，2003.