深層学習による多次元時系列データを用いた連続指文字認識システムの開発

白石優旗 [1)]，土屋智彦 [2)]，加藤伸子 [1)]，米山文雄 [1)]，設楽明寿 [3)]

筑波技術大学　産業技術学部　産業情報学科 [1)]
筑波技術大学大学院　技術科学研究科　産業技術学専攻 [2)]
筑波大学大学院　図書館情報メディア研究科　図書館情報メディア専攻 [3)]

キーワード：手話言語，センサグローブ，機械学習，Convolutional Neural Networks，Long Short-Term Memory

1．はじめに

昨今のダイバーシティ推進により，ろう・難聴者と聴者が共に生活する機会のさらなる増加が期待され，両者の円滑なコミュニケーションの実現は急務である。したがって，代替手段を用いて情報獲得の支援を行う「情報保障」は重要である。これまでの情報保障は，聴者の音声を文字（手話）に変換（通訳）して，ろう・難聴者に提示することが主流であった。しかしながら，コミュニケーションは双方向的であることが本質であるため，ろう・難聴者の発話（手話・指文字）を聴者に文字（音声）で提示することも必要となる。

1 対 1 の局面では，双方が文字を使う（筆談）等の工夫でコミュニケーション自体は可能であるものの，そのような対応がそもそも困難な状況がある。例えば，講演・会議・グループでの話し合い等，ろう・難聴者が多数の聴者に対し手話を用いて発話するケースである。本研究ではこれらの状況を第一義的に想定する。

本研究では，本課題の解決のため，動的指文字（手指を動かしながら提示する指文字）を含む全指文字（76 文字）に対し，それら指文字が連続的に提示された状態（連続指文字）における認識システムを開発する。認識手法には，近年高い認識精度の報告が多数なされている深層学習を用いる。また，センサグローブを採用することにより，カメラと異なり環境光やオクルージョン（手と手の重なり）の影響を受けずに，またカメラの存在を意識せずに，聴講者に対して発話することができる（図 1）。開発システムの詳細については，参考文献 [1, 2] を参照されたし。

昨年度は，センサグローブの改良，データ収集実験に最適な単語の選定，実際のデータ収集実験，並びに，認識アルゴリズムの改良を行った [3]。本年度は，昨年度に採取したデータを用いて，実際に連続指文字認識システムを開発し識別実験を行ったため，本稿ではその結果について簡潔に述べる。

2．関連研究

現在の手話認識研究は，カメラ（RGB カメラ・深度センサ）を用いたものが主流であるものの，約 1,000 単語に対する単語誤り率 22.9%（訓練データと同一手話者）であり，実用的なレベルで「文」を認識するまでには至っていない [4]。センサグローブとカメラを共に利用したものについて Ong らの報告 [5] があるが，隠れマルコフモデルを活用した認識により，英数字の指文字 37 文字と手話 30 単語の識別率 80% 程度に留まっている。センサグローブのみを用いた手法について Mummadi らの報告 [6] があるが，様々な機械学習による手法（深層学習は含まれていない）を比較・検討することで識別率 92%を達成したと報告されているものの，識別対象はフランス手話の内，手指を動かさずに提示する静的指文字 22 文字に限定されている。

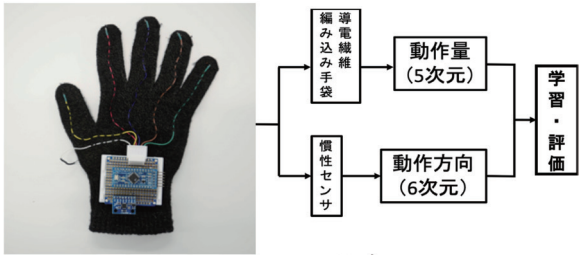

図 1　システム構成

3．学習・評価用連続指文字データ

学習・評価用データは，連続的な指文字の提示動作に起因する他の指文字への認識ミスのしやすさを考慮して選定した 64 単語（2 ～ 5 文字）を対象とし，日常的に手話を使用している，ろう・難聴の実験協力者 32 名（20 ～ 23

歳）から1人あたり5回繰り返して採取した。その際，指の動作量（5次元）と手の動作方向（加速度3次元，角速度3次元）について，120 sps で計測データ（多次元時系列データ）を取得した。本実験は筑波技術大学倫理審査委員会の承認を得て実施した。

その後，加速度と角速度から Madgwick フィルタにより角度を算出した。次に，データクリーニングにより欠損データを取り除き，ELAN を用いて動画を確認しながら提示時間毎に指文字のラベリングを手動で行った。その際，指文字を提示していない区間（わたり）はφとラベリングした。5本の手指の折り曲げデータを可視化したものを図2に示す。

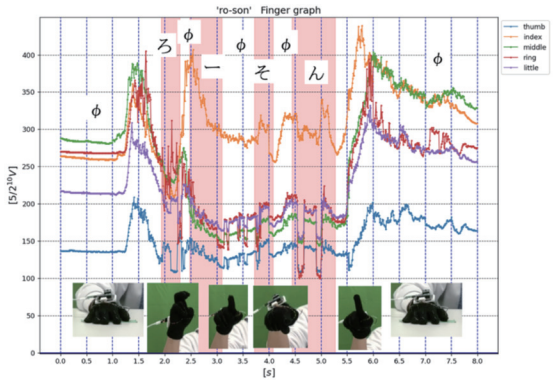

図2 「ろーそん」の5本の手指の折り曲げデータ

4. 連続指文字認識実験

連続指文字認識は，音声認識と同様に，連続的に提示された指文字に対して各時刻での指文字を認識するものである。したがって，音声認識で活用されている Long Short-Term Memory（LSTM）をベースラインとし，我々がこれまでに単一指文字認識で使用してきた Convolutional Neural Networks（CNN）を組み合わせた複数のモデル（7種類）を構築し，比較実験を行った。なお，NN への入力には，正規化を行った後，移動平均により 4sps にデータ圧縮したものを用いた。

すべてのサンプルに対して識別したものを 5-foldCV で評価した結果，図3に示す CNN と LSTM を組み合わせたモデルで，92.1%の認識率を確認した。なお，結果の詳細については，学術論文として投稿中である。

5. まとめと今後の課題

本研究では，深層学習を用いて動的指文字を含む全指文字の連続認識システムを開発し，実験を行った。結果，認識率 92.1%を確認したものの，手指の接触への対応，φに起因する教師データの偏り，単一時刻の識別から単語の識別への統合処理等の課題が残されている。今後は，手話言語の特徴を考慮した認識モデルを構築することで，手話認識システムの実現を目指す。

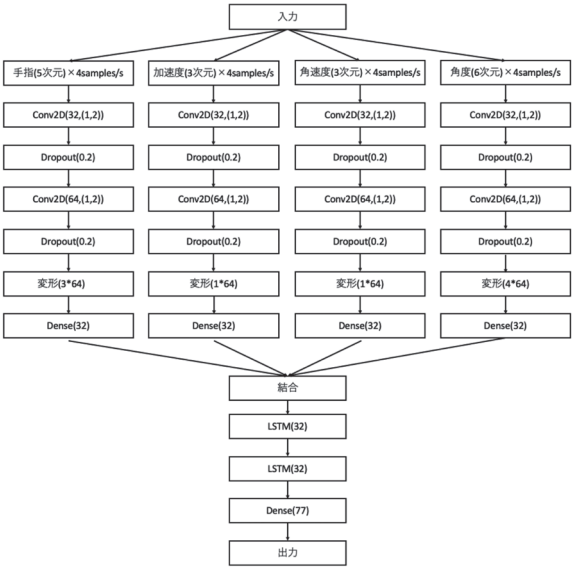

図3 CNNとLSTMを組み合わせた学習モデル

謝辞

本研究は筑波技術大学教育研究等高度化推進事業並びに JSPS 科研費 19K11411 の助成を受けたものです。

参照文献

[1] 土屋智彦，白石優旗，深層学習を用いたセンサグローブによる指文字認識の改良．情報処理学会アクセシビリティ研究会（IPSJ SIG AAC）第9回研究会；2019-3.

[2] T. Tsuchiya, et. al., Sensor Glove Approach for Japanese Fingerspelling Recognition System Using Convolutional Neural Networks, Proc. of the 13th Int. Conf. on ACHI 2020, pp.152-157, Valencia, Spain, Mar. 2020.

[3] 白石優旗他，深層学習による多次元時系列データを用いた連続指文字認識手法の検討．筑波技術大学テクノレポート ,pp.77-78, vol.20, 2021.

[4] D. Bragg, et. al., Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective, pp.16-31, ASSETS '19, 2019.

[5] C. Ong, et. al., Sign-Language Recognition Through Gesture & Movement Analysis (SIGMA), In: Billingsley J., Brett P. (eds) Mechatronics and Machine Vision in Practice 3. Springer, Cham, 2018.

[6] C. K. Mummadi, et. al., Real-time Embedded Recognition of Sign Language Alphabet Fingerspelling in an IMU-Based Glove, Proc. iWOAR 2017, Rostock, Germany, Sep. 2017.