講義における履修生の状態認識システムの試作と評価

Prototyping and evaluation of a students' state recognition system in lectures for active learning

松田 晃一
[1]大妻女子大学 社会情報学部

Kouichi Matsuda[1]
[1] Faculty of Social Information Studies, Otsuma Women's University
12 Sanbancho, Chiyoda-ku, Tokyo 102-8357, Japan

キーワード：アクティブラーニング，講義，履修生状態認識，深層学習
Key words：Active learning, Student's state recognition, Deep learning

抄録

近年重要度が増しているアクティブラーニングは，学生と教員の双方向性や学生らの能動的な学習を促進するといったスタイルの講義形式であるが，このような講義では，学生らの状態を教員やTA (Teaching Assistant) が把握できることが重要である．しかしながら，履修人数多い講義の場合，教員とTAだけでは，講義を行いながら，個々の学生の状態を適宜把握することは難しい．そこで本研究では，近年発展が著しい人工知能(AI)や，関連技術である深層学習，コンピュータビジョンなどを活用し，学生らの状況をリアルタイムで把握する基盤システムを構築，評価した．具体的には，着席した学生らの机上における動作に着目し，資料閲覧，スマートフォン操作，居眠りといった状態を識別するためのネットワークモデルを深層学習の手法を用いることで約4万枚の画像から獲得した．

1.　はじめに

近年，講義形態におけるアクティブラーニングの重要性が増している．これは教員から学生への一方的な講義ではなく教員と学生との双方向性や学生らの能動的な学習を促進するというスタイルである．このようなスタイルの講義では，学生らの状態を教員やTAが把握し，状況に応じて学生とのやり取りや，講義の進行速度や内容を調整することが望ましい．

しかしながら，1名の教員と1名程度のTAに対して履修生が多数という形態の講義では，講義中に，個々の学生の状態を常時把握することは難しい．

一方，カメラの高解像度化やコンピュータの高性能化に伴い，コンピュータがリアルタイムに外界の状況を認識することができるようになってきている．これらを講義で用いれば，教員やTAに替わって学生らの状態を把握し，それを元に教員やTAの行動を支援することで質の高い教育効果を上げられることが期待できる．

そこで本研究では，学生らの状況をリアルタイムに把握するため，着席した学生らの机上における動作に着目し，資料閲覧，スマートフォン操作，居眠りといった状態を認識するために，深層学習の手法により状態を推定する基盤システムを構築，評価した．本稿では，関連研究，想定環境，システムの概要，精度向上の工夫，まとめと今後の課題について述べる．

2.　関連研究

アクティブラーニングの補助システムに関する研究として，西野らは教室内に固定されたカメラで撮影された講義映像から移動物体を自動的に検出する手法を提案した[1]．西野らは，FAST(Features from Accelerated Segment Test)やSIFT　(Scale-Invariant Feature Transform)の特徴点によるフロー検出により，グループワーク学習時の教室内の移動人物をオフライン処理で検出した．また肖凌らは，室内の複数人の挙手動作を，レーザーレンジファインダとステレオカメラにより検出した[2]．

肖凌らは主に肌色情報とエッジ情報を用い，これらの関係がある条件を満たしたときに挙手していると判定させた．

これらの手法に対し，本提案システムでは深層学習手法を活用することで，各状態(クラス)を識別するためにどのような特徴を用いればよいかも含めた最適化を行った．また，後述のように既存の画像データを用い深層学習で得られたモデルを利用すれば，新規画像に対するクラス分類は比較的高速に実行されるため，リアルタイムでの運用が可能である．

深層学習に関しては，特に近年の急速な進展により，一般画像に何が映っているか(人物や物体など)の推定が活発に研究されており，高い精度が達成されている．たとえば静止画像認識では，VGGNet[3]，GoogLeNet[4]，ResNet[5]などの深層学習アーキテクチャの登場により，一般物体(1,000 カテゴリ)の Top-5 Error と呼ばれる判定基準（推定結果上位 5 クラスに正解が 1 つでも含まれる場合に正解とカウント）のエラー率が 3.5%程度にまで向上している．現在は，静止画像だけではなく動画像解析においても，その応用分野拡大と精度向上が求められている．

本研究では，講義の撮影映像に対して各種の深層学習のパラメータチューニングを行い，最終的に学生の受講状態（スマートフォン操作，資料閲覧，居眠り，その他）を認識するモデルを獲得することができた．それらの手順の概要を以下に示す．

3.　システムが想定する環境

本システムは，深層学習手法を活用し，教室内で受講している学生らの状態をリアルタイムに認識することを可能にする．状態把握手法として用いる深層学習には大量の訓練用データが事前に必要である．本研究では，訓練用データは次のような手順で取得した．

(1)　教室側面の上部に撮影用カメラを設置
(2)　被験者 8 名（女性 5 名，男性 3 名）が着席
(3)　訓練データ用に，被験者に「資料閲覧(text)」，「スマートフォン操作(phone)」，「居眠り(sleep)」の動作を各 1 分程度実行してもらい撮影した．実験者が状態の切り替えを合図した．

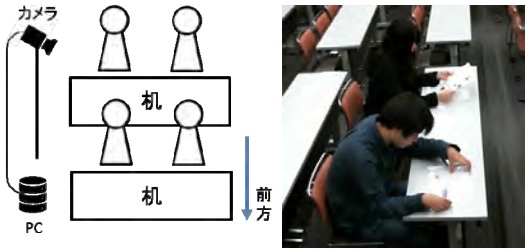

図 1．深層学習用訓練データ撮影

(4)　テスト用に上記 3 状態をランダムに実行したケースを撮影した．
(5)　映像(30 fps)の 1 フレーム中の各被験者の手元の 4 点を指定し，全フレームを正方形の画像に射影変換した．

図 2. 射影変換後の訓練データの一部

4.　システム概要と深層学習

本システムは，Python，TensorFlow，Keras，OpenCV などのプログラミング言語やライブラリを使用し Visual Studio Code (Microsoft)上で実装した．TensorFlow とは Google が開発しているオープンソースの機械学習ライブラリであり，Keras は TensorFlow 上で動作する，より高いレベルで抽象化された機械学習ライブラリである．

4.1.　訓練画像

前述の元映像から抽出した 41,264 枚の訓練画像(text 15,202 枚, phone 23,424 枚,sleep 2,638 枚)と 1,133 枚のバリデーション画像(text 229 枚, phone 176 枚,sleep 728 枚)を用いた．テスト画像は，前節(4)のランダムに実行された映像データから抽出した 158 枚を用いた．

図 3.　訓練画像（上）とテスト画像（下）の一部

4.2.　使用したネットワーク

取得した訓練画像に対し，Keras で提供されていた VGG16[3]モデルのネットワーク構造をベースとした機械学習を実施した．VGG16 モデルは ImageNet の 120 万枚の画像を 1000 カテゴリに分類する際に用いられた代表的な畳み込みニューラルネットワークモデルであり，13 層の畳み込み層，3 層の全結合層から構成される．ここで用いた VGG16 のネットワーク構成を図 4 に示す．

VGG16 は 1000 カテゴリに分類するものであるため，最後の全結合層(図 4 の fc の部分)を削除し，3 カテゴリ(クラス)に分類できるように，新しい全結合層を追加した．その際に過学習を防ぐ Dropout 層を追加している．今回用いたネットワークの諸元を表 1 に示す．

表 1. 深層学習の各種パラメータ

訓練画像	41264 枚
バリデーション画像	1133 枚
テスト画像	158 枚
学習させた層	すべての層
最適化手法	Adam 法（学習率 0.00001）
追加した最後の全結合層の Drop Out 率	0.5
入力サイズ	64×64×3
バッチサイズ	256
エポック数	10
訓練画像生成パラメータ	shear=5, zoom=0.1, h_shift=0.2, w_shift=0.2

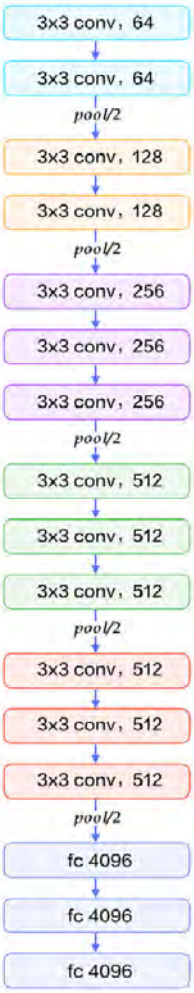

図 4. VGG16 のネットワーク構成

5.　学習結果

今回は VGG16 が ImageNet からの学習により獲得されたモデル(重み)を使用して訓練を行った．訓練画像やバリデーション画像とは独立に撮影して得られたテスト画像に適用した際の，資料閲覧(text)，スマートフォン操作(phone)，居眠り(sleep)を認識させ予測した結果の確率(probability)変化のグラフを図 5 に示す．

図中の truth（グラフ上部の帯）は人間が目視で判断した 3 状態，すなわち「正解」であり，変化する実線が，モデルが推定した状態の確率変化を示す．これらのグラフより，訓練画像やバリデーション画像とは独立したデータに対して，text，phone，sleep ともに，おおむね正しく推定できていることが分る．

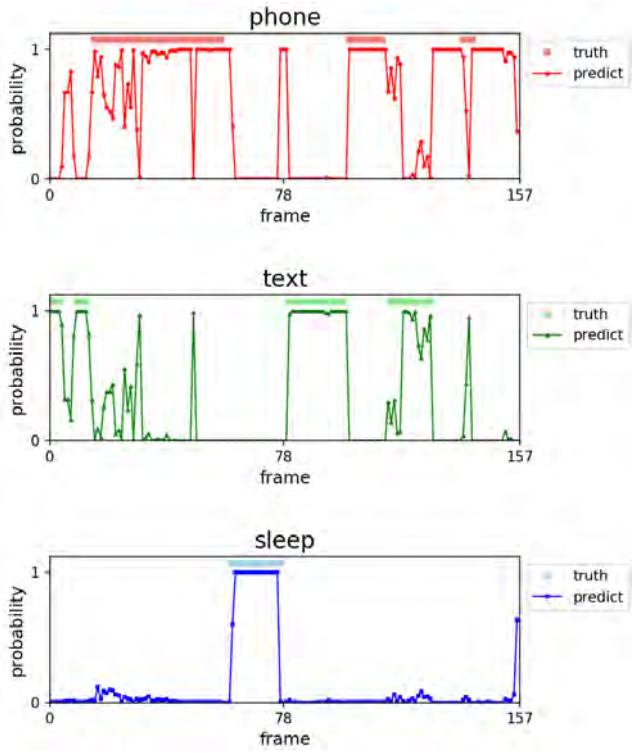

図 5. 3 クラス（phone/text/sleep）の確率変化

図 5 にこれから計算される混合行列を示す．これは縦軸に正解のクラス，横軸に本システムにより予測された結果のクラス(ラベル)を示したものである．値は精度を表す(1.0 が最大).

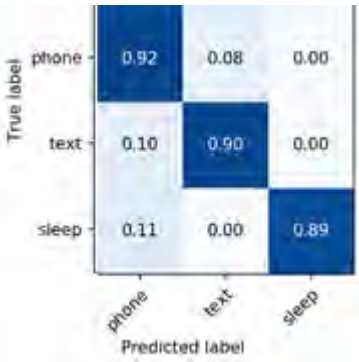

図 5. 精度変化（3 クラス）と混合行列

この表より，text は 0.90，phone は 0.92，sleep は 0.89 という高い精度で予測することができた．また，今回得られたモデルでは，学生 1 人の 158 枚のテスト画像の全状態推定を約 5 秒で行えたことから，実際の講義室で本システムを運用する場合，例えば 32 人の学生に対して 1 人当たり約 1 秒間隔で全員の状態推定を行えることが分った．

6.　訓練時の精度向上の工夫

本研究では，精度向上のために次に述べる 2 つの施策を行った．

6.1 ネットワーク可視化

精度向上にあたり，ネットワークが反応しやすい箇所(2 次元的位置)を特定するため Grad-CAM[6] の手法により，獲得されたネットワークを可視化した．可視化により得られたヒートマップを図 7 に示す．左側が画像，右側がそのヒートマップである．ヒートマップ上で色が赤い部分が，コンピュータが注目している箇所である．

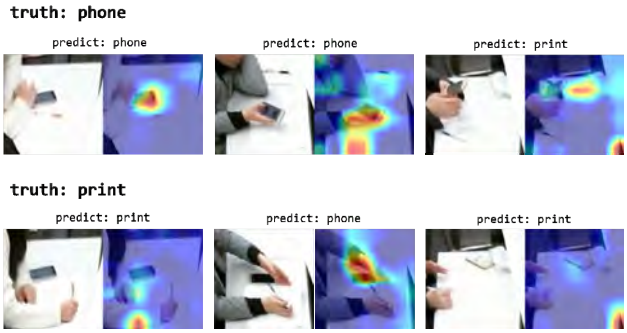

図 7. Grad-CAM[6]によるネットワーク可視化

これらの可視化により得られたヒートマップから，人間の目視により，明らかに訓練時の正解ラベルに関連が無いと考えられる箇所(例えば，スマートフォン操作画像に対して右下隅の黒い領域が反応している等)が分った．今回は，訓練画像の対応する箇所を白く塗りつぶし，再度訓練を実施し精度を向上させることができた．このように深層学習で獲得したネットワーク状態を可視化することで，再学習へのフィードバック，チューニングの有効な補助となることが分った．

6.2 訓練データ加工（水増し）

DC-GAN[7]の手法により phone 画像・非 phone 画像の深層学習を行い，獲得されたネットワーク(Generator)により疑似 phone 画像を生成した．これらを訓練画像として用いることにより，訓練画像数やバリエーションを疑似的に増加させた．

図 7. DC-GAN[7]による生成画像の一部

7. まとめと今後の展開

本論文では，本学の講義の履修生を撮影した映像を訓練画像として深層学習を行い，講義における学生の状態認識を行うシステムについて提案し評価した．今回得られたモデルでは，学生 1 人の 158 枚のテスト画像の状態推定を約 5 秒で行えたことから，実際の講義室で本システムを運用する場合，例えば 32 人の学生に対して 1 人当たり約 1 秒間隔で全員の状態推定を行えることが分った．

これにより，従来は人間が行っていた，履修生の動的な状態把握といった作業の自動化に応用でき，教育現場での教員や TA の支援が可能になると考えられる．

実際の運用に向けた応用化にあたっては，さらに得られた情報をどのように教員に提示するのかや，多くの人数や細かい時間間隔が必要な場合などがあるが，後者はコンピュータとカメラを増設することで解決できると考えられる．

また，今回得られたモデルによる学生の状態推定結果を講義の評価指標（居眠り率や活発度の定量データ）とすることで，従来の教師や TA の負担を軽減し，アクティブラーニングの実現に向けた補助システムとして活用でき，TA の代わりにロボットなどの行動に反映[8-13]させたり（例：ロボットが「スマートフォンをしまってね」などと発話）することなどが考えられる．

謝辞

本研究は大妻女子大学戦略的個人研究費 (S3003)の助成を受けた．

引用文献

[1] 西野博貴ほか．講義映像における移動物体の自動検出．電子情報通信学会技術報告. 2013, 07, pp. 47-52.

[2] 肖凌ほか．複数人物が任意の向きで着席した室内における挙手動作認識．映像情報メディア学会技術報告, 2011, vol.35, no.8, pp. 103-106

[3] K. Simonyan, et.al. Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR, 2014.

[4] C. Szegedy, et.al. Going Deeper with Convolutions. CVPR, 2015.

[5] Kaiming He, et.al. Deep Residual Learning for Image Recognition. ILSVRC, 2015.

[6] Ramprasaath R. Selvaraju, et.al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626.

[7] A. Radford, et.al. Unsupervised representationlearning with deep convolutional generative adversarial networks. ICLR, 2016.

[8] 永田雅人，松田晃一．マルチロボットによる分散型イベント制御システムの試作と実証実験．信学技報, 2018, vol.118, no.306, p.61-66.

[9] H.Ishiguro, et.al. Robovie: An interactive humanoid robot. International Journal of Industrial Robotics, 2001, vol.28, no.6, pp. 498-503.

[10] Y.Kondo, et.al. A gesture-centric android system for multi-party human-robot interaction. J. Human Robot Interaction, 2013, vol.2, no.1, pp. 133-151.

[11] 小林宏．表情豊かな顔ロボットの開発と受付システムの実現．日本ロボット学会誌, 2006, vol.24, no.6, pp. 708-711.

[12] 田中文英. 幼児教育現場におけるソーシャルロボット研究とその応用. 日本ロボット学会誌, 2011, vol.29, no.1, pp. 19-22.

[13]平野愛理, 松田晃一. ヒューマノイド型ロボットを用いた褒める行為に着目した学習支援システムの試作と評価, 情報処理学会研究会報告, 2019, Vol.2019-CE-148, No.15, pp.1-8

（受付日：2019 年 6 月 4 日，受理日：2019 年 6 月 12 日）

松田　晃一（まつだ　こういち）
現職：大妻女子大学社会情報学部教授

東京農工大学大学院工学研究科数理情報工学専攻修了．博士（工学，東京大学）．
専門はヒューマンコンピュータインタラクションやユーザエクスペリエンス．現在はヒューマンコンピュータインタラクションへの AI の導入などに焦点をあてた研究を行っている．
主な著書：Personal Agent-Oriented Virtual Society（単著，Advanced Knowledge International），WebGL Programming Guide（共著，Addison-Wesley Professional）