L-32

第12回画像センシングシンポジウム予稿集, pp.535-540, Yokohama, Jun., 2006

# 講義イベント検出に基づく短縮講義ビデオの自動生成

## A Method for Generating a Time Shrunk Lecture Video by Event Detection

横井 隆雄† 桐井 孝嘉† 藤吉 弘亘†

Takao Yokoi†, Takayoshi Kirii† Hironobu Fujiyoshi†

†中部大学工学部情報工学科

†Department of Computer Science, Chubu University

E-mail: {taka,kiry,hf}@vision.cs.chubu.ac.jp

### Abstract

本稿では，講義中のイベントを自動検出し，その結果から時間短縮講義ビデオを生成する手法を提案する．提案手法では，時間短縮講義ビデオを生成するために，発話区間と板書区間の検出を行う．発話区間は，事前に抽出した複数の講師の発話・無発話での音声特徴 (16次LPC ケプストラム・パワースペクトル 40Hz-900Hz) と，入力された音声データとのマハラノビス距離を計算し判定する．板書区間の検出では，板書の変化を抽出するために講師領域を正確に求める必要がある．本研究では，グラフカットによる講師領域の正確なセグメンテーション結果を用いて，講師消去画像を生成する．講師消去画像のフレーム間差分を毎フレーム求めることで，板書区間の検出が可能となる．検出した講義イベントから必要のない区間はカット，板書イベントのみの区間は 3 倍速にすることで時間短縮講義ビデオを自動生成する．評価実験の結果，提案手法は従来手法に比べ，人が編集したビデオと同程度の時間短縮ビデオを自動生成できることを確認した．

## 1 まえがき

近年，国内外の教育機関において Web Based Training(WBT) による e-learning の実施が増加している．特に，撮影した講義をインターネットを介して配信する事は，遠隔地等での教育格差の是正や復習 (リプレイ) による教育効果が期待できる．しかし，配信する講義映像の作成に，専門のカメラマンや編集者の雇用は，コスト面から難しい．

このような問題に対して，我々は 1 台の高解像度映像カメラから講義ビデオを自動的に生成する手法を提案している [1]．ハイビジョンカメラで撮影した講義室前方の高解像度映像から，講師を追従するようなトリミングを行うことで講義ビデオを生成する．トリミングの際に放送カメラマンの撮影技術に基づく仮想カメラワークを実現することで，臨場感ある映像を生成することが可能となる．しかし，学生にとって一度受講した講義を初めから終わりまで視聴することは，時間を要する．そこで本研究では，効率よく講義ビデオを再生するために，講義に関係ないシーン (例えば，講師が動作していない状態) を削除することで，時間短縮した講義ビデオの自動生成を目的とする．

映像要約に関しては，これまでに多くの研究が報告されている [2, 3, 4, 5]．Smith 等は映像解析や音声処理などによりシーン変化などの特徴を検出し，実際の映像からスキミングを行い要約映像を生成する [2]．三浦等は，料理映像の特徴を調査し，画像処理により，特徴的なシーンを自動検出し要約を行っている [3]．しかし，これらは一般的な映像や料理映像を対象とした映像に対しての要約手法であるため，本研究のの対象である講義映像には適さない．

教育を対象とした映像のインデキシングとして，Liu 等は，パワーポイントを用いた講義を対象とし，スライドの変化により講義の要約を行っている [4]．しかし，通常，講義においては，パワーポイントのみを用いた講義だけではなく，板書の場合が多い．石塚等は，板書を考慮した講師の位置や音声からの講義状態予測に基づく講義映像のインデキシングを行っている [5]．しかし，講義のインデキシングからの短縮映像の生成までは行われていない．

本稿では，講義イベント検出に基づく時間短縮講義ビデオの自動生成法について提案する．講義において，出現頻度の高い発話区間と板書区間に着目し，これら 2 つの講義イベント区間検出の結果に基づいて時間短縮講義ビデオを生成する．発話区間検出には，LPC ケプストラム係数とパワースペクトルを特徴量としたマハラノビス距離により，入力音声データを発話もしくは無発話区間に判別する．板書区間検出では，黒板上に新しく板書されたかどうかを判定する際，講師領域を消去する必要がある．本稿では，より正確な講師領域

抽出として，フレーム間差分とグラフカットによるセグメンテーション法を提案する．

## 2 高解像度映像からの講義ビデオ生成

本研究では，ハイビジョンカメラを用いて講義の撮影を行う．黒板全体が入り，かつ黒板の板書文字を読むことができるように HDV(1080$i$) カメラ一台を講義室の後方部に設置し講義映像を取得する．しかし，閲覧者のユーザが持つノート PC の表示解像度は一般に XGA が多く，HDV カメラで撮影した高解像度映像 (1,440×810) をそのまま表示することができない．この問題を解決するために，図 1 に示すように講師等の注目対象に追従するように高解像度映像からトリミングを行い，講義ビデオを自動生成する．トリミングを行う際，放送カメラマンが撮影するような仮想カメラワークを実現することで臨場感ある講義映像生成を行う [1]．高解像度画像から仮想カメラワークによる講義映像の生成手順を以下に示す．

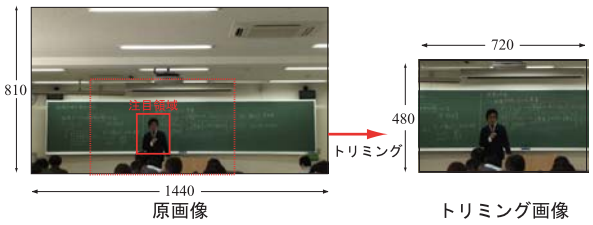



図 **1**　高解像画像からのトリミングによる画像生成

**step1 講師位置検出**

講師等の移動体領域をフレーム間差分法により求める．しかし，フレーム間差分によって得られた注目対象の位置座標は，講師の敏速な動きに追従して激しく変動する場合がある (図 2(a))．従って，求めた座標値を基にトリミングを行うと，映像が激しく横にゆれる問題がおこる．そこで，バイラテラルフィルタを用いて変動の抑制を行う．バイラテラルフィルタを用いることで，エッジを保存しつつ，細かな振動を抑制させることができる (図 2(b))．

**step2 カメラワークタイミング**

バイラテラルフィルタにより抑制された講師領域座標に対して，零交差処理を施し講師の動作特徴点を求める．隣り合う各特徴点間の位置座標の変化より，変化の激しい区間をパンニング区間，変化の少ない区間をズーミング区間と判定する (図 2)．ズーム区間を設定する理由は，映像が長期的に変化がない状態を防ぎ，立ち止まって話す講師の発話動作や表情をよく見えるようにするためである．

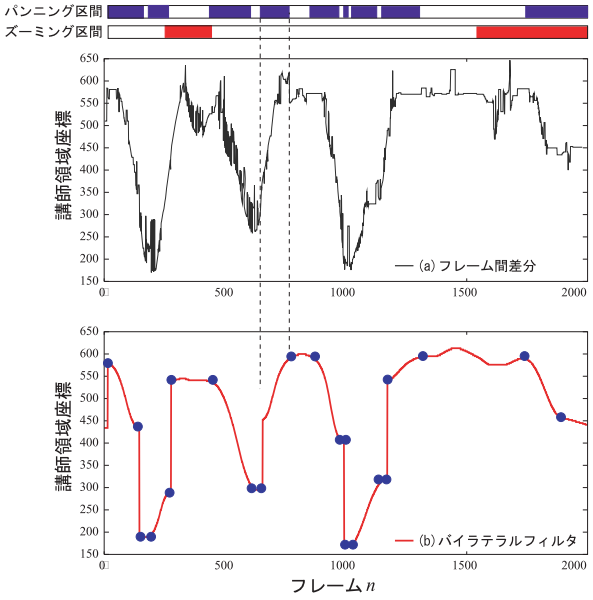



図 **2**　カメラワークタイミング

**step3 仮想カメラワーク生成**

検出した区間に対して，以下に示す放送カメラマンの撮影技術による知見 [6, 7] から仮想カメラワークモデルを算出し，トリミング位置を決定することで，臨場感ある講義映像を生成する (図 3)．

- パンニング速度曲線非対称型で，減速時間が加速時間に比べて 6 割程度長い．
- 最大パンニング速度が生じる時間が加速時と減速時で異なる．
- 最大拡大変化率は，ズームイン時には後半に，ズームアウト時には前半に発生する．

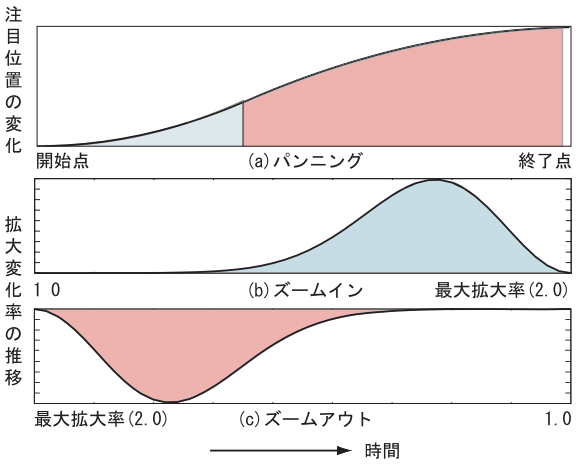



図 **3**　仮想カメラワークモデル

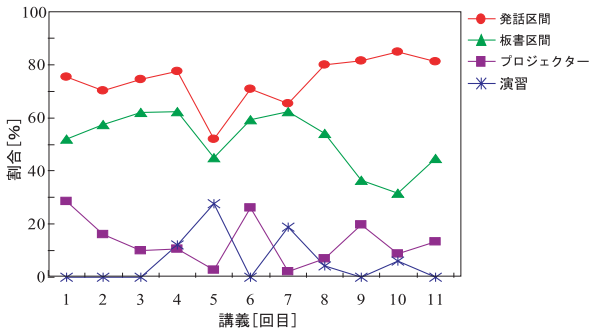



図 4　講義の構成要素

## 3　講義イベント検出による短縮ビデオ生成

講義映像を再生する際，学生にとって重要な区間だけを視聴できることが望ましい．図 4 は，黒板を使用する一般的な講義 11 回分 (1 回 90 分) の講義における講義要素 (発話，板書，プロジェクタを用いた説明，演習) の頻度を表したグラフである．黒板を用いる講義において，特に板書と発話の時間が多く存在し，講義において重要な要素であると考えられる．そこで，本手法では，短縮講義ビデオを生成するために，講義における講師の発話と板書区間の検出を行う．

### 3.1　発話区間検出

講義における発話検出は，事前に複数の講師（10 人分) の講義における音声データから発話・無発話での音声特徴 (表 1 参照) を抽出する [8]．入力された音声データと発話クラス・無発話クラスとのマハラノビス距離を求め，発話か無発話かを判定する．

表 1　分析条件

| | |
|---|---|
| サンプリング周波数 | 11kHz |
| 分析窓 | 48 msec<br>ハミング窓 |
| フレームシフト | 18 msec |
| 特徴量 | 16 次 LPC ケプストラム<br>パワースペクトラム (150-900Hz) |

### 3.2　板書区間検出

板書区間は黒板上に新しく板書されたかどうかを判定する．新規板書の判定は，入力画像から黒板上の講師部分を消去した画像のフレーム間差分を取ることで検出することができる．

#### 3.2.1　講師領域の検出

西口等は，フレーム間差分を利用した板書区間検出を提案している [10]．フレーム間差分を用いた場合，背景画像を用意する必要がなく，照明変動に対してロバストに移動領域を検出できるというメリットがある．しかし，講師領域全域を前景ピクセルとして得ることができないため，フレーム間差分で得られた結果に対して，一定領域の大きさの矩形領域を講師領域としている．この場合，得られる講師領域には，背景画像も含まるため，正確な板書領域を検出することができない．そこで，フレーム間差分の結果を利用して，グラフカット [11] による正確な前景ピクセルの抽出を提案する．

グラフカットは，ピクセルをノードとしたグラフを作成し，そのグラフの最小カットを求めることで，物体と背景のセグメンテーションを行う．グラフ $\mathcal{G}$ は，頂点 $\mathcal{V}$（ノード）とそれらを結ぶ線 $\mathcal{E}$（エッジ）の関係を表したものであり，$\mathcal{G} =< \mathcal{V}, \mathcal{E} >$ と表す．画像からグラフを作成する場合，ピクセルの集合を $\mathcal{P}$，近傍ピクセルの集合を $\mathcal{N}$，またターミナルを表現するノードである “souce(object)”を $s$，“sink(background)”を $t$ とした際，ノード $\mathcal{V}$ とエッジ $\mathcal{E}$ 以下の式で表される．

$$\mathcal{V} = \mathcal{P} \cup \{s, t\} \quad (1)$$

$$\mathcal{E} = \mathcal{N} \cup_{p \in \mathcal{P}} \{\{p, s\}, \{p, t\}\} \quad (2)$$

このようなグラフ $\mathcal{G}$ をフローネットワークと呼び，その構成を図 5(a) に示す．近傍ピクセル $\mathcal{P}$ 間のエッジを n-link，各ピクセル $p$ から $s$ や $t$ に接続しているエッジを t-link と呼ぶ．注目ピクセルを $p \in \mathcal{P}$，その近傍ピクセルを $q \in \mathcal{N}$，ノード $p$ と $q$ をつなぐエッジを $\{p, q\}$ とするとき，$\{p, q\}$ のエッジコスト $B_{\{p,q\}}$ は次式により求められる．

$$B_{\{p,q\}} = \exp\left(-\frac{(I_p - I_q)^2}{2\sigma^2}\right) \cdot \frac{1}{dist(p, q)} \quad (3)$$

ここで，$I_p$ はピクセル $p$ の輝度値，$dist(p, q)$ はピクセル間の距離，$\sigma$ は近傍との連結度に関連するパラメータであり経験的に決定する．t-link は，あらかじめ物体領域と背景領域のヒント (seed) となるピクセル $\mathcal{O}$(“object” seed) と $\mathcal{B}$(“background” seed) を与えることにより，次式から決定される．

$$\{p, s\} = \begin{cases} \lambda \cdot R_p(\text{“}bkg\text{”}) & , p \notin \mathcal{O} \cup \mathcal{B} \\ K & , p \in \mathcal{O} \\ 0 & , p \in \mathcal{B} \end{cases} \quad (4)$$

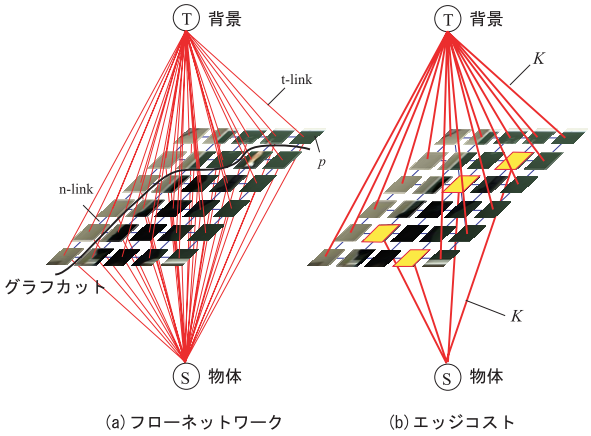
$$\{p, t\} = \begin{cases} \lambda \cdot R_p(\text{“}obj\text{”}) & , p \notin \mathcal{O} \cup \mathcal{B} \\ 0 & , p \in \mathcal{O} \\ K & , p \in \mathcal{B} \end{cases} \quad (5)$$

$$R_p(\text{“}obj\text{”}) = -\ln \Pr(I_p|\mathcal{O})$$

$$R_p(\text{“}bkg\text{”}) = -\ln \Pr(I_p|\mathcal{B})$$

$$K = 1 + \max_{p \in \mathcal{P}} \sum_{q:\{p,q\} \in \mathcal{N}} B_{\{p,q\}}$$

$\lambda$ は n-link との関係を表す比例係数である．




(a) フローネットワーク　(b) エッジコスト

図 5　グラフカット

グラフカットによるセグメンテーションを行う場合，t-link のエッジコストを求めるために seed を決定しなければならない．文献 [11] では，seed をインタラクティブにユーザが与えるためセグメンテーションのプロセスは自動化されていない．そこで、本稿では，フレーム間差分の結果から物体と背景の seed を決定する．まず，現画像フレームに対して，次式に示すフレーム間差分により前景ピクセルを検出する [9]．

$$\Delta I^t = \max\{|\ I^t - I^{t-k}\ |\} \qquad 0 < k < 5 \qquad (6)$$

$I^t$ は現在の入力画像，$I^{t-k}$ は $k$ フレーム前の画像とする．急激な輝度値の変化がピクセル上に生じたとき，変化量 $\Delta I^t$ の値は大きくなる．輝度変化 $\Delta I^t$ をしきい値処理することにより，移動体と判定したピクセルに対応するノードを物体 seed($p \in \mathcal{O}$) とする．また，検出領域を膨張した拡散処理により，矩形領域の上，左，右端の領域に対応するノードを背景 seed として設定する (図 5(b) 参照)．各エッジコストを式 (4),(5) から求めることによりグラフ $\hat{\mathcal{G}}$ が完成する (図 5(b))．作成した $\hat{\mathcal{G}}$ において，$s$ と $t$ を分割するエッジコストの和が最小となるようにエッジを切断し，2 つのグラフへ分割する．グラフカットによるセグメンテーションの結果，図 6(d) のような物体マスクパターン $O - t_M$ を得る．

$$O_M^t = \begin{cases} 1 & : object \\ 0 & : background \end{cases} \qquad (7)$$

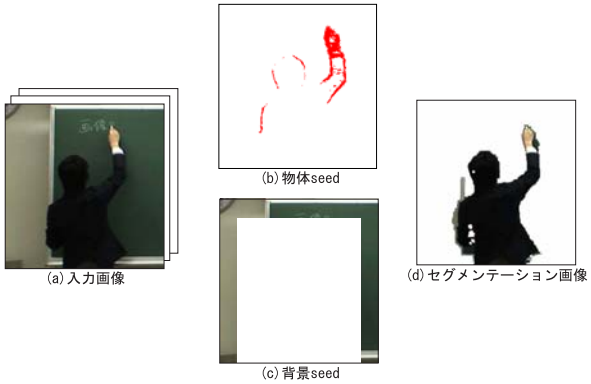



図 6　講師セグメンテーション

### 3.2.2　講師消去画像の生成

講師が含まれていない画像を初期講師消去画像 $I_c^0$(図 7(c)) とし，3.2.1 で得られた講師領域以外のピクセル $\bar{O}_M$(黒板領域) を更新することでフレーム $t$ における講師消去画像 $I_c^t$ を生成する．

$$I_C^t = I^t \cdot \bar{O}_M^t + I_C^{t-1} \cdot O_M^t \qquad (8)$$

ここで，$I^t$ は講師ピクセルを含む現フレーム画像を表す．図 7(b) に生成した講師消去画像例を示す．

### 3.2.3　黒板の変化に基づく板書検出

生成した講師消去画像をもとに，新しく板書が書かれたかどうかの判定を行う．$I_C^t$ と $I_C^{t-1}$ のフレーム間差分を求め，黒板上に輝度変化が生じたかを検出し，フレーム $t$ を板書区間と判定する．

$$\Delta I_C = \sum_{(i,j)\in I} |I_C^t(i,j) - I_C^{t-1}(i,j)| \qquad (9)$$

図 7(d) に提案手法による板書検出例を示す．

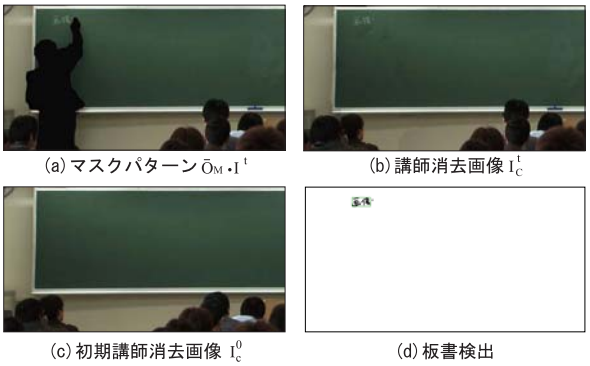



図 7　板書検出例

### 3.2.4　短縮講義ビデオの生成

検出した講義イベント結果を用いて時間短縮講義ビデオを生成する．図 8 に示すブロック A や D(無発話，無動作) のような何も講義要素が存在しない区間は，講義の内容に関する情報量が少ないと考えられ，このようなシーンブロックは消去する．また，ブロック F のよ

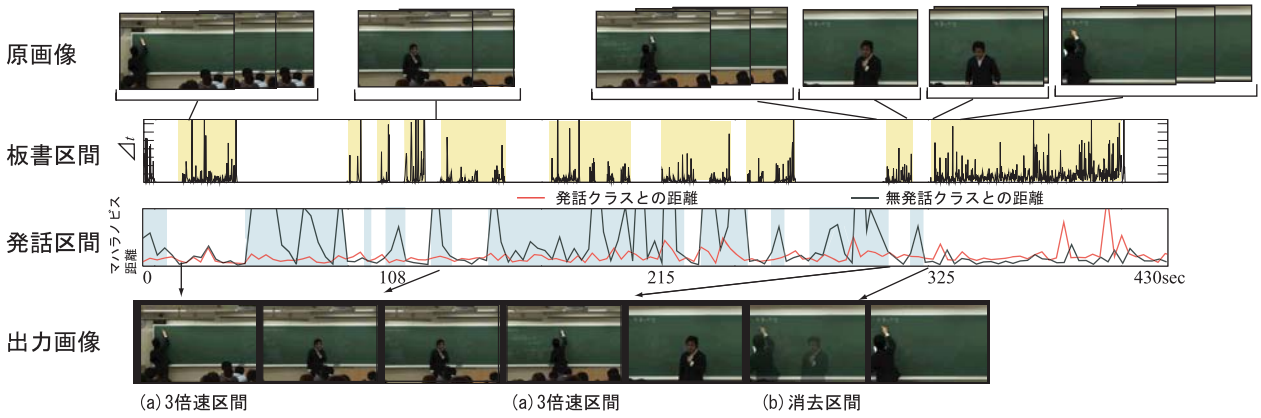



図 **9**　講義イベント検出による講義短縮例

うに板書区間で発話していない区間は，講師の板書動作を確認できれば良く，早送りが可能である．そこで，板書のみの区間は 3 倍速で再生を行う．発話と板書をしていない割合は，講義時間に対して 4% ～ 10%，板書のみは 10% ～ 30%であるため，90 分の講義を 60~70 分に短縮することが可能となる．講義要素が存在しないシーン区間を消去した場合，映像が急激に変化するため違和感のある映像となる場合がある．そこで，シーンチェンジにおける違和感を取り除くために，切り替わり前後数フレームで，クロスフェード処理を施し，自然な映像の切り替わりを実現する (図 9(b))．

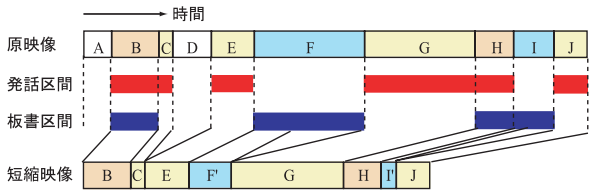



図 **8**　講義短縮概要

## 4　評価実験

HDV カメラを用いて撮影した講義映像 3 本 (各 80 分) に対して，本手法を適用し板書区間の検出を行う．撮影した映像には，説明，板書，板書文字を消すという動作が含まれている．提案手法により自動検出した発話区間ならびに板書区間の評価には再現率と適合率を用いる．再現率は，検出漏れの少なさを表す指標であり，適合率は，誤検出の少なさを表す．

$$\text{再現率} = \frac{\text{正検出}}{\text{正検出} + \text{検出漏れ}} \tag{10}$$

$$\text{適合率} = \frac{\text{正検出}}{\text{正検出} + \text{誤検出}} \tag{11}$$

### 4.1　講義イベントの検出結果

発話検出結果を表 2 に示す．表 2 より，講義中の発話区間検出において，約 96%の再現率と適合率を得ることができた．発話区間の未検出の理由は，講師の声が小さかった場合である．誤検出は，学生の声によりざわついていた時であった．共にマイクを使用しない講義であった．表 3 に，板書検出結果を示す．板書検出

表 **2**　発話区間検出結果

| | 再現率 [%] | 適合率 [%] |
|---|---|---|
| Movie1 | 95.6 | 97.5 |
| Movie2 | 97.5 | 96.0 |
| Movie3 | 97.0 | 94.6 |
| 平均 | 96.7 | 96.0 |

においては，平均 85.1%の再現率, 平均 95.7%の適合率を得ることができた．提案手法は，従来法 [10] の板書検出手法より講師領域をより正確なセグメンテーションが可能であるため，板書区間を高い検出率で得ることができた．再現率が適合率に比べて値が低い理由は，講師が黒板に向かって文字を書く際に，文字と講師が重なるために発生したオクルージョンと，黒板に書く文字の色の薄さによる検出漏れが原因である．しかし，講師と文字が重なりによる板書検出の遅延は 5 秒程度であり，また文字の薄さに関する検出漏れについては，講義中において長期間連続して発生することはなく，短縮する際に大きな問題とはならない．

### 4.2　時間短縮講義ビデオの生成結果

図 10 に，人手により編集した場合 (A)，本手法により生成した場合 (B) 及び従来の板書検出法 [10](C) による短縮率を示す．提案手法により生成される映像の短縮率は，従来法に比べて手動での短縮率に近く，手動での編集に近い短縮ビデオの自動生成が実現できたこ

表 **3**　板書区間検出

| | 再現率 [%] | | 適合率 [%] | |
|---|---|---|---|---|
| | 本手法 | 手法 [10] | 本手法 | 手法 [10] |
| Movie1 | 84.8 | 53.9 | 96.7 | 84.6 |
| Movie2 | 82.4 | 54.7 | 93.9 | 74.9 |
| Movie3 | 88.1 | 57.0 | 96.4 | 87.5 |
| 平均 | 85.1 | 55.2 | 95.7 | 82.3 |

とが分かる．

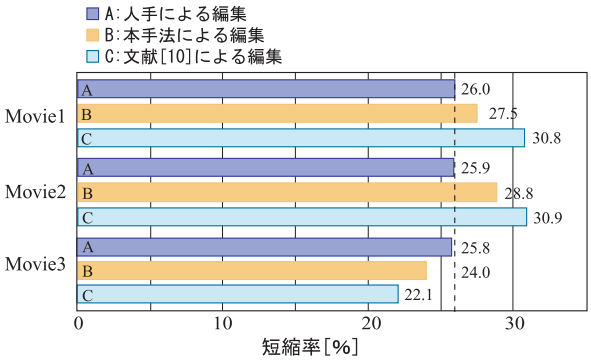



図 **10**　各手法による短縮率

## 5 まとめ

本稿では，講義イベントである発話区間と板書区間検出に基づく時間短縮講義映像生成法について提案した．フレーム間差分とグラフカットによる自動セグメンテーション法を提案し，従来法に比べ精度の高い板書区間検出が可能となった．また，提案手法は人が編集したビデオと同等の時間短縮映像を自動生成することができた．

今後の課題は，ユーザの入力に対してカメラワークの変更を可能とするインタラクティブ機能の追加や，コンテンツに応じた講義のインデキシング等が挙げられる．

## 参考文献

[1] T. Yokoi and H. Fujiyoshi, “Virtual Camerawork for Generating Lecture Video from High Resolution Images”, Proc. of IEEE ICME 2005, July, 2005.

[2] M. Smith and T. Kanade, “Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques”, Proc. of CVPR, 1997.

[3] 三浦宏一，浜田玲子，井手一郎，坂井修一，田中英彦, “動きに基づく料理映像の自動要約”, 情報処理学会コンピュータビジョンとイメージメディア研究会論文誌, vol.44, no.SIG9, pp.21-29, 2003.

[4] T. Liu, R. Hjelsvold and J. Kender, “Analysis and enhancement of videos of electronic slide presentations”, Proc of ICME’02, 2002.

[5] 石塚 健太郎, 亀田 能成, 美濃 導彦, “講義の自動撮影系における音声・映像インデキシング”, 電子情報通信学会 技術研究報告 PRMU, Vol.99, No.709, PRMU99-258, pp.91-98, 2000.

[6] 加藤大一郎，山田光穂，阿部一雄, “スタジオ番組における放送カメラマンのカメラワークと視線の動きの分析”, テレビジョン学会誌, Vol.49, No.8, pp.1023-1031,1995.

[7] 石川秋男, 加藤大一郎, 津田貴生, 福島宏, 下田茂, 阿部一雄, “放送カメラマンのズーミング計測法の検討と静止している被写体を撮影するときのズーミング解析”, 映像情報メディア学会誌, Vol.53, No.5, pp.749-757, 1999.

[8] 古井貞熙, “音声情報処理”，北森出版，　1998.

[9] A. Lipton, H. Fujiyoshi, and R. Patil, “Moving target classification and tracking from real-time video”, Proc. of WACV, IEEE, pp. 8-14, 1998.

[10] 西口敏司，仙田修司，美濃導彦，池田克夫, ‘首振りカメラによる黒板の記録手法”, 画像の認識・理解シンポジウム (MIRU’ 96) 講演論文集 I, pp. 37-42, 1996.

[11] Y. Boykov, M. Jolly, “Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images”, Proc. of ICCV, vol. I, pp. 105-112, 2001.