Vol.2018-AAC-6 No.4
2018/3/9

深層学習を用いた センサグローブによる指文字認識の検討

小松周生[†1]　白石優旗[†1]

概要： 聴覚障害者と健聴者のスムーズなコミュニケーションを実現するため，それぞれの発話内容を自動で認識する，手話認識システム及び音声認識システムが求められている．しかし，日本における手話認識システムは，音声認識システムと比較して認識精度が低いのが現状である．本論文では，認識精度の向上を目指し，深層学習を用いたセンサグローブによる指文字認識システムを新たに提案し，評価実験によりその有効性について検討する．

キーワード：手話，機械学習，コミュニケーション支援，聴覚障害，情報保障システム

1. はじめに

近年，音声認識，およびその関連技術についての研究が進み，音声による入力機能を備えた情報機器が広く普及してきている．それに伴い，音声認識による聴覚障害者のための情報保障システムとして，「こえとら」[1]や「UD トーク」[2]などのスマートフォンのアプリが公開されている．これにより，健聴者の音声を聴覚障害者が読み取ることが可能になり，両者の円滑なコミュニケーションの第一歩を踏み出したと言える．

一方，聴覚障害者が発する手話を健聴者が読み取ることができる様になるための，手話認識による情報保障システムについての研究は，音声認識による情報保障システムと比較して認識精度が低いのが現状である．なお，聴覚障害者同士の日常会話では，主要なコミュニケーション手段として手話が使われている．

手話言語は音声言語とは異なる特徴を持つ言語であり，健聴者が手話言語を習得し，読み取りができるようになるのは容易ではない．したがって，手話を音声情報，文字情報に変換し，健聴者に対する情報保障，すなわち手話認識システムが求められている．

手話認識システムを実現するためには，手の位置，向き，および動きを含む手の形を認識する必要がある．現状の手話を認識する方法は，大きく分けて非接触式センサであるカメラによる認識[3-4]と，接触式センサであるセンサグローブによる認識[5-6]とに大別できる．例えば，Luzhnica ら[5]が提案したセンサグローブによる手話認識では識別精度は98.5%達成したと報告されているものの，識別対象クラスは 30 程度であり，実用化されるまでには至っていない．

一方で，近年深層学習という技術が注目されている．ここで深層学習とは，機械学習の一種であるニューラルネットワークの隠れ層を増やしたものであり，識別率向上に貢献をしている．例えば，画像認識によるハンドジェスチャー認識精度の向上のために，深層学習を適用した手法[3]が報告されている．

†1 筑波技術大学
Tsukuba University of Technology

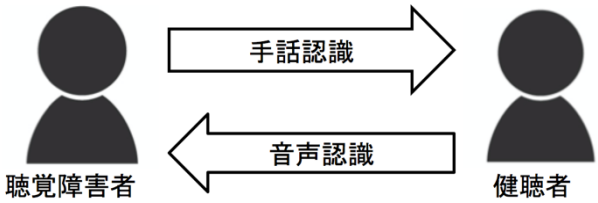

図 1　情報保障システム

本研究では，健聴者と聴覚障害者のコミュニケーションを円滑にするための手話認識への第一歩として，日本語対応の指文字（Japanese Finger Spelling，JFS）入力インターフェースの実現を目指す．ここで JFS とは，平仮名に 1 対 1 対応した手指の表現のことである．

センサには，接触式センサであるセンサグローブを採用する．非接触式センサであるカメラを使用する場合は，カメラに手が映るように設置する必要があり，またカメラは環境の影響を受けやすいため，外出先などでの利用が困難である．一方，センサグローブなど接触式センサによる手形状認識は手に装着するだけでできるため，利用が容易である．

我々は，センサグローブの重量とコストを抑え，着用者が手を動かしやすい手法である導電繊維編み込み手法を採用し，従来の指文字認識手法に深層学習を用いることで認識精度の向上目指す．

本論文では，深層学習による JFS 認識手法の提案，並びに認識システムを開発し，その他の機械学習による認識手法との比較評価実験を通して，提案手法の実現可能性について検討する．

2. 関連研究

これまでの指文字認識に関する研究には，カメラセンサを用い，指文字の一連の動作を画像にし，認識を行う方法と，センサグローブなどを用いた接触式センサを用いる方法がある．以下に，それぞれの方法について述べる．

2.1. 画像認識による認識

カメラを用いて指文字を撮影し画像処理を行い手形認識する手法が複数提案されている．原田ら[7]は指文字を撮影

したカラー画像を 2 値化して手の領域を抽出し，手形を認識する手法を提案している．三宅ら[8]が提案する手法は距離値を格納した距離画像に対して画像処理を行い非接触で認識を行う方法である．長嶋ら[9]は撮影された指文字の画像から「手首・拳の方向，孔の有無，指の本数」の情報を取得し，分類アルゴリズムとして決定木を採用し，日本語の静的指文字（1 つの文字の表現時に手指を動かさない指文字）41 文字の判別を行っている．このシステムにより，一人の指文字画像データ（41 文字× 7 セット）に対して，96.5%の認識率が得られている．Hosoe ら[10]は認識アルゴリズムに深層学習を採用し，日本語の静的指文字の認識を行なっており，認識率は 93%と報告されている．

このように，カメラを用いた非接触式センサにより指文字を認識する方法は、手の平にデータグローブなど装着する必要がなく，自然な指文字で比較的高い精度で認識されている．しかし，実用化のためには認識精度はまだ十分とは言えず，また動的指文字（1 つの文字の表現時に手指を動かす指文字）についての識別結果はほとんど報告されていない．

2.2. センサグローブによる認識

センサグローブなどの接触式センサを用い，得られた計測データを元に手形状認識する手法が複数提案されている．例えば，センサにより 5 本指の屈折，手の位置・方向の計測を行い，そのデータを PC やマイコンに計測データを送り，分類アルゴリズムにより手形状を認識する[11]というものである．Cabrera ら[11]は、5DT Data Glove 5 Ultra と加速度センサを組み合わせ，各指の屈曲度と手首の向きに関する情報を取得している．得られた測定値は，オフラインかつニューラルネットワークにより，24 の American Signe Language（ASL）の静的指文字の分類のテストを行なった．ニューラルネットワークに 5,300 パターンで訓練を行い，1,200 のテストパターンで 94.07%の精度が得られたとしている．Mummadi ら[12]は，手袋に複数の小さな慣性センサーユニット（IMUs）を埋め込んだセンサグローブプロトタイプの提案をしている．57 人のフランス手話（LSF）の指文字のデータを集め，テストを行なった結果，提案されたシステムにより平均認識率 92%，F 値 91%を得られたとしている．JFS 認識を行った手法の中で導電繊維編み込み手法[13]では曲げセンサの代わりに導電繊維が編み込まれている手袋自体をセンサとして利用し，さらに 3 方向ジャイロセンサを組み込む事で手形状と手の動きを認識する手法を提案している．ユーグリッド距離による指文字（「あ」～「お」）の認識率は 60%と報告されている．

3. 開発システム概要

本研究では，いつでもどこでも円滑なコミュニケーションを実現するため，軽量で着用しやすいセンサグローブを用い，指文字の動作を行うと同時にリアルタイムで高い精度により識別し情報伝達するシステムを設計する．

認識システムは，センサ値計測部と識別部で構成されている．本研究で開発した JFS 認識システムの外観を図 2 に示し，ソフトウェアの全体構成を図 3 に示す．

図 2　開発したプロトタイプ

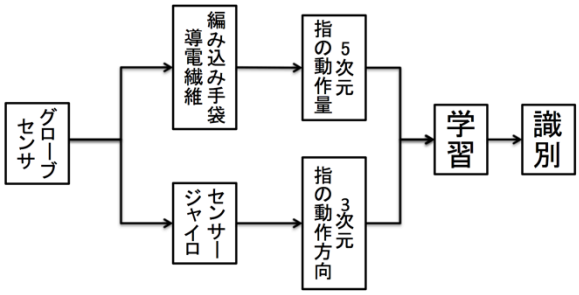

図 3　ソフトウェア全体構成

3.1. センサグローブ

手形の動きのある指文字を効率良く認識するために，センサグローブを用いることで，動作方向と動作量を検出して認識を行う．

本研究では，長時間身につけていて負担が少ない，かつ安価であり，比較的軽量である導線繊維編み込み手法[13]を採用する．動作方向はジャイロセンサを用い，動作量はグローブに編み込まれた導線繊維の抵抗の変化を利用し，手の動作方向と指の動作量を検出する．

具体的には，動作検出ボードには Arduino を採用し，PC と IC2 接続を行うことで検出されたセンサグローブの計測値を Arduino から PC に転送し，CSV 形式で保存する．機械学習並びに動作認識は PC 上で Python 言語を用いて行う．

JFS の動作のデータグローブのセンサ読取値は，JFS を行う人によって異なるスケールを有する．したがって，個人の手の動きの違いを考慮した線形正規化[13]を行う．その際，本システムの活性化関数，尤度関数は確率を扱うため，ネットワーク入力の前処理として，0~1 にスケール変換を行う．

3.2. 識別アルゴリズム

本研究では，深層ニューラルネットワークの有用性を検討するため，以下の学習・識別に 5 つの機械学習を用いて比較する．k-最近傍法（k-nearest neighbor, k-NN），ナイーブベイズ（naive Bayes classifier, NB），ロジスティック回帰（logistic recogression，LR），サポートベクタマージン（support vector machine，SVM），ランダムフォレスト（random forest，RF），深層ニューラルネットワーク（deep neural network，DNN），コンボルーションニューラルネットワーク（convolution neural network，CNN）である．k-NN，NB，LR，SVM，RF にはオープンソースの機械学習ライブラリ scikit-learn[10] を用い，DNN，CNN にはオープンソースライブラリである TensorFlow[11] を用いる．DNN，CNN の活性化関数として ReLU 関数

$$f(u) = max(u, 0) \qquad (式1)$$

を，誤差関数には交差エントロピー関数

$$E = -\sum_{k} t_k \log y_k \qquad (式2)$$

（式 2）を，学習アルゴリズムには Adam[12]を用いる．ここで t_k は正解ラベル（one-hot 表現）を，y_k はネットワークの出力を表す．

CNN は画像認識で識別に使われることが多く，一般に高い認識率を得ることができている．畳み込み層とプーリング層があることが特徴であり，学習過程で自ら特徴量抽出を行いながら学習することができる．CNN を採用するため，我々はデータグローブから得られた計測データを 2 次元データに変形し， CNN で学習・評価することを試みた．CNN のシステム概要を図 5 に示す．

4. 実験

本実験では，開発したプロトタイプを用いて，センサグローブから得られた動作量・動作方向データを用いて指文字が正しく識別できるか検証する．

4.1. 指文字データ収集

JFS 46 文字全てを対象としてデータ収集を行った．実験協力者 9 名（手話歴 4〜18 年の聴覚障害者）に対してセンサグローブを着用してもらい，指文字の動作を一文字に対して 1 秒間行ってもらった．結果，46 文字全てに対して 20 個／秒のセンサグローブデータ（動作量 5 次元，動作方向 3 次元の計 8 次元）を取得した．このとき，1 サンプルを 160 次元（20 個×計測データ 8 次元）とした．また，データの収集と同時に手動でデータのラベル付けを行った．この一連の動作を 10 回繰り返した結果，一人当たり 1 文字当たり 10 個，全指文字 46 文字に対して，合計 460 個の動作計測データを収集し，被験者 9 名から合計指文字データ 4,140 サンプルを集めることができた．

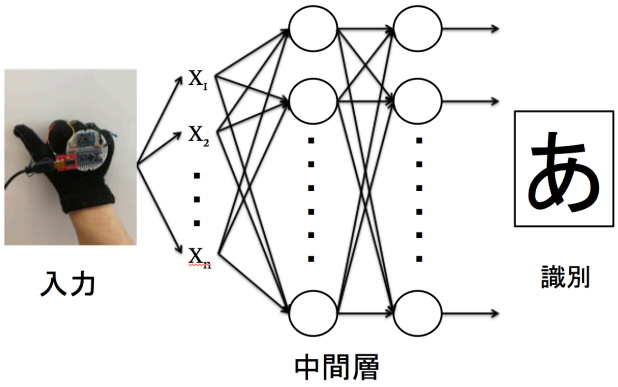

図 4 ニューラルネットワーク

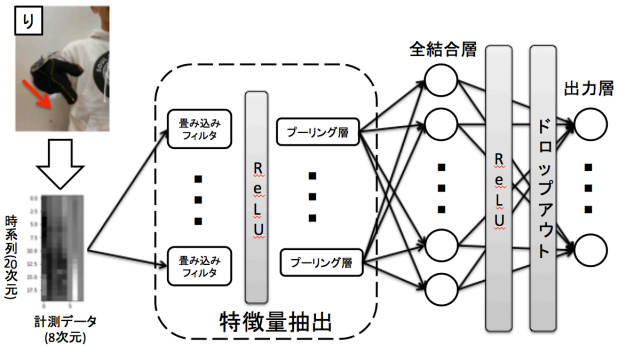

図 5 畳み込みニューラルネットワーク

表 1 静的指文字の識別結果（8 次元）

学習モデル	識別率	識別時間(s)
k-NN	0.952	1.19752
LR	0.432	0.01971
GNB	0.45	0.05450
SVM	0.336	49.14906
RF	0.992	1.01181
6DNN	0.957	0.56578

表 2 指文字（動的指文字を含む）の識別結果

学習モデル	識別率	識別時間(S)
k-NN	0.6684	0.3592
LR	0.6556	0.009
GNB	0.5103	0.326
SVM	0.6379	0.4542
RF	0.8007	0.0538
4DNN	0.7925	0.0396
CNN	0.8365	0.1479

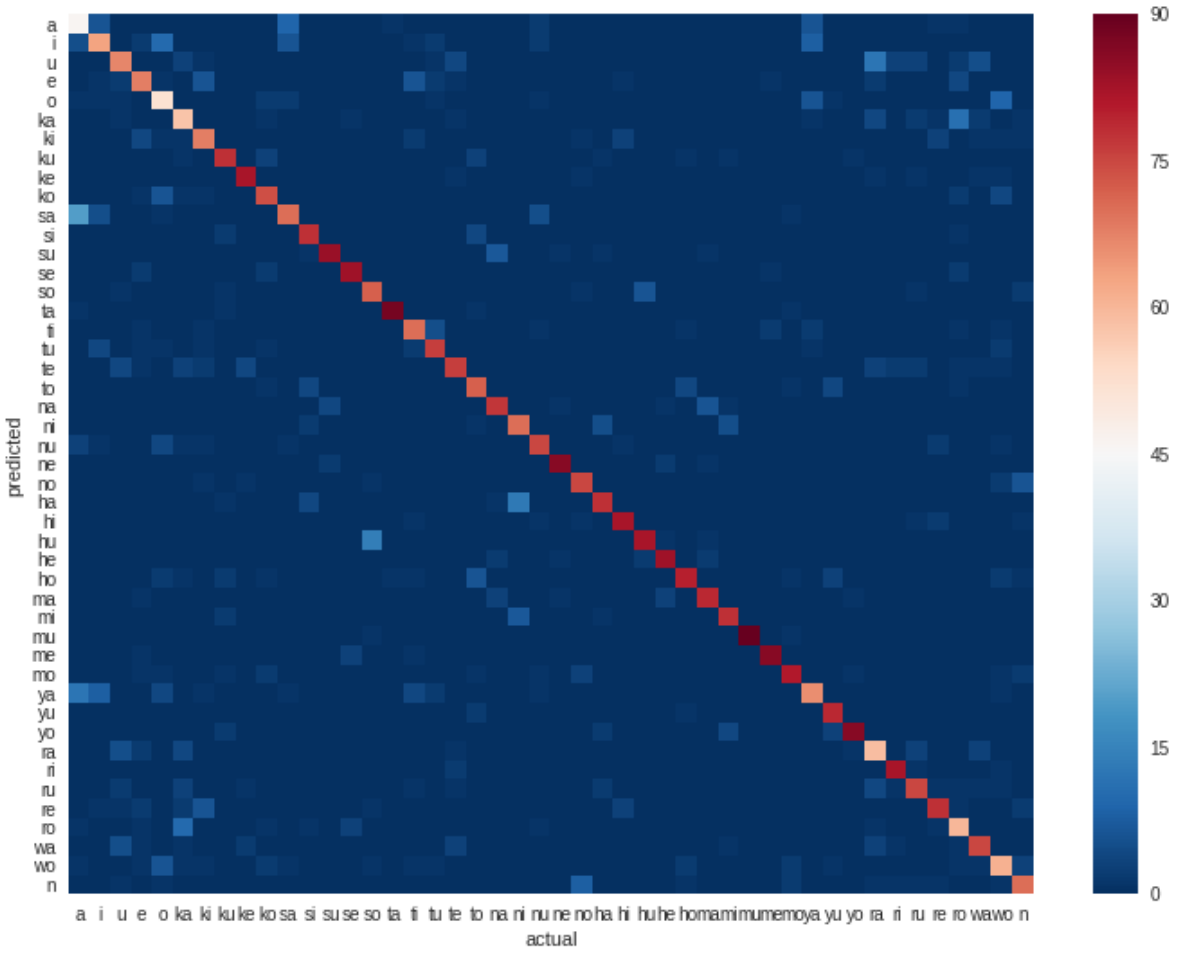

図 6　DNN における混同行列

4.2.　学習と識別

4.2.1.　静的指文字

静的指文字（動的指文字を除く）1 文字 10 個，41 文字合計 8,200 個の 5 本指の動作計測データ（8 次元）に対して学習・評価を行う．9 人のサンプルデータは合計 73,800 になる．k-NN，SVM，NB，LR，RF，DNN に対して 9-fold CV 法でモデルの識別評価を行なった結果を表 1 に示す．

表 1 から分かる通り，k-NN が 95.2%，6DNN が 95.7%，RF が 99.2%と高い識別率を得られた．RF の識別率が最も高くなった理由は，今回使用したセンサグローブから得られる計測データが 8 次元と低次元のため，判別式の学習が容易であったためと考えられる．

一方，RF と DNN で共通して，「そ」を「ふ」，「な」を「す」と誤って識別することがあった．これらの指文字の違いは，親指の屈折の有無であり，人によっては完全に屈折をしない人がいるため，導電繊維の短絡による電圧変化が起こり辛く，親指の電圧が類似していたためと考えられる．また，「ち」と「つ」と誤認識することもあった．これらの違いは，中指と薬指の密着の有無であり，「ち」の場合，薬指は第一関節のみ屈折をするため，電圧が似ていることから誤認識したと考えられる．

4.2.1.　指文字（動的指文字を含む）

動的指文字を含んだ JFS 46 文字合計 4,140 個の動作計測データ（160 次元）に対し，先と同様に k-NN，SVM，NB，LR，RF，DNN，CNN，について 9-fold CV 法でモデルの識別評価を行なった．結果を表 2 に示す．また DNN の正誤表（混同行列）を図 6 に示す．

表 2 から分かる通り，CNN の識別率が 83.7%で最も高くなり，他の手法と比較して有効であることを確認できた．また，静的指文字の識別結果と比較すると，RF，DNN 共に識別率が 15%程度低くなっていることが分かった．今回，入力が 160 次元と高次元のため，次元数に比べてデータ数が少なく，学習が困難になったためと考えられる．よって更なるデータ収集による学習データを増やすことで，識別率向上が期待できると考える．

一方，CNN，RF，DNN で共通して「あ」を「さ」と誤認識した．これらの指文字の違いは親指と人差し指の密着の有無であり，親指の屈折を行わないため，静的指文字と同様，親指の判定が困難であることが原因と考えられる．これについては，親指のみ慣性センサを付け加え，指の動作情報を増やすことで，対策ができる可能性がある．

5. まとめと今後の課題

本研究では，聴覚障害者と健聴者が円滑なコミュニケーションを実現すべく，軽量なセンサグローブを採用し，リアルタイムかつ高い識別率をもつシステム開発の第一歩として，様々な機械学習を用いた識別率の検証を行なった．結果，静的指文字に関しては k-NN，RF，DNN がそれぞれの識別率 95.2%，99.2%，95.7%となり，他の手法と比較して有用性を確認した．一方，動的指文字を含む全ての指文字の認識に関しては，CNN の識別率が 83.7%と最も高く，他の手法と比較して有効性を確認した．ただし，RF，DNN についても，それぞれの識別率 80.1%，79.3%となり，比較的高い精度で認識している．

今後は, 更なるデータ収集並びに学習方法の改善により，識別率の向上を目指す．同時に，センサの数を増やすことも検討する．また，濁音，半濁音，小書文字並びに手話に対応した識別モデル開発をする．更に，センサグローブとスマートフォンを Bluetooth で接続し、リアルタイムに識別できるシステムを開発することを目指す予定である.

謝辞　本研究の一部は，筑波技術大学平成 29 年度学長のリーダーシップによる教育研究等高度化推進事業による助成，並びに JSPS 科研費 JP16K16460 の成果であり，ここに記して謝意を表すものとする.

参考文献

[1] 国立研究開発法人情報通信研究機構(NICT): こえとら, FEAT(online),available from<http://www.koetra.jp/> (accessed,Junuary,2018)
[2] Shamrock Records，Inc．:UD トーク , Shamrock Records, Inc.(online),available-from<http://udtalk.jp/> (accessed,Junuary,2018)
[3] Srujana Gattupal: ”Evaluation of Deep Learning based pose Estimaition for Sign Language Recognition”, Prodeedings of the 9th ACM International Conference on Pervasive Technologies Related to AssistiveEnviroments Article No.12, 2016
[4] 西村洋介, 今村大輔, 堀内靖男, 篠崎隆宏, 黒岩眞 吾:“Kinect とパーティクルフィルタを用いた HMM 手話 認識手法の検討”, 電子情報通信学会, pp.161-166 (2012)
[5] Granit Luzhnica, Elizabeth Lex, Viktoria Pammer. A Sliding Window Approach to Natural Hand Gesture Recognition using a Custom Data Glove. In: 3D User Interfaces (3DUI); 2016 IEEE Symposium on ; 2016 Mar 19 ; New York : IEEE; 2016 ; p.81-90.
[6] K. Murakami and H. Taguchi. Gesture recognition using recurrent neural networks. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’91, pages 237–242, New York, NY, USA, 1991. ACM.
[7] 原田直人，張英夏，向井信彦：“カラー画像による指文字認識に関する基礎検討”，テレビジョン学会技術報告,Vol.17, No.14, pp.19-24,1993.
[8] 三宅太一,若月大輔,内藤一郎：“距離画像を用いた動きのある指文字の非接触認識手法の検討”，電子情報通信学会 2012,pp.270-275,2012.
[9] 長嶋裕二,藤井昌紀,長嶋秀世：“決定木を用いた日本語手話における指文字の認識”, 映像情報メディア学会年次会,ROMBUNNO.7-3,2014.
[10] Hana Hosoe, Shinji Sako and Bogdan Kwolek, “Recognition of JSL Finger Spelling Using Convolutional Neural Networks”, 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA) Nagoya University, Nagoya, Japan, May 8-12, 2017.
[11] Maria E Cabrera, Juan Manuel Bogado, Leonardo Fermin, Raul Acuna, and Dimitar Ralev, “Glove-based gesture recognition system”, In *Proc. of Intl. Conf. on climbing and walking robots and the support technologies for mobile machines*.747–753, 2012.
[12] Chaithanya Kumar Mummadi,, Frederic Philips Peter Leo,Keshav Deep Verma, Shivaji Kasireddy, Philipp Marcel Scholl,Kristof Van Laerhoven ,“Real-time Embedded Recognition of Sign Language Alphabet Fingerspelling in an IMU-Based Glove”, the 4th international Workshop on Sensor-based Activity Recognition and Interaction, Rostock Germany, ISBN:978-1-4503-5223-9, 2017
[13] 高田介,志築文太郎,高橋伸 :“導線繊維編み込み手袋を用いた指の曲げ計測手法, 情報処理学会 2017, Vol.2017-HCI-171, No.25, 2017
[14] David Cournapeau:scikit-learn, INLIA(online), availablefrom<https://www.scikit-learn.org/stable/> (accessed November 2017)
[15] Google Brain:TensorFlow, Google(online), available from<https://www.tensorflow.org>(accessedNovember 2017)
[16] Diederik Kingma, Jimmy Ba, “Adam: A Method for Stochastic Optimization”, the 3rd International Conference for Learning Representations, San Diego, arXiv:1412.6980, 2015