H-034

DNN モデルの違いによる手話認識の精度比較に関する検討
Accuracy Comparison on Sign Language Recognition by Different DNN Models

菅野　成希†　　渡邊　滉大†　　亀山　渉‡

Naruki KANNO　Koudai WATANABE　Wataru KAMEYAMA

1. はじめに

手話認識の手法として、2DCNN と LSTM を組み合わせる手法と 3DCNN を用いる手法が存在する。

2DCNN と LSTM を用いる手法では、動画中のいくつかのフレームに対してそれぞれ 2DCNN で画像特徴量を抽出し、時系列順に LSTM へ入力し認識を行う。文献[1]では、RGB 画像に加えてオプティカルフロー画像、LSTM の代わりに Bidirectional-LSTM を用いて認識精度を比較している。

3DCNN を用いる手法では、動画中のいくつかのフレームに対し画像情報と時系列情報を含むデータを作成し、3DCNN へ入力することで認識を行う。文献[2]では、3DCNN に残差ブロック（Residual Block）[3]を適用し、RGB 画像とオプティカルフロー画像を用いた認識手法を検討している。

本稿では、手話単語認識において、上記の DNN モデル及び提案手法を比較検討する。

2. 比較実験概要

2.1 比較モデル

本稿では、2DCNN と LSTM を用いる文献[1]の手法に SPP（Spatial Pyramid Pooling）[4]を追加した SPP-RCNN と、文献[2]の手法に ConvLSTM2D[5]を追加した 3DCNN-ConvLSTM2D を提案手法として、以下の 7 個のモデルを用意して比較実験を行った。

入力が RGB 画像のモデル
- ① SPP-RCNN（図 1）
- ② 3DCNN[2]（図 2）
- ③ 3DCNN-ConvLSTM2D（図 3）

入力がオプティカルフロー画像のモデル
- ④ Temporal Stream ConvNet[6]（図 4）

入力が RGB 画像とオプティカルフロー画像のモデル
- ⑤ ①と④の統合モデル
- ⑥ ②と④の統合モデル
- ⑦ ③と④の統合モデル

①から④のそれぞれモデルの層構造について、図 1 から図 4 に示す。

図中の Conv2D、Conv3D、ConvLSTM2D は畳み込み層の種類を表しており、括弧内には順に畳み込みのウィンドウの大きさ（Kernel Size）、出力フィルタの枚数（Filter）、ウィンドウのストライド（Stride）を示している。また、Global Average Pooling、MP、SPP は、プーリング層の種類を表しており、括弧内には順にフィルタの大きさ（Filter）、フィルタのストライド（Stride）を示している。SPP では、入力を 1,4,16 とそれぞれ分割した格子内でマックスプーリングを行っている。Fully Connected は全結合層、LSTM は LSTM 層を表しており、括弧内には層中のノード数を示している。BN はバッチ正規化層、Softmax は Softmax 層を示している。⊕は、出力の足し合わせを示している。

オプティカルフロー画像を用いるモデルでは、Temporal Stream Convnet の Optical Flow Stacking を用いた。また、オプティカルフロー画像は、TV-L1[7]を適用して求めた。

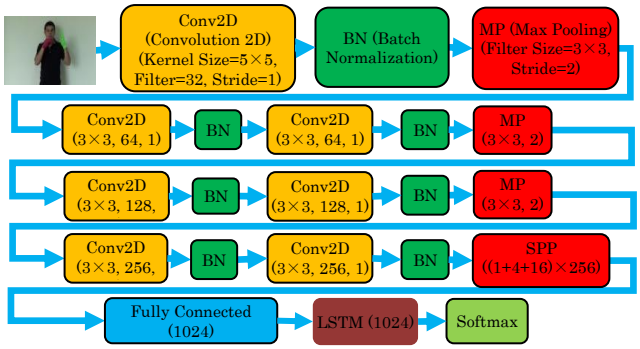

図 1　SPP-RCNN の層構造

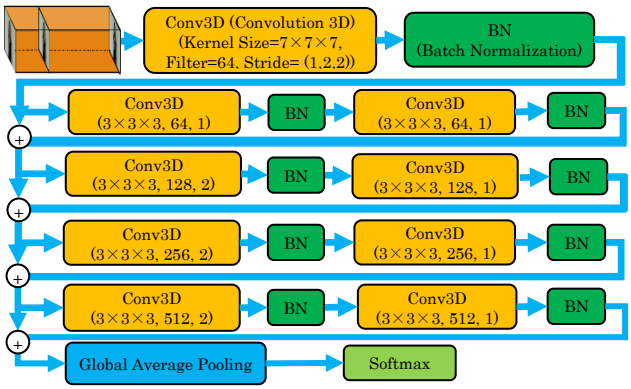

図 2　3DCNN の層構造

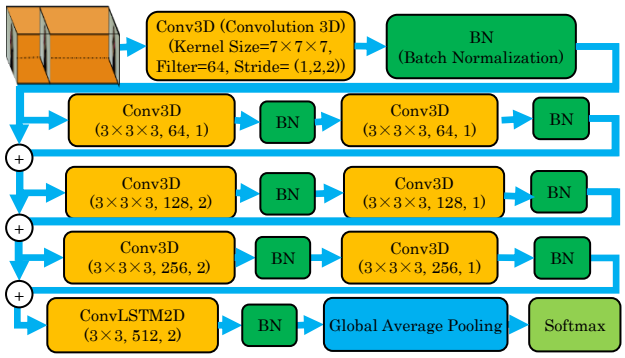

図 3　3DCNN-ConvLSTM2D の層構造

† 早稲田大学大学院基幹理工学研究科, Waseda Univ.
‡ 早稲田大学理工学術院, Waseda Univ.

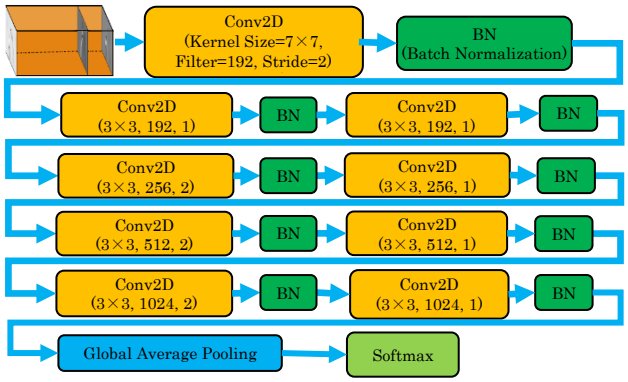

図 4　Temporal Stream ConvNet[6]の層構造

2.2 比較条件

比較実験のデータセットには、LSA64[8]の Still Moment Cut Version を使用し、各単語のうち 1 人目の手話者を検証用、2 人目から 10 人目の手話者を学習用とした。

モデルの作成には Keras のライブラリを使用し、ネットワークの学習は GeForce GTX 1080 を用いた。バッチサイズは 20 で、イテレーション数は 144 とした。活性化関数は ReLu、最適化関数は Adam を用いた。学習率は 10^{-4} とし、学習時の検証ロスが 10 エポック停滞するたびに学習率を 0.1 倍した。

入力は、RGB 画像の場合は、動画中の等間隔な 10 フレーム、オプティカルフロー画像の場合は動画中の等間隔な 20 フレームとした。モデルに入力する際は、それぞれの画像を 100×200 にリサイズした。

また、統合モデル（⑤から⑦）では、それぞれ学習済みのモデルの Softmax 層の一つ手前の層の LSTM 層または Global Average Pooling 層の出力の平均値を取り、ファインチューニングを行った。ただし、②と③のモデルを用いる場合では、Temporal Stream Convnet[6]の最終層のノード数と合わせるため、②または③側の Global Average Pooling 層にノード数 1024 の全結合層を追加した。

3. 実験結果と考察

全モデルでの検証正解率の比較を表 1 に示す。

RGB 画像のみを用いる認識手法では、③の 3DCNN-ConvLSTM2D が最も高い検証正解率で 97.5%となった。また、RGB 画像とオプティカルフロー画像を用いる認識手法では⑤の SPP-RCNN を用いる統合モデルが最も高い検証正解率で 97.8%となった。

ここで、RGB 画像を入力とした場合では SPP-RCNN が最も検証正解率が低いのに対し、RGB 画像とオプティカルフロー画像を入力とした場合では、SPP-RCNN を用いる場合が最も検証正解率が高くなった。これは、RGB 画像を入力とした場合では学習可能であった重みが、②と③の統合モデルを構築する際に全結合層を追加したことにより、最終層までに誤差伝播ができずに勾配消失してしまった可能性が考えられる。

4. まとめと今後の課題

手話単語認識では、SPP-RCNN と 3DCNN-ConvLSTM2D が有効であることを確認した。

しかし、本稿のオプティカルフロー画像の算出方法では、1 個の動画あたり平均 7 分もの処理時間がかかってしまうため、高速なオプティカルフロー画像の算出方法またはその他特徴量の検討が必要である。

また、本稿で使用したデータセットは、カラー手袋を身に着けていることにより、比較的識別が容易であった。しかし、これは一般的な条件ではない。そのため、特殊な制限を設けずに、より数の多い既存データセットの利用、または、それを作成することが今後の課題である

表 1　全モデルの検証正解率の比較

モデル	検証正解率 (%)
RCNN[1]:	
RGB 画像	注84
RGB 画像+オプティカルフロー画像	注91
RGB 画像+オプティカルフロー画像 +Bidirectional LSTM	注94
RGB 画像:	
① SPP-RCNN	93.8
② 3DCNN[2]	95.3
③ 3DCNN-ConvLSTM2D	97.5
オプティカルフロー画像:	
④ Temporal Stream ConvNet[6]	88.8
RGB 画像＋オプティカルフロー画像:	
⑤ 統合モデル（①＋④）	97.8
⑥ 統合モデル（②＋④）	97.2
⑦ 統合モデル（③＋④）	97.5

注:文献[1]の記述から

参考文献

[1] 松田 啓佑，山本 雅人，飯塚 博幸，“手話動作分類における RCNN モデルの性能評価と内部状態解析”，2018 年度人工知能学会全国大会論文集，4F1-OS-11c-04，2018 年
[2] 渡邊 滉大，亀山 渉，“RGB 画像とオプティカルフローを用いた 3DCNN による手話認識に関する検討”，信学技法，Vol.118, No.501，IE2018-152，pp.251-255，2019 年
[3] Kaiming He, et al, “Deep Residual Learning for Image Recognition”, IEEE CVPR, pp.770-778, 2016
[4] K. He, X. Zhang, S. Ren, J. Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”, IEEE TPAMI, Vol.37, No.9, pp.1904-1916, 2015
[5] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, Wang-chun Woo, “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting”, NIPS 2015, pp.802-810, 2015
[6] K. Simonyan, A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos”, NIPS 2014, pp.568-576, 2014
[7] Javier Sánchez Pérez, Enric Meinhardt-Llopis, Gabriele Facciolo, “TV-L1 Optical Flow Estimation”, Image Processing On Line, pp.137-150, 2013
[8] Ronchetti, Franco, et.al. “LSA64: A Dataset of Argentinian Sign Language”, XX II Congreso Argentino de Ciencias de la Computación (CACIC), 2016