社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－CVIM－153 (28)
2006／3／17

講師行動の統計的性質に基づいた講義撮影のための講義状況の認識手法

杉本吉隆† 丸谷宜史† 角所考†† 美濃導彦††

† 京都大学 情報学研究科
†† 京都大学 学術情報メディアセンター

あらまし 講義を撮影した映像 (講義映像) を遠隔講義や講義アーカイブなどで利用する試みが行われるのに伴い, 講義を自動撮影する研究が行われている. 講義映像では講義内容を理解できるものである必要があるため, 自動撮影では適切な撮影対象を選択することが重要である. 本研究では, 講義内容を理解するために映すべき被写体を捉えた講義映像を作成することを考える. 講義内容を理解するために映すべき被写体の組合せを講義状況と定義すると, そのような講義映像作成を行うためには各時刻での講義状況を認識する必要がある. そこで本研究ではそのような講義状況の認識を目的とする. 従来の講義自動撮影の研究では講義状況として講師位置や講師行動を考え, 画像などの観測情報からそのような講義状況を認識してきた. この場合, 講師位置や講師行動は各時刻または決まった時間区間の観測情報から一意に認識されていた. しかしながら本研究で扱うような講義状況では, 同じ講義状況でもおきる講師行動や講師位置は様々であるため, 従来のように観測情報から一意に決定する手法では講義状況を高い精度で認識することが難しい. そこで本研究では講師行動の頻度, 講義状況間の遷移確率という統計的性質に着目し, この統計的性質を表現したHMM に基づく講義状況認識手法を提案する. 実験では実際に行われた講義を対象とし, 提案手法の有効性, 講義状況に基づいて作成した講義映像の有用性を確認する.

キーワード 講義自動撮影, 講義状況認識, 統計的性質, HMM

Lecture Context Recognition based on Statistical Feature of Lecturer Action for Automatic Video Recording

Yoshitaka SUGIMOTO†, Takafumi MARUTANI†, Koh KAKUSHO††, and Michihiko MINOH††

† Graduate School of Informatics, Kyoto Univ.
†† Academic Center for Computing and Media Studies, Kyoto Univ.

Abstract In this paper, we discuss the problem of recognizing situations of lectures for shooting lecture videos automatically. We classify the situations of the lectures based on the subjects to be gazed at for understanding the lectures in a lecture room. However it is not easy to recognize the lecture situations by conventional approaches based on the features of the sensory data obtained in a lecture room because the same sensory data occurs in the different lecture situations. We employ statistical properties for the occurrence of sensory data in each lecture situation together with the transition probability between lecture situations. This process is realized by Hidden Markov Model that represents those statistical properties.

Key words Automatic shooting video, Recognizing lecture situation, Statistical properties, Hidden Markov Model

1. はじめに

教育分野へのマルチメディア技術の普及に伴い, 講義の映像を撮影して遠隔講義や講義アーカイブなどで利用することが数多く試行されるようになってきた. これに伴い, このような講義映像の作成を省力化するために講義の自動撮影に関する研究が行われている.

講義映像は講義内容を理解できるものである必要があり, このような映像を撮影するためには講義の状況に応じて, 講師のみ, 講師とスライドなど適切に撮影対象が選択されることが重要である. 講義内容を理解できるためにどのような対象を撮影すべきかについては様々な提案があるが, いずれの場合でも撮影対象は講義の状況によって様々に変化することから, 講義状況の自動撮影のためには, 逆に講義理解に必要な対象が同一のものをそれぞれ講義状況と定義し, 各講義状況を認識する処理が必要となる. 本稿ではこのような講義映像の自動撮影のため

の講義状況認識について議論する.

まず2章では本研究で目的とする講義の自動撮影やそのための講義状況認識の具体的内容について議論する. 続く3章では2章で述べた講義状況認識を実現する上で考慮すべきデータの特性やそれに対する従来手法の問題点について, 実際の講義データの分析結果を示しながら述べる. さらに4章では3章の結果を踏まえ講義データの統計的性質を表現したHMMに基づく講義状況認識の手法を提案する. 5章では本手法の有効性を確認するために行った実験結果について示し, 6章でまとめと今後の課題について議論する.

2. 講義自動撮影のための講義状況

2.1 講義の形態

大学などの高等教育機関で行われる授業形態には講義・セミナー・実験・フィールドワークなどさまざまなものがあるが, 本研究ではこのうち講義を対象とする. 大学における講義の形態として次のようなものを考える.

(1) 一人の講師が多人数の生徒に対して一斉に講義内容を口述する.

(2) (1) において教材の提示が必要な場合にはPCによるスライドや板書が用いられる.

(3) (2) の教材を指し示すための指示棒が用いられる.

2.2 講義理解のための被写体

2.1節で述べた講義においては講師は講義内容について話すのに加え, 必要に応じて話している内容に関連する教材をスライドや板書によって提示する. したがってこのような講義内容を理解するためには, 話している主体である講師自身に加え, スライドや板書が提示されている状況では, それらも併せて撮影する必要がある.

本研究では, 講義室にいる学生が講義の理解のために視線を向ける可能性がある講義室内の対象である講師, スライド, 白板の集合を"被写体候補"と呼んで $\Omega = \{$ 講師, スライド, 白板 $\}$ で表す. さらに講義中の時刻 t においてそのときの講師の話している内容を理解するために捉えるべき被写体候補 Ω の部分集合を"焦点化被写体"と呼び, $\omega(t) \subset \Omega$ で表す. なお, 講義では話題の切り替わりや, 板書に用いるペンのインクが切れたといったようなハプニングなどによって, 講義が行われていない瞬間もしばしば生じ, このようなときの焦点化被写体は空集合 $(\omega(t) = \phi)$ となる.

講義の理解に必要な撮影すべき対象については従来から様々な提案があるが, 本研究では上の議論に基づき, 各時刻における焦点化被写体を画面に捉えた映像を撮影することを目指す. 焦点化被写体は理論上は被写体集合の任意の部分集合の数だけ存在するはずであるが, 実際の講義においてスライドと白板を完全に同時に提示して説明するといったことは起きない (両方を交互に提示することはありうる) といったことを考慮し, 本研究では焦点化被写体として次の4種類を考える.

(1) 「語りかけ」: 講師が特に教材を利用せずにジェスチャなどを交えながら説明している状況 $(\omega(t) = \omega_1 = \{$ 講師 $\})$

(2) 「スライド説明」: 講師がスライドを用いて教材を提示し, 適宜指示棒で指しながら説明している状況 $(\omega(t) = \omega_2 = \{$ 講師, スライド $\})$

(3) 「板書説明」 :講師が白板を用いて教材を提示し, 適宜板書したり指示棒で指しながら説明している状況 $(\omega(t) = \omega_3 = \{$ 講師, ホワイトボード (白板) $\})$

(4) 「説明無し」: 話題の切り替わりや, 板書に用いるペンのインクが切れたといったようなハプニングなど $(\omega(t) = \omega_4 = \phi)$

2.3 固定カメラによる撮影

講義の自動撮影のための手法には,PTZカメラ (カメラの向きや焦点をリモート操作で自由に変更できるカメラ) によって講師を追跡撮影するもの [1], [2] や, 複数の固定カメラを切り替えてスライドや講師を撮影するもの [3], [4] がある. 本研究では2.2節で述べたように講義理解のための映像として焦点化被写体を画面内に捉えるような撮影を目指すため, これが固定カメラの切り替えのみで可能か, あるいはPTZカメラによる追跡撮影が必要かを実際の講義データに基づいて調べた. 図1のような講義室で行われている実際の講義を対象として, そのときの講師の位置を講義室の前方付近に位置センサを設置することで獲得した. なお位置センサは超音波方式のものを用い, 超音波発信子を講師の両肩に装着して両肩の位置の中心を講師位置とし,0.5秒間隔で取得した.

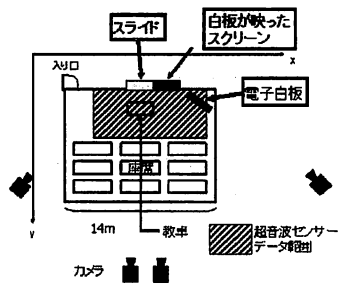

図1 講義室全体の概要

実験に用いたデータは3名分の講師によるそれぞれ90分の講義を各講師につき2回分収集し, 講義の各時刻の焦点化被写体が2.2節の $\omega_1, \ldots, \omega_4$ のいずれであるかを手作業で分類した. 図2はこのデータに基づいて, 焦点化被写体 $\omega_1, \omega_2, \omega_3$ のときの3名のうち1名の講師位置の分布を示したものである. 図2の各グラフの底面が図1のデータ範囲として示している網掛けの領域に相当する. 棒グラフの高さはその領域内の50cm×50cmの範囲を単位として, その範囲に滞在した時間の長さを表す. なお他の2名の講師に関しても同様の結果を得た. この結果より焦点化被写体が $\omega_1, \omega_2, \omega_3$ のそれぞれであるときの, 講師の位置にはばらつきが少ないことが分かる. このことから, 焦点化被写体が同じであるときはあらかじめ定めたカメ

ラパラメータによる固定カメラで, 焦点化被写体をカメラの画面内に捉えるような撮影が可能であるといえる. そこで本研究

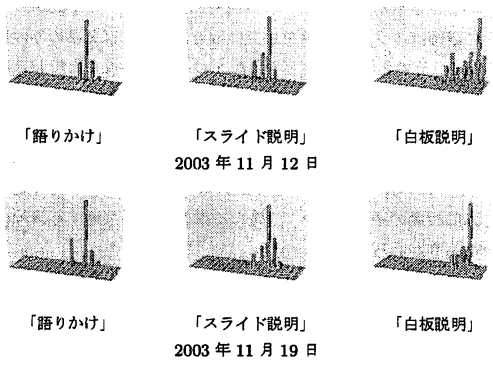

図 2　各講義状況での講師位置分布

では, 焦点化被写体が $\omega_1, \omega_2, \omega_3$ の場合ごとに, そのときの焦点化被写体を画面に捉えられるように専用の固定カメラ (3 台) を設置し, それぞれのカメラを切り替えて, 各時刻の焦点化被写体を捉えた映像を撮影することを目指す. なお,$\omega(t) = \omega_4$ の場合については講義の理解が必要でないため, 講義室の様子を捉えられるように講義室全体を撮影するものとする.

2.4　講義状況認識

2.3 節で述べたような固定カメラによる切り替え撮影のためには講義の各時刻 t において焦点化被写体 $\omega(t)$ が 2.3 節の $\omega_1, \ldots, \omega_4$ のうちのどれであるかを認識する必要がある. そこで本研究では,$\omega(t)$ が $\omega_1, \ldots, \omega_4$ のどれであるかを“講義状況”と呼ぶ.

従来から講義自動撮影のために様々な講義状況が定義されており, その認識のために用いられるセンサや, それによって得られる講義の観測情報にも様々なものが用いられてきた. 特に近年はセンサの低価格化により, 講義室に多数のセンサを設置することが非現実的ではなくなりつつあることから, 本研究では文献 [5] などと同様に, 表 1 のようなセンサを利用できるものとする. さらに指示棒の先端, 講師の両肩, 講師の両手首にそれぞれ位置センサをつけておくことで表 1 のようなセンサデータ特徴が利用できるものとする. そこで, このようなセンサデータ特徴から次のような規則で算出される 2 値の論理変数 (スライド指示 a_s, 白板指示 a_w, 白板が写ったスクリーン指示 a_d, 白板記入 a_k, ジェスチャ a_g, 発話 a_u) を考え, 本研究ではこのような観測情報を“講師行動”と呼ぶ. ただし, スライドに対応する 3 次元空間中の平面領域を S_{sl}, 白板に対応する 3 次元空間中の平面領域を S_{wh}, 白板が映ったスクリーンに対応する 3 次元空間中の平面領域を S_{whs} と表記する.

(1)　スライド指示 a_s

表 1　セ　ン　サ

センサ	データ特徴
マイク	音の有無 v $\{ v = 1$ (音あり),$v = 0$ (音なし)$\}$
超音波位置センサー	三次元位置 指示棒先端の三次元位置 $\boldsymbol{p_{bar}}$ 肩の高さ h_s 左右手首の高さ $h_{w,r}, h_{w,l}$
電子白板	ホワイトボードに記入中か否か w\{1:記入中 0:記入していない\}

$$a_s = \begin{cases} 1 \quad if \quad \exists \boldsymbol{p} \in S_{sl} \\ \qquad (\|\boldsymbol{p_{bar}} - \boldsymbol{p}\| < T) \\ \qquad \cap \ (\|\boldsymbol{p_{bar}} - \boldsymbol{p}\| < \min_{\boldsymbol{q}} \|\boldsymbol{p_{bar}} - \boldsymbol{q}\|) \\ \qquad \boldsymbol{q} \in \{S_{wh} \cup S_{whs}\} \\ 0 \quad otherwise \end{cases} \tag{1}$$

(2)　白板指示 a_w

$$a_w = \begin{cases} 1 \quad if \quad \exists \boldsymbol{p} \in S_{wh} \\ \qquad (\|\boldsymbol{p_{bar}} - \boldsymbol{p}\| < T) \\ \qquad \cap \ (\|\boldsymbol{p_{bar}} - \boldsymbol{p}\| < \min_{\boldsymbol{q}} \|\boldsymbol{p_{bar}} - \boldsymbol{q}\|) \\ \qquad \boldsymbol{q} \in \{S_{sl} \cup S_{whs}\} \\ 0 \quad otherwise \end{cases} \tag{2}$$

(3)　白板が映ったスクリーン指示 a_d

$$a_d = \begin{cases} 1 \quad if \quad \exists \boldsymbol{p} \in S_{whs} \\ \qquad (\|\boldsymbol{p_{bar}} - \boldsymbol{p}\| < T) \\ \qquad \cap \ (\|\boldsymbol{p_{bar}} - \boldsymbol{p}\| < \min_{\boldsymbol{q}} \|\boldsymbol{p_{bar}} - \boldsymbol{q}\|) \\ \qquad \boldsymbol{q} \in \{S_{sl} \cup S_{wh}\} \\ 0 \quad otherwise \end{cases} \tag{3}$$

(4)　白板記入 $a_k : w$

(5)　ジェスチャ a_g

$$a_g = \begin{cases} 1 \quad if \quad min\{|h_{w,r} - h_s|, |h_{w,l} - h_s|\} < T_G \\ \qquad \cap \ \neg(o_s \cap o_w \cap o_d) \\ 0 \quad otherwise \end{cases} \tag{4}$$

(6)　発話の有無 $a_u : v$

このようなセンサデータ特徴から実際に講師行動がどの程度正確に獲得できるのかを調べるために, 図 2 の 3 名の講師のうち

1名に対して,図2の講義とは別の講義に対して実験を行った.その結果,スライド指示は抽出率:97.3%,適合率:95.4%,ジェスチャも同様に抽出率:90.5%,適合率:74.3% で求められた.この結果より,表1のようなセンサーデータを用いて,講師行動が実際に高い精度で得られることが確認できた.以上より,本研究で目標とする2.3節のような映像の自動撮影のためには,講義における講師行動 (スライド指示 a_s, 白板指示 a_w, 白板が写ったスクリーン指示 a_d, 白板記入 a_k, ジェスチャ a_g, 発話 a_u) から,各時刻 t の4つの講義状況を認識することが目標となる.

3. 講義状況認識のためのアプローチ

3.1 講義状況における講師行動の多様性

従来の講義自動撮影に関する研究では,画像などの観測情報に基づいて講師の位置や行動を講義状況として認識することが行われてきた.石塚[6]らは講師位置と発話有無の組合せを講義状況と考え,カメラ画像とマイクからその講義状況を認識するための手法を提案している.そこではカメラ画像から講師の高さ制約などコンピュータビジョン技術を用いて講師位置を推定し,マイクから発話の有無を検出している.大西[3]らは講師の行動を講義状況と考え,カメラ画像からその講義状況を認識する手法を提案している.そこではカメラ画像からエッジ抽出や肌色抽出などの画像処理で体の動き,顔の向き,手の動き,板書文字の変化を求め,それらの組合せによって講師の行動を認識している.山口ら[1]も講師の行動を講義状況と考え,カメラ画像からその講義状況を認識する手法を提案している.そこではカメラ画像からエッジ抽出や背景差分などの画像処理で顔の向き,肩幅,腕の位置,手の動き,体の動きを獲得し,それらをもとに試行錯誤的に作成したファジールールによって講師の行動を推定している.先山ら[4]は典型的な行動系列を講義状況と考え,カメラ画像とマイクからその講義状況を認識する手法を提案している.そこではカメラ画像から背景差分やテンプレートマッチングなどの画像処理で講師の位置や顔の向きなどを獲得し,それらの組合せによって講師行動を獲得し,あらかじめ決めておいた講師行動の系列が獲得できればその行動系列が認識されていた.

これらの手法では各時刻における観測情報,または観測情報の決まった系列からその時刻の講義状況を認識するというアプローチが採られている.本研究ではこのようなアプローチを“決定論的アプローチ”と呼ぶ.決定論的アプローチによる講義状況認識が成功するには以下の条件が満足される必要がある.

(1) 各時刻,または決まった区間の観測情報が与えられれば,講義状況が一意に決まる

(2) 観測情報のあらかじめ決まった系列が与えられれば,講義状況が一意に決まる

従来のような具体性の高い講義状況を認識する場合にはこのような条件が成り立つことが多いため決定論的なアプローチが有効である.しかし本研究で目指すような,センサなどで観測可能な物理量との関係が明らかでない講義状況を認識する場合,同じ講義状況においても決まった講師行動が起きるとは限らず,同じ講義状況においても決まった講師行動の系列が起きるとは限らない.例えば「語りかけ」の時に講師はジェスチャをしているときもあれば発話のみの時もあり,「スライド説明」の時に講師はスライド指示をしているときもあれば発話のみの時もある.そのため本研究で考えた4つの講義状況ごとでおきる講師行動は様々で,異なる講義状況に共通する講師行動もある.また,「語りかけ」の時に発話の後にジェスチャがおきることもあれば,「スライド説明」の時に発話の後にジェスチャが起きるときもある.そのため本研究で考えた4つの講義状況で起きる講師行動の系列は様々で,異なる講義状況に共通する講師行動の系列がある.

このことを確認するために2.3節で利用したのと同じ講義データに対して各講義状況にどのような講師行動がおきるかを調べた.なおこの実験における講師行動は認識誤りの影響を避けるために2.4節の方針に従って手作業で与えた.このデータに基づいて,講師位置および講師行動と講義状況の関係を調べた.

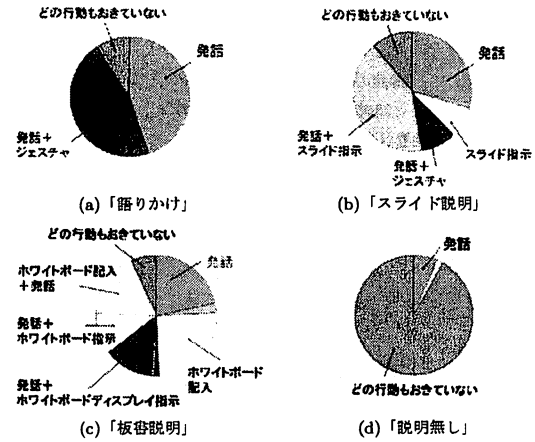

図3 講師行動頻度 (2003年11月12日)

3名のうちの1名の講師で,各講義状況で発生した講師行動の組 $o = \{a_s, a_w, a_d, a_k, a_g, a_u\}$ の頻度を調べた結果を図3に示す.図中の発話とは発話以外の講師行動がおきていないことで,図中の発話+ジェスチャとは発話とジェスチャという講師行動が同時におき,それ以外の講師行動はおきていないことを意味する.図3の(a)から講義状況「語りかけ」は発話が約40%,発話+ジェスチャが約40 % おきていることが読み取れる.図3中(a),(b),(c),(d)から特に発話,ジェスチャ,発話+ジェスチャ,行動無しなどは複数の講義状況で共通しておきていることが分かる.

また図2にあるように,例えば「語りかけ」と「スライド説明」に共通する位置が多く,「スライド説明」と「ホワイトボード説明」でさえ共通する位置が多い.これは他の講師でも同様であった.同じ講師位置であっても講義状況は異なっているた

め, 講師行動の変わりに講師位置を用いた場合でも, 同様に難しいと考えられる.

このことから本研究で目指すような講義状況認識は従来の決定論的なアプローチでは高い精度で認識することが難しいと考えられる.

3.2 講義状況と講師行動の統計的性質の利用

3.2.1 講師行動の生起頻度

3.1 節の図 3 では, 前述のように同じ講義状況に様々な講師行動の組が含まれるが, その一方, 各講義状況でおきる講師行動の組の頻度は異なっている. さらに図 3 と同じ講師の別の日の講義のデータについても調べてみたところ図 4 のようになった. これらのデータから, 「講義状況無し」は他の講義状況に比べて“発話”の頻度が小さく, 「語りかけ」では他の講義状況に比べて“ジェスチャ”の頻度が大きいといったように, 各講義状況で発生するそれぞれの講師行動には頻度に違いが大きいことが分かる. また, 図 3 中にある (a) と図 4 中の (e), 図 3 中 (b) と図 4 中 (f), 図 3 中 (c) と図 4 中 (g), 図 3 中 (d) と図 4 中 (h) を見比べると, 同じ講師であれば同じ講義状況でおきる各講師行動の頻度は異なる日でもほとんど変わらないことが分かる.

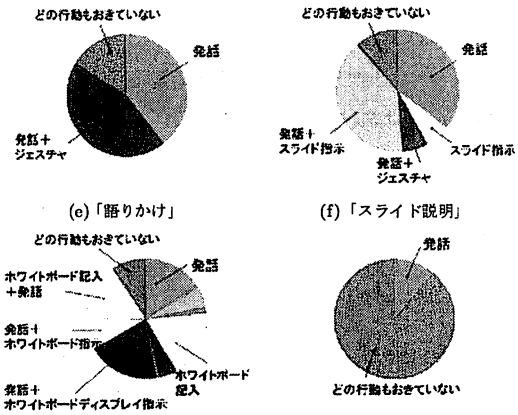

図 4 講師行動頻度 (2003 年 11 月 19 日)

このような特徴は他の講師の場合についてもほぼ同様であった. このような特徴から講師行動の生起頻度に基づいて講義状況が認識できる可能性がある.

3.2.2 講義状況間の遷移

講義はある講義内容をスライドや板書を教材として用いながら順序だてて説明されるものであるから, 講義の進行に伴って現れる講義状況の生起順序には何らかの依存関係があると考えられ, その依存関係の違いがそれぞれの講師の講義スタイルとなっていると考えられる. 例えば, 「スライド説明」中心に講義を進めながら「語りかけ」や「ホワイトボード説明」を適宜行う講義スタイルや, 「語りかけ」中心に講義を進めながら「スライド説明」や「ホワイトボード説明」を適宜行う講義スタイルなどが考えられる. そこでこのことを確認するために 3.1 節で利用したのと同じ講義データに対して講義状況間の遷移確率を調べた.

表 2 講義状況間の遷移確率 α_{ij}(2003 年 11 月 12 日 講師 B)

		講義状況 j			
		語りかけ	スライド説明	板書説明	説明無し
講義状況 i	語りかけ	96.55 %	1.19%	0.05 %	2.20 %
	スライド説明	1.03%	97.85%	0.18%	0.94%
	板書説明	0.56%	0.52%	98.16%	0.76%
	説明無し	5.45 %	2.99 %	0.56 %	91.00%

表 3 講義状況間の遷移確率 α_{ij}(2003 年 11 月 19 日 講師 B)

		講義状況 j			
		語りかけ	スライド説明	板書説明	説明無し
講義状況 i	語りかけ	96.33%	0.95%	0.08%	2.64%
	スライド説明	0.68%	98.08%	0.09%	1.14%
	板書説明	0.27%	0.50%	98.32%	0.91%
	説明無し	4.02%	2.02%	0.38 %	93.58%

表 4 講義状況間の遷移確率 α_{ij}(2003 年 10 月 1 日 講師 A)

		講義状況 j			
		語りかけ	スライド説明	板書説明	説明無し
講義状況 i	語りかけ	97.48 %	0.28%	0%	2.24 %
	スライド説明	1.62%	96.44%	0.03%	1.92%
	板書説明	0%	0%	92.3%	7.69%
	説明無し	7.28 %	0.86 %	0 %	91.86 %

講師 A,B の講義状況間の遷移確率を調べてみた結果を表 2〜表 4 に示す. それぞれの表では確率は 0.5 秒を単位として講義状況が自己遷移を含めて次の講義状況へ遷移するときの遷移確率を示している. ここで講義状況 i から講義状況 j への遷移確率を $\alpha_{i,j}$ と書くと,

$$\alpha_{i,j} = \frac{\text{講義状況 } i \text{ の次に講義状況 } j \text{ に遷移した回数}}{\text{講義状況 } i \text{ の回数}}$$

表では縦が i, 横が j をあらわしている. 例えば表 2 では「講義状況無し」の次に「スライド説明」である確率は 2.99 % であることを表す. これらの表から, 各講義状況からそれぞれ異なった遷移確率で別の講義状況へ遷移しており, かつ同じ講師であれば別の日の講義でも遷移確率は似ていることが確認できる. 特に 1:2 程度であるのに対して表 4 では 1:9 程度になっており, このような遷移確率の違いは講師 B が「スライド説明」中心に講義を進めているのに対して講師 A は「語りかけ」中心に講義を進めていることを表している.

このような性質から, 講義状況の遷移確率を講義状況の認識に利用可能であると考えられる. 以上から本研究では講義状況の遷移確率と各講義状況での講師行動の生起確率という 2 つの統計的性質を用いて講義状況を認識することを考える. ここで, 時系列データに対してその時系列データのカテゴリを認識する手法として隠れマルコフモデル (HMM) がある. そこで本研究では講義状況の遷移確率を内部状態の遷移確率, 講師行動の発

生頻度を各内部状態での出力確率に対応させたHMMを用いて講師行動から講義状況の認識手法について次に述べる.

4. 講義状況認識のためのHMMの利用

HMMは内部状態と外部信号があり,内部状態間の遷移確率と外部信号の出力確率を持つモデルである.本研究では図5のようにHMMの内部状態に講義状況を,HMMの外部信号に講師行動を対応付ける.このとき講義状況間の遷移確率は内部状態間の遷移確率に,各講義状況における講師行動の生起頻度は内部状態における外部信号の出力確率に対応することになる.3.2節で述べたようなどの講義状況からも異なった遷移確率で次の講義状況へ遷移するという性質から,このようなHMMでは各内部状態からそれぞれ異なった遷移確率で次の内部状態へ遷移することになる.同様に3.2節で述べたような各講義状況で発生する講師行動の頻度が異なるという性質から,HMMでは各内部状態が出力する外部信号が異なることになる.さらに本研究において講師行動から講義状況は決定論的に決まらないという性質は,対応付けされたHMMでは外部信号から内部状態が特定できないということに対応する.

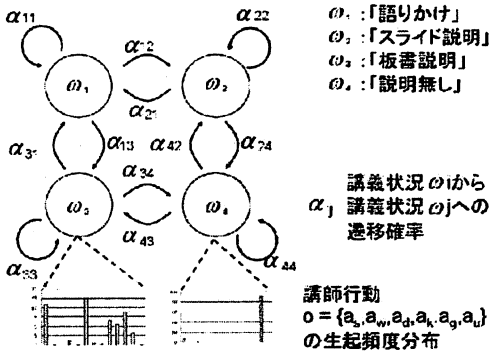

図5 講義状況と講師行動を対応付けたHMM

4.1 HMMのパラメータ学習

$F(S_iS_j)$ を内部状態 S_i から内部状態 S_j へ遷移が起こった回数を表すとする.$E(S_i,o_a)$ を内部状態 S_i で外部信号 o_a を出力した回数を表すとする.また内部状態の数を M とする.通常のHMMの学習では,内部状態が観測できないため,α,β の学習にはEMアルゴリズムなどが用いられるが,本研究では内部状態が講義状況として直接観測できるので,次のように直接計算できる[7].

$$\alpha_{ij}=\frac{F(S_iS_j)}{\sum_{m=0}^{M}F(S_iS_m)}$$

$$\beta_i(o_a)=\frac{E(S_i,o_a)}{\sum_{o_m}E(S_i,o_m)}$$

4.2 講義状況の認識

4.2.1 オンライン処理のための認識手法

講義状況に基づいた講義映像を遠隔講義で利用する際,講義状況の認識はリアルタイム性が求められる.そのため遠隔講義を目的としたオンライン処理での講義状況認識とは,時刻 t までの出力信号系列から時刻 t における最も確率の高い内部状態を求めることであると考えられる.出力信号系列長を L,2.4節で定義した講師行動の組 $o=\{a_s,a_w,a_d,a_k,a_g,a_u\}$ が各時刻ごとに並んだ出力信号系列 $O_t=\{o_1,o_2,\ldots,o_t\}$,HMM内部状態間の遷移確率を α,外部信号の出力確率を β とする.また求めたい時刻 t の内部状態を $s_{t,online}$,ただし $s_{t,online}\in\{\omega_1,\ldots,\omega_4\}$ とする.求めたい $s_{t,online}$ を定式化すると以下のようになる.

$$s_{t,online}=\arg\max_{\omega_m}P(\omega_m|O_t,\alpha,\beta)$$

これを変形すると

$$s_{t,online}=\arg\max_{\omega_m}\frac{P(\omega_m,O_t|\alpha,\beta)}{P(O_t|\alpha,\beta)}$$

$P(O_t|\alpha,\beta)$ は S_m に関係ないので

$$s_{t,online}=\arg\max_{\omega_m}P(\omega_m,O_t|\alpha,\beta)$$

となる.これによって得られた内部状態が講義状況に対応するためそれを認識結果とする.

4.2.2 オフライン処理のための認識手法

講義状況に基づいた講義映像を講義アーカイブで利用する際は必ずしもリアルタイム性は必要がない.つまり講義アーカイブを目的としたオフラインでの講義状況認識の場合は,全時刻でのデータを考慮した上で認識が可能である.またHMMのパラメータは学習によって求まっているため,オフライン処理における講義状況認識は,このような条件の下で確率が最も高い内部状態系列を求める処理であると考えられる.すなわち出力信号系列長を L,2.4節で定義した講師行動の組 $o=\{a_s,a_w,a_d,a_k,a_g,a_u\}$ が各時刻ごとに並んだ出力信号系列 $O=\{o_1,o_2,\ldots,o_L\}$,HMM内部状態間の遷移確率を α,外部信号の出力確率を β とする.任意の内部状態系列を $S=\{s_1,s_2,\ldots,s_L\}$ ただし $s_i\in\{\omega_1,\ldots,\omega_4\}$,求めたい内部状態系列を $S_{offline}=\{s_{1,offline},s_{2,offline},\ldots,s_{L,offline}\}$,ただし $s_{i,offline}\in\{\omega_1,\ldots,\omega_4\}$ とする.求めたい $S_{offline}$ を定式化すると以下のようになる.

$$S_{offline}=\arg\max_{S}P(S|O,\alpha,\beta)$$

これによって得られた内部状態系列は講義状況系列に対応するためそれを認識結果とした.

5. 実　　験

5.1 実験環境

本手法の有効性を確認するために,実際の講義を対象として実験を行った.実験は図1の講義室で2005年後期に実際に行われた講義で,講師A:3回(11月2日,11月9日,11月16日),講師B:3回(11月30日,12月7日,12月14日)の講師2名,計6回分の講義を対象とすることにする.実験では映像を見て人手で判断した講義状況を正解とした.ただし,人間が見ても講義状況が曖昧な部分は4つの講義状況と異なるフラグをつけ評価からは除外した.

5.2 提案手法と従来の決定論的手法

提案手法の有効性を確認するために実験を行う．ただし実験ではテストと同じ講師で他の2回分の講義で学習し，そのパラメータを用いる．また，オンライン認識は $\arg\max_{\omega_m} P(\omega_m, O_t|\alpha,\beta)$ を forward アルゴリズム [7] を用いて計算し，オフライン認識では $\arg\max_S P(S|O,\alpha,\beta)$ を viterbi アルゴリズム [7] を用いて計算した．6回の講義における各講義状況の抽出率，適合率を提案手法，従来手法のそれぞれで算出し，その値を比較した．その実験結果を表5に示す．ただし実験結果は紙面の都合上，講師Aの結果は11月2日,11月9日,11月16日の3回分の講義をテストした結果の平均，講師Bの結果は11月30日,12月7日,12月14日の3回分の講義をテストした結果の平均とした．また講義状況 i の抽出率と適合率はデータの間隔である0.5秒ごとに講義状況が起きたと考え，以下のように算出した．

$$\text{抽出率} = \frac{100 \times (\text{認識結果が正しい講義状況と一致した個数})}{(\text{正しい講義状況 } i \text{ の個数})}$$

$$\text{適合率} = \frac{100 \times (\text{認識結果が正しい講義状況と一致した個数})}{(\text{講義状況 } i \text{ と認識した個数})}$$

表5にあるようにオフライン認識での各講義状況の抽出率，適合率は90% 前後，オンライン認識での抽出率，適合率が85% 前後となり良好な結果が得られたと考えられる．オンライン認識とオフライン認識を比較すると表5にあるように，オンライン認識はオフライン認識に比べて各講義状況の抽出率が5% ～ 10% 程度低下している．この原因は本研究の認識手法は統計量を用いた認識手法であり，オンライン認識では未来のデータは未知であるため，講義状況が変化した時に新しい講義状況に変化したと認識するのが遅れるためである．

表5 提案手法の認識率および従来手法の認識率 (講義三回平均)

	語りかけ		スライド説明		板書説明	
	抽出率	適合率	抽出率	適合率	抽出率	適合率
提案手法 オンライン認識 講師A	89.5 %	90.6%	87.1%	81.9%	86.0%	95.1%
提案手法 オフライン認識 講師A	93.1%	95.5%	94.0%	88.2%	90.2%	97.5%
従来手法 講師A	71.4 %	90.1%	74.2%	69.0%	77.8%	89.7%
提案手法 オンライン認識 講師B	83.4 %	73.8%	89.4%	94.6%	72.2%	67.9%
提案手法 オフライン認識 講師B	93.5%	87.9%	93.8%	98.0%	86.6%	72.2%
従来手法 講師B	74.3 %	39.0%	39.7%	90.1%	45.8%	60.3%

次に本手法で述べた統計的性質を用いた認識手法が従来用いられていた決定論的な手法に比べ，本研究で目指す講義状況認識にどれだけ有効であるかを検証する．ただし従来研究は講師位置，講師の向き，講師の行動を主な特徴量としそれに基づいて決定論的な手法で講義状況を認識していたため，本研究では従来手法として次の手法を想定する．

(1) 特徴量：従来の代表的な特徴量である講師位置，講師の向き，講師行動を特徴量とする．

(2) 認識手法：従来の決定論的な規則に相当する手法として，パターン認識の基本的な決定論的手法である NN 法を用いる．

また従来手法も提案手法と同じ条件で学習したパラメータを用いた．提案手法と従来手法を比較すると表5中の講師Bでは，オフライン認識と従来手法の抽出率の差が40% 近く出ている．それに対して表5中の講師Aでは，講師Bよりも差が小さく20% 程度の差となっている．これは講師Aが各講義状況で存在する位置があまり共通していないためであると考えられる．しかし講師A，講師Bともにどの講義状況においても提案手法 (オンライン認識，オフライン認識) のほうが抽出率，適合率ともに高いため提案手法の有効が確認された．

5.3 同じ講師で学習した場合と異なる講師で学習した場合の比較

3.2節で述べたように講師ごとに講義スタイルがあると考え，5.2節における実験では HMM のパラメータを求める学習を講師ごとに行った．本節ではそのことを検証するために，認識する講義の講師と同じ講師の別の日の講義で学習した場合 (同じ講師で学習と呼ぶ) と，認識する講義の講師とは異なる講師の別の日の講義で学習した場合 (異なる講師で学習と呼ぶ) を考え，両者による認識の差を調べる実験を行った．同じ講師で学習した場合と，異なる講師で学習した場合を比較した結果が表6である．

表6 同じ講師で学習する場合と異なる講師で学習する場合の比較 (3回講義平均)

	語りかけ		スライド説明		板書説明	
	抽出率	適合率	抽出率	適合率	抽出率	適合率
テスト:講師A 学習:講師A	93.1%	95.5%	94.0%	88.2%	90.2%	97.5%
テスト:講師A 学習:講師B	97.8 %	90.8%	79.2%	98.6%	94.1%	97.5%
テスト:講師B 学習:講師B	93.5%	87.9%	93.8%	98.0%	86.6%	72.2%
テスト:講師B 学習:講師A	87.5 %	84.0%	95.0%	93.3%	27.2%	98.5%

表6中の講師Aで学習し講師Bの講義を認識する場合の「ホワイトボード説明」の抽出率を見ると27.2% とほとんど抽出できていないことが分かる．これは講師Bがホワイトボードが映ったディスプレイ指示を頻繁に行うのに対して，講師Aはホワイトボード直接指示でホワイトボード説明を行う．そのため，講師Aで学習しても「ホワイトボード説明」でホワイトボード

が映ったディスプレイ指示という講師行動が学習されないために認識が出来ていない.

表 6 中の「スライド説明」の抽出率が講師 B で学習して講師 A の講義を認識したとき, 講師 A で学習して講師 A の講義を認識した時に比べ 15% 近く低下している. 講師 B の講義では「語りかけ」の時に指示棒を直線にたとえる, 指示棒を動かして移動にたとえるというようなことがみられ,「語りかけ」のときにスライド指示が起きることがある. それに対して講師 A は語りかけのときにスライド指示がほとんど起きない. そのため講師 B で学習して講師 A の講義を認識すると, スライド説明の抽出率が低下することになる. この講師行動の違いが講師 A で学習して講師 B の講義を認識するときにも現れる. それは表 6 中にある「スライド説明」の適合率が 5% 程度低下していること,「語りかけ」の抽出率が 5% 程度低下していることである.「語りかけ」の時に起きる指示棒を使った例示がスライド指示になってしまい, それを講師 A で学習した場合は「スライド説明」と誤認識してしまっているためである.

以上のように異なる講師で学習すると. 講師による講師行動の発生頻度の違いから, 一部の講義状況で抽出率, 適合率が著しく低下する場合がある. そのため, 同じ講師で学習するほうが望ましいと考えられる. このことは逆に学習した HMM を用いて講義状況だけでなく講師の認識が可能となることを期待させるものとも言える.

5.4 講義状況に基づいた映像作成

本研究で考えた講義状況の有効性を確認するために, そのような講義状況に基づいて作成した講義映像を従来の講師行動に基づいて作成した講義映像と比較するで有用性を確かめるための実際の講義を対象とした実験を行った. 提案手法による講義

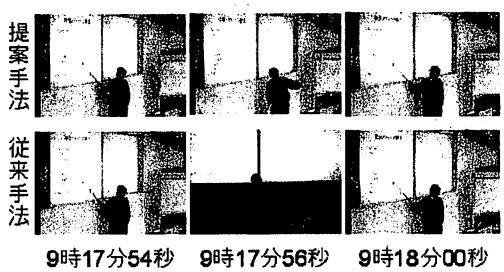

図 6 講義映像

状況の認識結果に基づいて作成した講義映像の結果と従来の講師行動に基づいて作成した講義映像の結果の一部を図 6 に示す. 図中の 9 時 17 分 54 秒から 9 時 18 分 00 秒までの間, 講師はスライドを説明しており講義状況は「スライド説明」の場面である. 本研究で考えた講義状況に基づく講義映像では, 講義理解に必要な対象であるスライドと講師が映った映像が作成されている. しかしながら従来の講師行動に基づく講義映像では, 9 時 17 分 56 秒のところではスライドが映っていない講義映像となる. このような講義映像は例えば映像中のスライド文字を読んでいる視聴者が, 話している内容に変化が無いにもかかわらず読むのを中断される場合が考えられる. このような講義理解に必要な被写体が映っていないという問題点に加え, 講師の行動は頻繁に変化するので画面が頻繁に切り替わり見にくいということも問題として考えられる. このようなことから本研究で考えた講義状況に基づいて作成された講義映像は有用性があると考えられる. そのため本研究で考えた講義状況の有効性が確認された.

6. まとめと今後の課題

本研究では講義の自動撮影を対象とする. 講師の話している内容を理解するために捉えるべき対象を“焦点化被写体”と定義し, そのような焦点化被写体を画面上に捉えた講義映像の作成を実現することを目的とする. それを実現するために焦点化被写体に対応して“講義状況”を考え, センサから取得できる講師行動から講義状況を認識するための手法について提案した. 従来は講師位置や講師行動を講義状況と考え, 画像などのセンサから決定論的に認識していた. このとき各講義状でおきる講師行動が多様であることから, 従来のような決定論的なアプローチではなく, 講義状況間の遷移確率, 各講義状況でおきる講師行動の頻度という統計的な性質を用いて講義状況を認識する手法を提案した. 実験では, 提案手法の認識率が 90% 程度と従来の決定論的な手法に比べて十分に高い結果が得られ, 提案手法の有効性が確認した. また本研究で考えた講義状況に基づいて講義映像を実際に作成し従来の講師行動に基づいて作成した講義映像に比べて有用な点があることを確かめることで, 本研究で考えた講義状況の有効性についても確認した.

本研究において講義の分析を進める中で, スライドが図の場合, 文字だけのスライドに比べて講師が指示することが多いことが分かってきた. このようなことから, 講師行動の頻度は講義状況だけでなくスライドの種類にも依存すると考えられるため, 講師の行動頻度や講義状況の遷移確率の精度をより高めるために, スライドの種類も考慮に入れることも考えられる. また実験によって可能性が示されたように, 講師ごとに講義スタイルが異なることから講義状況の認識だけでなく講師の認識についても考えられる.

文 献

[1] 山口達, 吉川大弘, 篠木剛, 鶴岡信治, “講師の動作認識に基づいた遠隔授業映像の自動撮影”, PRMU, pp.149-156, Jan.2001
[2] 島田敬士, 菅谷明, 谷口倫一朗, “講義中の教師の動作に基づく説明対象の抽出”, 画像の認識・理解シンポジウム (MIRU), vol.2, pp.353-358, Jul. 2004.
[3] 大西正輝, 村上昌史, 福永邦雄, “状況理解と映像評価に基づく講義の知的自動撮影”, 信学論 D-II, Vol.J85, N0.4, pp.594-603, Apr. 2002.
[4] 先山卓朗, 大野直樹, 椋木雅之, 池田克夫, “遠隔講義における講義状況に応じた送信映像選択”, 信学論 D-II, Vol.J84, No.2, pp.248-257, Feb. 2001.
[5] 西口 敏司, 亀田 能成, 角所 考, 美濃 導彦 “大学における実運用のための講義自動アーカイブシステムの開発”, 信学論 D-II, Vol.J88-D-II, No.3, pp.530-540, 2005.
[6] 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.
[7] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄 編著 “音声認識システム”, オーム社, 2001.