2E-02

日本語の難読文字検出を目的とした顔動画像解析に関する検討

伊藤　悠大　　石沢　千佳子　　景山　陽一

秋田大学

1. 背景・目的

授業中に使用する電子テキストなどにおいて，補足情報が必要となるタイミングでは，分からない用語に対して学生の視線が留まり(注視) [1]，顔をしかめるといった表情の変化(困惑表情)が起こると予想される．図 1 に示すように，注視や困惑表情の表出タイミングで視線が向けられた用語の補足説明を自動的に提示することができれば，学生が説明を探したり表示したりする操作が不要になるため，集中力を低下させずに授業の理解度の向上を図ることが可能と考える．

そこで本研究は，文章を読んだ際に読者が抱く「読みづらい」などの心情を自動的に収集し，心情に基づいた適切な補足説明を自動的に提示可能なシステムの開発を目的とする．これまでに，視線が留まる(停留)位置とその時間の長さ(停留時間)を用いることで，文章を黙読する人が「読みづらい」と感じて注視した文字(難読文字)を検出する手法について検討を加えてきた[2]．ここで，視線の停留情報に加えて，読者の顔を撮影した顔動画像を併用することは，難読文字の高精度な特定を可能にすると考える．しかしながら，既存の顔表情解析の研究は，作業を行っていない理想的な状態下で撮影された動画像を対象としており，文章黙読時の微細な顔表情の変化を対象とした事例は殆ど存在しない[3]．

そこで本稿では，文章黙読中の顔動画像を対象に，難読文字を黙読している顔表情を識別可能な手法の開発を目的とする．具体的には，顔表情を識別可能な機械学習モデルを複数の種類作成し，難読文字を黙読しているときの顔表情の識別精度に関して比較・検討を加えた．

2. 使用データ

2.1. 取得環境

Web カメラ (Logicool C920)を用いて，PC 画面上の文章を黙読している間の読者の顔動画像を取得した．カメラの解像度は 1920×1080pixel であり，フレームレートは 30fps である．被験者が黙読中にキーボード操作を行うことで，難読文字を黙読している顔画像かどうかの正解データを取得した．具体的には，黙読中に読みづらいと感じた場合にキーボードを押下させ，キーボード押下時の顔動画像は難読顔画像，非押下時の顔動画像は定常顔画像として取得した．顔動画像の例を図 2 に示す．

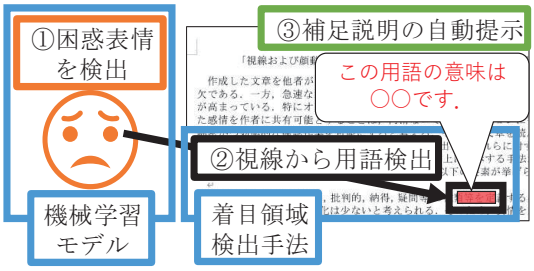

図 1. 本研究が目的とするシステムのイメージ（困惑表情が表出された用語を補足説明）

被験者は日本語を母国語とし，日常的に PC を使用している 20 代の大学生 7 名である(A～G)．なお，データ取得は「秋田大学手形地区における人を対象とした研究に関する倫理規程第 6 条第 2 項」に基づき，被験者の同意の下で行った．

使用した文章の例を図 3 に示す．使用した文章は，平均 71 文字の 5 種類の文章である．いずれの種類の文章も日本漢字能力検定試験[4]の読み問題から引用した．文章の難易度は，事前の調査で被験者が読み飛ばしなく黙読できる難易度の中で最も難しいと回答した準 2 級である．フォントは，等幅フォントである MS 明朝を用いた．1 文字の大きさは約 10mm 四方，画面と被験者の距離は約 50cm であり，1 文字あたりの視野角は約 1.15°である．

2.2. 前処理

取得した黙読顔画像は定常顔画像よりも枚数が多いため，定常顔画像をランダムに除外し，黙読顔画像と定常顔画像の枚数を被験者ごとにそろえた．画像の総数は 3760 枚である．

3. 難読顔画像の識別手法

本稿では，時系列データの解析に用いられる BiLSTM モデル[5]，顔領域内の特徴領域の検出に優れている POSTER モデル[6]，BiLSTM モデルと POSTER モデルの利点を組み合わせた POSTER-LSTM モデルを用いて，3 種類の表情識別モデルを作成した．

図 2. 撮影した顔動画像の例（被験者 B）（左：難読顔画像，右：定常顔画像）

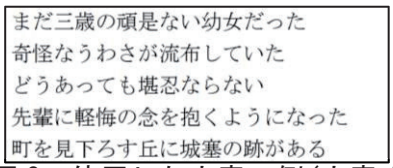

図 3. 使用した文章の例（文章 1）

An Analysis of Facial Video for Detecting Hard-to-read Characters in Japanese
Yudai Ito, Chikako Ishizawa, Yoichi Kageyama
Akita University

3.1. BiLSTM モデル

BiLSTM モデルは，時系列データを未来と過去の双方向に学習可能なモデルである．本稿では，顔特徴点の座標値の時系列変化(60 フレーム分)を用いて難読顔画像の識別を行うモデルを作成した．入力する顔特徴点は 8 種類である．具体的には，InsightFace[7]に基づいて，左右の各まぶたの開度，左右の各眉頭と鼻の頂点との距離，左右眉頭間の幅，顔全体の 3 軸方向の回転を算出し，それぞれ標準化処理を施して入力した．

3.2. POSTER モデル

POSTER モデルは，入力された 1 枚の画像内に写る人物の感情を顔表情に基づいて分類するモデルである．BiLSTM モデルと比較した場合，顔表情の時系列変化は考慮できない一方，表情識別に最適な特徴量を顔画像から自動的に抽出することが可能である．本稿では，既存のベンチマーク用データセットを用いて事前学習済みの POSTER モデルを使用した．

3.3. POSTER-LSTM モデル

POSTER-LSTM モデルは，POSTER モデルを用いて画像から抽出した特徴量を LSTM モデルに入力することで，画像の時系列変化(動画)を直接学習し，動画像を分類することが可能である．本稿では，入力された 3 枚の時系列顔動画像を用いて難読顔画像の識別を行うモデルを作成した．具体的には，既存のベンチマーク用データセットを用いて事前学習済みの POSTER モデルに対し，未学習の LSTM 層 1 層を追加した．

4. 実験方法

難読顔画像の識別に有用な手法を明らかにするために，3 種類の表情識別モデルに対して学習処理を施し，テストデータに対する各モデルの判別精度を算出した．BiLSTM モデルは取得データを用いて学習を行い，POSTER モデルと POSTER-LSTM モデルは，取得データを用いて重みを固定しない追加学習(ファインチューニング)を施した．なお，学習率は，各手法のデフォルト値を用い，学習の収束が認められない場合に値を変更した．バッチサイズは，設定可能な最大値を用いた．エポック数は，過学習が発生するまで学習させた際に最もテストデータに対する正解率の高い値をデータセット毎に用いた．取得データの偏りを考慮するため，各手法の精度の算出には交差検証を用いた．交差検証では，5 種類の文章中 4 種類を学習用データとし，残りの 1 種類をテスト用データとするデータセットを 5 種類の組み合わせに対して作成し(全 5 組)，テストデータに対する平均精度を算出した．モデルの識別精度として，正解率(Accuracy)，再現率(Recall)，適合率(Precision)を算出した．

5. 実験結果および考察

各モデルの識別結果を図 4 に示す．Recall は最大で 0.60 の値を得た．停留時間に基づく研究で得られた Recall の 0.39 を上回る値であり，顔動画像に基づく手法は，視線の停留時間に基づく手法よりも難読文字検出に有用である可能性が示唆された．POSTER-LSTM モデルは，Accuracy，および Precision においても，それぞれ最大値の 0.77 および 0.91 を示した．この要因は，画像の時系列変化を直接学習可能であったためであると考える．実際の検出結果において，時系列変化を考慮可能な LSTM 層の有無を比較した例を図 5 に示す．図 5 に示す一連のフレームは，被験者 D が難読文字の黙読に差し掛かり，思考のために顔の動きおよび表情が静止した箇所である．POSTER モデルは難読顔画像を全て定常顔画像と誤識別したことに対し，POSTER-LSTM モデルは，時系列変化を考慮し，静止状態を認識可能であったため，難読顔画像を正しく識別可能な場合が認められた．これは，難読顔画像の識別における時系列的な解析の重要性を示唆している．

今後は，提案手法を応用して難読文字の検出を行った場合の検出精度についても検証を行う．

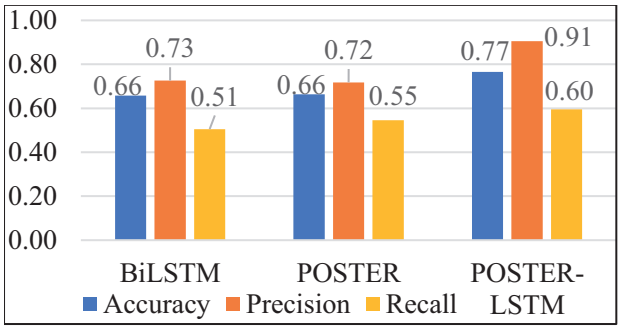

図 4. 各補正手法における難読文字検出精度

フレーム番号	458	459	460	461	462	463	464	465	466	467	468	469	470
POSTER													
POSTER-LSTM													

：識別成功　：識別失敗

図 5. 難読顔画像の識別結果例(LSTM 層の有無)

6. 謝辞

本研究は，JSPS 科研費 JP22H04165 および JP23H05147 の助成を受けて行われた．

参考文献

[1] X. Ma, Y. Liu, R. Clariana, C. Gu, P. Li: “From eye movements to scanpath networks: A method for studying individual differences in expository text reading,” Behavior Research Methods, Vol.55, pp. 730-750 (2023)
[2] Y. Ito, C. Ishizawa, and Y. Kageyama: “Threshold Determination Method for Detecting the Gazing Areas on a PC Screen While Silent Reading Japanese Texts,” IEEJ Transactions on Electrical and Electronic Engineering, Vol. 18, pp. 714-721 (2023)
[3] M. Sajjad, F. U. M. Ullah, M. Ullah, G. Christodoulou, F. A. Cheikh, M. Hijji, K. Muhammad, J. J. P. C. Rodrigues: “A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines,” Alexandria Engineering Journal, Vol. 68, pp. 817-840 (2023)
[4] 日本漢字能力検定協会，“日本漢字能力検定”, https://www.kanken.or.jp/kanken/ (Accessed: 2024/1/4)
[5] Z. Hameed, B. Garcia-Zapirain: “Sentiment Classification Using a Single-Layered BiLSTM Model,” IEEE Access, Vol. 8, pp. 73992-74001 (2020)
[6] C. Zheng, M. Mendieta, C. Chen, “POSTER: A Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition,” arXiv, Vol.2204.04083 (2022)
[7] “InsightFace,” https://insightface.ai/ (Accessed: 2024/1/4)