6ZA-01

ビデオ講義における映像・音声情報を用いた受講生の視線予測モデル

植木 康介 † 川嶋 宏彰 † 松山 隆司 †

† 京都大学 大学院情報学研究科

1. 視線に基づく内部状態推定

MOOCs (Massive Open Online Courses) に代表される講義映像視聴型の個別学習が近年急速に普及し，多様かつ質の高い講義コンテンツを誰もが視聴できるようになった．一方で，映像視聴型の学習は一方向的になりがちであるため，映像視聴中の受講生の状態を，観測される振る舞いからリアルタイムに推定し，理解が十分でなかった箇所を支援するシステムの必要性が高まっている．

このとき，視線は学習者の状態を知る上で重要な情報であり，文章読解中の視線の動きから受講生の習熟度や集中度といった内部状態を調べる様々な研究が行われている [1, 2]．さらに，映像コンテンツと視線の動きとの関係から，視聴者の集中状態を推定する研究がある [3]．

しかし，講義映像視聴時においては，受講生はコンテンツの映像・画像情報だけでなく，講師の説明音声情報を受け取りながら，さらに自身の注意状態も時々刻々変化させる．このような複雑な状況，すなわちマルチモーダルな入力と内部的な注意の変化が視線の動きに反映される状況において，受講生の視線の振る舞いはいかにパターン化・モデル化することができるだろうか．

本研究では，受講生がどの程度講義映像に追従もしくは集中しているかといった状態を，あらかじめ構築した標準的な視線パターンモデルからの逸脱により認識するアプローチを目指す．そこで本稿ではその第一ステップとして，講義映像に含まれる映像情報および音声情報が与えられた時に，受講生の典型的な視線の振る舞いを予測できるような確率モデルの構築法を検討する．

2. 視線の確率的予測モデル

個別学習で用いられる一般的な講義映像には，講師が黒板やホワイトボードの前で講義を行っている様子を撮影したものや，スライドのみが写っており，マウスやペン，アニメーションによって情報が追加されるもの，講師の映像が一部に重畳されているものなど，様々なタイプがある．本研究では簡略化のために，講義映像は図表のないスライドベースのコンテンツとし，映像に講師は映っておらず，文字だけで構成されているものとする．講義はスライドの上から下の内容へと順番に進んでいき，講師の話がスライドに書かれている内容から脱線することはない場合を考える．また，受講生はテーブル上のディスプレイの前に座ってビデオ講義を受講する状況を想定し，講義映像以外の参考書やノートは持っておらず，講義映像のみに注目しているとする．

2.1 注意領域の確率的遷移モデル

受講生の視線は，講師が話している場所に注意を向けるような視線パターンが多い一方で，スライド全体を見るという視線パターンやスライドが切り替わりの前後で全体を見渡す動きも見られ，これらの状態が時間とともに移り変わる．ここでは受講生の「状態」をスライドのどの領域（もしくはその集合）に注意が向けられているかとしてモデル化し，各状態によってどの領域が注視されるかという傾向が変わるものとする．

このような，外部から観測できない状態を扱う数理モデルとして，音声認識等でもしばしば用いられている隠れマルコフモデル (Hidden Markov Model, HMM) がある．そこで本研究では，確率状態遷移モデルとして HMM を用い，音声情報から得られる講義の文脈情報を入力できるものへ拡張する．また，HMM の出力にはスライド上の領域を割り当てるものとし，出力が離散的記号となる離散 HMM（以下，単に HMM）を用いる．このとき，離散 HMM は 5 項組 $M = (Q, R, A, B, \pi)$ で定義される．

- $Q = \{q_1, \ldots, q_N\}$：状態の有限集合
- $R = \{1, \ldots, D\}$：コンテンツ上の領域集合
- $A = \{A_{ij} | i, j = 1, \ldots, N\}$：状態遷移確率分布
  A_{ij}は状態 q_iから状態 q_jへの遷移確率
- $B = \{B_i(d) | i = 1, \ldots, N, d \in R\}$：記号出力確率分布
  $B_i(d)$ は状態 q_iで記号 $d \in R$ を出力する確率
- $\pi = \{\pi_i | i = 1, \ldots, N\}$：初期状態確率分布
  π_iは状態 q_iが初期状態である確率

ここで，時刻 t は様々な定め方がありうるが，本研究では注視の切り替わり，すなわちフィクセーションの出現順序を表す添え字 $t \in T_s = (1, ..., T)$ とする．

2.2 音声文脈依存型隠れマルコフモデル

ビデオ講義の受講時において，音声の文脈情報が受講生の状態に影響を与えると想定できる．そこで，受講生の状態の遷移確率が講師の発話内容によって切り替わるように，HMM を以下のような文脈依存型のモデルに拡張する．まず，講師はスライドに表示されている内容についてのみ話すと仮定する．つまり，講師の発話内容は文脈ラベル集合 $\mathcal{P}$（非発話ラベルを含む）のいずれかに対応するとし，状態遷移確率が時刻 $t \in T_s$ での講師の発話内容 $a_t \in \mathcal{P}$ の関数 $A_{ij}(a_t)$ として表されるとする．ただし $A_{ij}(a_t)$ は $\sum_{j=1}^{N} A_{ij}(a_t) = 1$ を満たす．

HMM の出力確率は，スライド上の近い行に偏るとする．つまり，出力記号（コンテンツ上の領域）をスライドの行に対応させ，状態は「1 番目から 3 番目の行に注意している状態」や「4 番目から 6 番目の行に注意している状態」というような意味を持たせる．さらには，受講生の注意はスライドの上の端から下の端へと大きく飛ぶことは少なく，状態の遷移はすべての状態間で起こるわけではないと考えられる．そこで，状態 q_i は状態 q_{i-1} か q_{i+1} にのみ遷移するとし，それ以外には遷移しないと

する．ただし状態には「スライド全体を見回している状態」という場合も考えられる．この状態を q_N とし，状態 q_N との遷移はどの状態とでも起こるとする．状態 q_N にあるときは外因性の刺激により影響されやすいと考えられるため，スライドの各点の顕著度合いによって出力確率 $B_N(d)(d \in R)$ が定まるとする．

本研究では，あらかじめ同じ講義映像を視聴した受講生の視線データが利用可能であるとし，HMM のパラメータ A，B（B_N を除く），π は，Expectation Maximization (EM) アルゴリズムにより学習する．ただし EM アルゴリズムは初期値依存性があるため，以下で述べる実験では，初期値はランダムに 50 回与え，最も高い尤度を与えるモデルを選択した．

3. 視線予測モデルの評価実験

3.1 実験手順

実験参加者に提示する講義映像として，情報セキュリティに関する 10 分程度の講義映像（スライド 6 枚，音声による解説あり）を用意した．それぞれのスライドは，アニメーションやポインタの有無などによって異なるが，本稿では特にアニメーションやポインタの無いスライド（以下，第 1 スライドと呼ぶ）での結果を報告する．

受講生にはディスプレイに向かうように配置した椅子に着席してもらった．受講生とディスプレイとの距離は 850mm に設定し，より安定した視線計測のため顎台を使用して頭部を固定した．視線の計測には受講生の眼下に視線計測装置 (Tobii X120) を設置し，これによりディスプレイ平面上の注視位置を得た．なお，サンプリング周期は $t_s = 1/60 = 16.7$ms に設定して行った．また，受講生には全員集中して講義を受けてもらうことを前提とするため，「映像視聴後に講義内容に関する問題を解いてもらう」という教示を与えて映像を見てもらい，受講後に講義内容に関する問題を解いてもらった．

フィクセーションについては，本研究では 400 ピクセル/秒（11.6deg/秒）以上の変動があった場合をサッケードとみなした．視線データには計測に失敗したデータも含まれる．これらは欠損データとして扱い，そこでフィクセーションはとぎれる．しかし欠損データには瞬きも含まれており，瞬きの前後でサッケードが生じていないなら合わせてひとつのフィクセーション区間として扱いたい．そこで短時間の欠損についてはその前後のデータの平均を取って補正することで注視系列を得た．HMM の状態数 N は，大きいほど学習データに対する予測精度は増すが，学習データに過度に適合するオーバーフィッティング（過学習）の問題を避けるために，経験的に $N = 5$（段落数 4+スライド全体を見渡す状態）とした．

3.2 結果と考察

図 1 はモデルの予測性能を，予測と実際の注視系列との一致度を測る Normalized Scanpath Saliency（NSS）によって評価した結果である（leave-one-out による open test）．顕著性マップのみ（$B_N(d)$ による予測），もしくは音声の文脈ラベルから対応する行を直接注視領域として予測した場合と比較して提案モデルの予測精度が高い

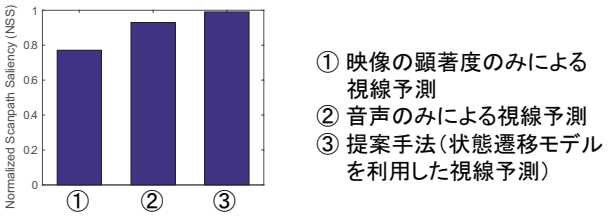

図 1: 視線の予測精度の比較

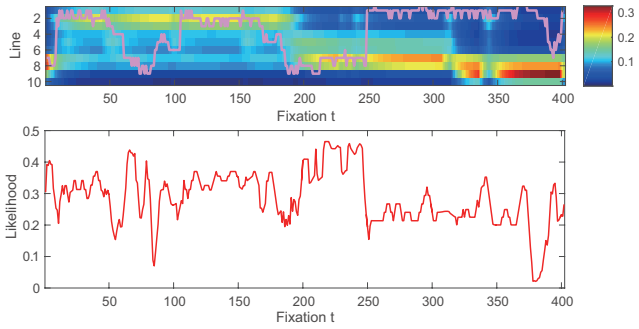

図 2: 上段：受講生 A の注視系列（紫線）および HMM で予測された標準的パターン（縦軸: 行，横軸: t，色: 確率），下段：受講生 A の注視系列に対する尤度の推移

ことが確認できる．

ここで，受講生のうち 1 名（受講生 A）については，後半でタイトル行を注視するという特徴的な視線パターンであった（図 2（上段）の紫線）．そこで，この 1 名分のデータを除いて学習された HMM を用いて予測した注視系列を，図 2（上段）のカラーマップで示す．確率の高さが色相で表されており，縦軸はスライド中の行を表す．これより視線が上から下へと移動していく様子が観測でき，このスライドを視聴する際の標準的な注視系列を予測しているといえる．一方で，受講生 A は $t = 250$ 以降でこの標準的パターンから特に逸脱していることも確認できる．受講生 A の注視系列に対する尤度変化を図 2（下段）に示す（窓サイズ 5 の移動平均で平滑化を行っている）．後半の特徴的な注視を行っている時間帯では特に尤度が低くなることが分かる．

以上の予備的実験より，本研究で提案した予測手法は，講義映像視聴時の標準的な視線パターンのモデルや，そこから逸脱するような視聴状態を検出する際に有効であることが示唆される．今後はより多くのデータで実験を行うことで，集中度やコンテンツへの追従度といった指標の定量化が可能であるかどうかを検証する予定である．

謝辞 本研究の一部は JST さきがけの補助を受けて行った．

参考文献

[1] 藤好 宏樹, 吉村 和代, K. Kunze, 黄瀬 浩一. 英文問題解答時の視点情報を用いた英語能力推定法. **信学技法**, volume PRMU2015-10, 2015.

[2] M. Rodrigue, J. Son, B. Giesbrecht, M. Turk, T. Höllerer. Spatio-Temporal Detection of Divided Attention in Reading Applications Using EEG and Eye Tracking. In *IUI*, pages 121–125. 2015.

[3] 米谷竜, 川嶋宏彰, 加藤丈和, 松山隆司. 映像の顕著性変動モデルを用いた視聴者の集中状態推定In **電子情報通信学会論文誌**, volume J96-D, number 8, pages 1675–1683, 2013.