Conbolutuonal Neural Networkを用いた手話判別に関する研究

竹内, 泰人 / Takeuchi, Yasuhito

(出版者 / Publisher)
法政大学大学院理工学・工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
58
(開始ページ / Start Page)
1
(終了ページ / End Page)
4
(発行年 / Year)
2017-03-31
(URL)
https://doi.org/10.15002/00014345

Convolutional Neural Network を用いた手話判別に関する研究

A Study of Several Gestures Recognition for Japanese Sign Languages by Applying Convolutional Neural Network

竹内泰人
Yasuhito Takeuchi
指導教員　小林一行教授

法政大学大学院理工学研究科システム理工学専攻修士課程

Sign languages are an important communication method for hearing impaired people. To communicate to hearing impaired people, normal people also have to learn the gesture of meanings. In order to achieve smooth communication between hearing impaired people, automatic sign recognition system is required. In this paper, we apply simple 2D camera with convolutional neural network (CNN) classification method is applied. The combination of simple 2D camera with CNN enable real-time recognition is achieved by using GPU based microcontroller board. Through preliminary experiments, we can classify three different types of gestures with accuracy of 70 %.

***Key Words** : Convolutional Neural Network, sign language, Hidden Markov Model*

１．はじめに

厚生労働省の調査によると日本にはおよそ32万4千人の聴覚・言語障害者がいるとみられている[1]．高齢になればなるほど聴覚・言語障害者の割合は増えており，高齢化社会がこれからますます加速する日本において，大きな社会問題になる危険性がある．

聴覚・言語障害者の内，コミュニケーションの手段として手話・手話通訳を用いている人はおよそ 6 万人と少ない[2]．そこには，聴覚・言語障害者同士であれば意思疎通ができるが，健常者とのコミュニケーション手段としては手話が機能しないという理由があると考えられる．そのため，手話を認識し，日本語に変換できるシステムがあれば，手話の汎用性が上がり，聴覚・言語障害者の不便が減ることが期待される．

これまでに，手話と日本語間を変換することで，コミュニケーションに役立てる研究事例は多く存在する．例えばみずほ情報総研レポート[2]では，モーションセンサ「KINECT」を用いて手話を認識する研究が発表されている．他にも株式会社日立製作所では手袋型の入力装置から手指動作を，ビデオカメラから手指以外の動作を認識する研究などがある[3]．また，近年は Deep Learning を使い，手話の指文字をカメラの画像から認識する研究もある[4][5]. Deep Learning は機械学習の過程で必要な特徴抽出を機械に任せることができるということもあり，特に画像処理の分野で目覚ましい成果をあげている．

手話を認識する際重要なのは，精度良く認識することはもちろんのこと，簡易的に行えることも必要になる．聴覚・言語障害者が実際に使用することを考えた場合，センサを体に取り付けることや，扱うためにコツが要るシステムでは実用化は難しい．そこで本研究では認識の際に被験者の負担にならないよう特殊なセンサは用いず，カメラから取得した画像処理によってのみ手話判別を行う．その際，認識の精度を担保するため，前述した Deep Learning を使用し，その中でも画像の判別に優れた Convolutional Neural Network(CNN)を適用する．

２．提案システム

本研究の提案システムを Fig. 1 に示す．

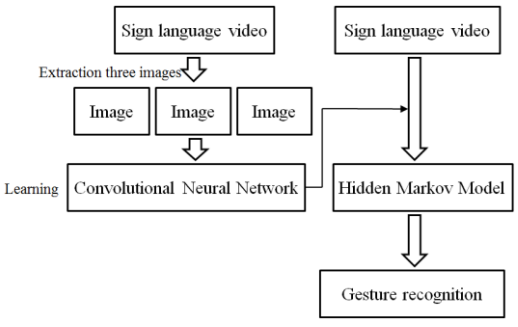

Fig. 1　Proposed System

まず，学習させたい手話の動画を撮影し，その動画から，人間の目で見たときに特徴的だと感じられる動作をしている瞬間の画像を 3 枚抜き出す．学習には膨大の画像が必要なため，同じ手話の動画を何度も撮影し，その度 3 枚の画像を抜き出す．抜き出した 3 枚の画像はそれぞれラベルづけをし，同じ瞬間の画像は同じラベルに分類する．画像が大量に集まったならば，その画像を Convolutional Neural Network に入れ学習させる．これにより，認識したい手話の特徴的な瞬間を学習することができる．つまり本研究は，撮影した手話に対して，学習画像との部分的な一致を見つけることで手話を認識する．

また，本研究は手話を認識する際，隠れマルコフモデルを適用する．隠れマルコフモデルは確率モデルの 1 つであり，出力列から最も尤度の高い状態列を計算するアルゴリズムである．

3． Convolutional Neural Network

Convolutional Neural Network は画像に対して Fig.2 に示すような Convolution と Max pooling を繰り返すことで画像の特徴を抽出していく手法である．

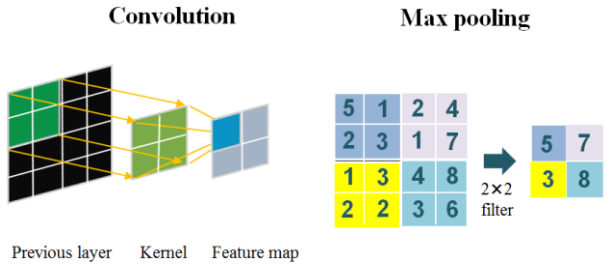

Fig.2 Convolution and max pooling

本研究で構築した Convolutional Neural Network を Fig.3 に示す．

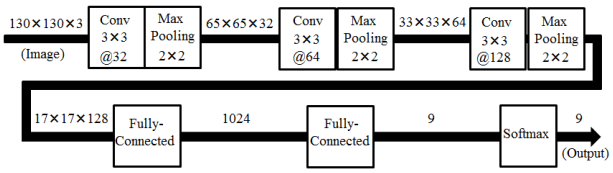

Fig.3 Convolutional Neural Network of this study

4． 本研究で判別する手話

日本手話には 50 音の「あ」「い」「う」「え」「お」などに対応する指文字がそれぞれ存在する．これらは指の形が固定的で動作を伴わない．一方で指や手を動かすことによって言葉を表現する場合もある．例えば「ありがとう」や「こんにちは」など日常会話で使う言葉は，動作を伴った手話で表現される．指文字を高い精度で認識する研究は既にあるため，本研究では動作を伴った手話を想定する．

具体的に，判別する単語は「ありがとう」，「ごめんなさい」，「よろしくお願いします」の 3 単語とした．この 3 単語を選んだ理由は 2 つある．1 つは日常的に誰もが使う言葉であり馴染みがあるため，もう 1 つはこの 3 つの単語に手刀を切るような似た動作が含まれているためである．画像から手話を判別する性質上，構築するシステムは似た動作を誤りなく区別することが求められると考えた．つまり，この 3 単語を判別することができれば他の似たような動作を含む手話も判別が可能であると推測した．

5． 学習

本研究では Google が公開する機械学習用ライブラリの Tensorflow を使用して学習を行った．

「ありがとう」，「ごめんなさい」，「よろしくお願いします」の 3 単語を学習させたときの学習精度を Fig.4 に示す．

なお，学習させた画像は約 2000 枚である．

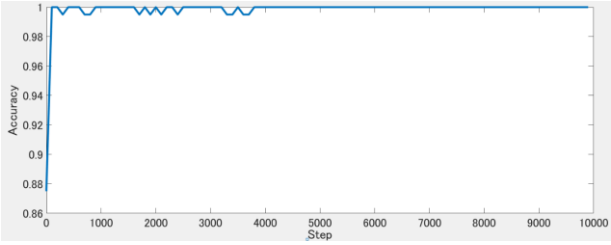

Fig.4 Learning Result

Fig.4 では 1 万 Step までの学習精度を表示しているが，実験を繰り返す過程で，誤判別をしてしまった画像については適切な正解を与えて追加学習を行ったため，合計すると 20 万 Step 程度の学習回数を行っている．

6． 検証

学習させた 3 単語について正しく学習できているか検証する．そのために Fig.5, Table1 に示す Web camera を使い，手話を撮影した．なお，学習のための画像も Fig.5 の Web camera を使用して用意している．

Fig.5 B910 HD WEBCAM

Table1 Specification of B910 HD WEBCAM

Size	34mm(H) × 94mm(W) × 80mm(D)
Weight	80g
Viewing Angle	78°
Max Frame Rate	720×1280pixel：30fps

本研究では学習する手話の画像それぞれにラベルを付

与している．「ありがとう」の画像については 0, 1, 2 のラベルを，「ごめんなさい」の画像については 3, 4, 5 を，「よろしくお願いします」は 6, 7, 8 という具合である．例えば「ありがとう」であれば，時間変化とともに 0, 1, 2 と手話が変化する．

手話を撮影して得られた判別結果を Fig.6 に示す．

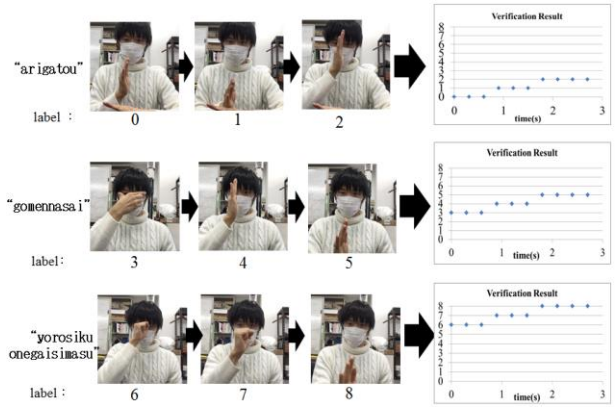

Fig.6 Verification Result

3 単語とも，時間変化と共に適切なラベルの数字に遷移しているため，手話の瞬間動作を正しく判別できていることがわかる．

7． 隠れマルコフモデル(HMM)

検証の項目で示した通り，人間の目で見れば判別結果の遷移から何の単語を示しているかが何となくわかるが，単語を推定するための客観的な根拠が要るため，HMM を採用した．HMM は近年音声認識の分野で大きな成果を挙げている確率的な状態遷移モデルである．本研究では Baum-welch アルゴリズムを用いてパラメータを推定し，Viterbi アルゴリズムによって手話を認識する．

Baum-welch アルゴリズムを本研究に適用するために，検証の項目で記した，判別結果の時間的遷移を利用する．例えば Fig.6 における「ありがとう」であれば，判別結果は 0, 0, 0, 1, 1, 1, 2, 2, 2, 2 と遷移している．この判別結果から，必要なパラメータである，初期状態確率 π，状態遷移確率 A，シンボル出力確率 B を計算し，HMM モデル $\lambda(=\{A, B, \pi\})$を作成する．ここで，状態の集合を $S=\{s_t\}$，シンボル列を $O=O_1, O_2, \ldots, O_T$，λ が O を出力する確率を $\Pr(\lambda|O)$，状態遷移が i から j に遷移するとしたとき各パラメータは以下の式で記述される[6].

$$\pi = \{\pi_i|\pi_i=\Pr(s_1=i)\} \quad (1)$$

$$A = \{a_{ij}|a_{ij}=\Pr(s_{t+1}=j|s_t=i)\} \quad (2)$$

$$B = \{b_j(O_t)|b_j(O_t)=\Pr(O_t=j|s_t=j)\} \quad (3)$$

認識対象の各単語毎の HMM を $\lambda_i(=\{A_i, B_i, \pi_i\})$とすると，手話の推定は以下の式で表される手話単語 R_k を認識結果として採用する．

$$k = \text{argmax}(\Pr(\lambda_i|O)) \quad (4)$$

8． 実験

撮影した手話が何の単語であるかを判別する実験を行う．方法は，Fig.5 の web camera で手話を行っている様子を動画で撮影し，出力列に対する尤度が最も高い単語を認識結果とする．

実験は，3 単語の手話をランダムに，合計 30 回行い，正しく認識できた回数から認識精度を求めた．なお，手話者は 3 人で実験を行った．その結果を Fig.7 に示す．

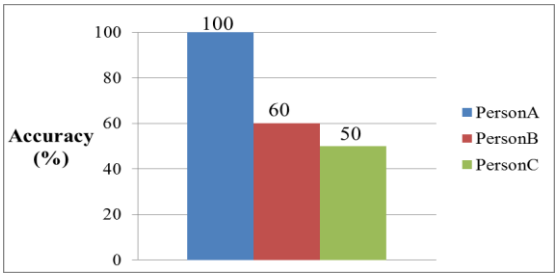

Fig.7 Recognition Accuracy

Fig.7 において PersonA の手話は精度 100%で認識できている．一方で，PersonB は 60%，PersonC は 50%と，認識精度にバラつきがでてしまっている．これは 1 人の人間(PersonA)の手話画像を学習させたためであり，多くの人間の手話画像を学習させれば，その分学習に時間がかかるなどの手間は増えるだろうが，精度は安定すると推測する．

Fig.7 において各単語ごとの認識精度を Fig.8 のようになった．

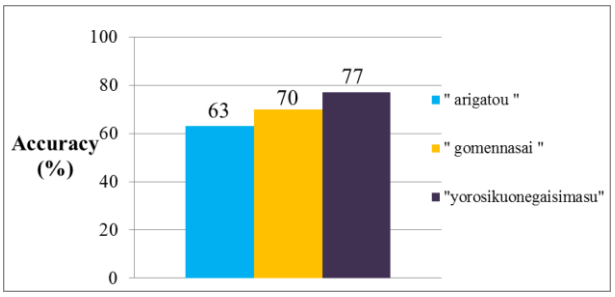

Fig.8 Recognition Accuracy by word

9． 結論

本研究では Convolutional Neural Network を用いて手話判別を行った．判別する手話は「ありがとう」，「ごめんなさい」，「よろしくお願いします」の 3 単語に絞り，その認識精度は，手話画像を学習させた人物であれば 100%であった．学習させていない人物が手話を行うと精度は下がってしまうが，学習させた人物においては 100%の精

度が出ていることから，多くの人物を学習させれば，手話者によって精度が変わることのないシステムが構築できると考えられる．

参考文献

1) 厚生労働省「平成 23 年生活のしづらさなどに関する調査（全国在宅障害児・者等実態調査）結果」，2013 年

2) みずほ情報総研ニュースリリース「手話者とのコミュニケーションを支援する手話認識システム」, 2014 年

3) 日立製作所ニュースリリース　手の動きと頭の動きを利用した「手話―日本語翻訳技術」を開発，2001 年

4) NAGI, Jawad, et al.:Max-pooling convolutional neural networks for vision-based hand gesture recognition, 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), p. 342-347, 2011.

5) KANG, Byeongkeun, et al.:Real-time sign language fingerspelling recognition using convolutional neural networks from depth map, 3rd IAPR Asian Conference, pp. 136-140, 2015

6) 大和淳司，大谷淳，石井健一郎:隠れマルコフモデルを用いた動画像からの人物の行動認識,電子情報通信学会論文誌 D, Vol.76, No.12,pp.2556-2563,1993.