関節軌跡情報を用いた手話認識手法の評価

Evaluation of Sign Language Recognition Method Using Joint Trajectory Information

我妻　嵩斗
指導教員　　坪川　宏

東京工科大学　工学部　電気電子工学科　坪川研究室

キーワード：手話識別，骨格認識，動作解析，機械学習

１．背景

国内における聴覚・言語障碍者は約 34 万人いるとされている．遠隔での会話が普及した昨今，他言語からの自動翻訳が求められている．その一方で手話の自動翻訳は普及が進まず，一般的に利用されているとはいえない．

２．目的

本研究では 1 台のカメラによる動画の手話識別を目的としてシステムの検討を行う．動画の背景などの変化に対応できる汎用性を持ったモデルを構築し複数動作の分類を行う．

３．関連研究

双方向 LSTM を RCNN に適用させモデルを構築した研究[1]では LSTM を用いている．これは時間の長さが固定され，指の形状を特徴として取り出すことが困難だといえる．CNN を用いたフレーム間差分画像による手話動作分類の研究[2]においても畳み込む次元数により手話動作の時間が固定である．画像差分のため撮影状況に左右され正答率の上がらないクラスが存在している．

４．研究概要

４．１　使用した機械学習モデル

本研究では一連の手話動作画像から OpenPose を用いて骨格座標を取得する．その座標から軌跡として画像に出力したものを学習データとする．生成した画像の分類は CNN を利用したモデルで行う．

本提案では一連の動作の中で動きのあるものを特徴として抽出できる点，色情報として時間方向の情報を保持できる点が挙げられる．さらに画像識別において不要な情報を排除することができ，識別率の向上が見込める．

Openpose とはカーネギーメロン大学で開発されたソフトウェアの名称である．特徴として 1 枚の画像から高精度な骨格の推定が行える点にある．

CNN は畳み込みニューラルネットワークの略称である．画像情報が格納された多次元配列から隣接する情報を保持できるため，色情報を含めた特徴を捉えることができる．この畳み込み層を中間層に持つニューラルネットワークである．

またバッチ学習と呼ばれる，学習画像を複数枚に分け学習効率を高める手法を用いている．

４．２ 学習用データ・テストデータ

カメラの画角の基準となる焦点距離は 14mm，18mm，25mm の 3 通りで撮影した．シャッタスピードは OpenPose の手形状識別が十分に行えると判断した 1/200 秒に固定した．手話動作の時間分解能の目安となるフレームレートは 60fps とした．

骨格情報から生成される画像には 2 種類存在する．画像面における関節の座標変化を時間ごとに輝度を変え繋いだ軌跡画像，同フレームにおいて取得した関節を骨格の相対関係がわかるように繋ぎ，時間ごとに輝度を変えつつ重ね合わせたもの

を重ね合わせ画像と表記する．

本検証では予備動作が似ていると判断した「上る」「理解する」，「重い」「大人しい」[3]を分類の対象とする．さらに手話動作以外の分類できない動作を「その他」としデータを作成した．図１，図２にそれぞれ動画から作成した手話動作から作成した画像を示す．

図１　軌跡画像

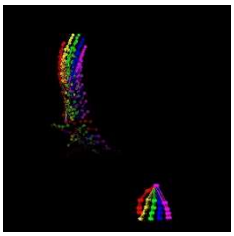

図２　重ね合わせ画像

学習の精度を確認するためのテストデータには，撮影された学習用データの中から学習に用いない群を抽出し各クラス 10 枚計 50 枚を使用した．

５．検証

軌跡画像と重ね合わせ画像それぞれでモデルのパラメータを検討し，その精度が最も高くなるモデルの比較をした．両手法の学習データと検証データについての学習曲線を以下に示す．

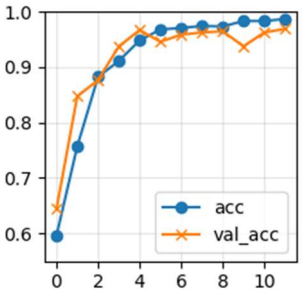

図３　軌跡画像

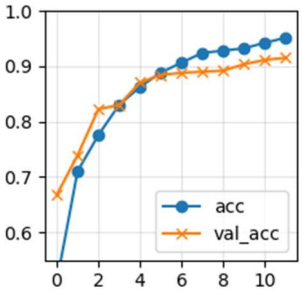

図４　重ね合わせ画像

このモデルでは「上る」「理解する」各 250 枚と 500 枚の「その他」画像を含めた学習データを用いている．上記画像から軌跡画像では検証データと学習データの正答率に差はない．重ね合わせ画像では正答率が 95%を超えず，過学習が起きている．

骨格情報から得られる画像のうち，両手法について同様の検討を行った．事前に用意した 50 枚のテストデータでの検証を以下の表に示す．

結果から片手動作の「上る」と「理解する」の予備動作の区別がついておらず，わずかな変化を区別することができていないと考えられる．

表１　軌跡画像で作成したモデルの評価
（縦軸：推論結果，横軸：実際のクラス）

軌跡	上る	理解する	重い	大人しい	その他
上る	10	0	0	0	0
理解する	0	9	0	0	0
重い	0	0	10	0	0
大人しい	0	0	0	10	0
その他	0	1	0	0	10

表２　重ね合わせ画像で作成したモデルの評価

重ね合わせ	上る	理解する	重い	大人しい	その他
上る	9	0	0	0	4
理解する	0	9	0	0	0
重い	0	0	10	0	0
大人しい	0	0	0	10	0
その他	1	0	0	0	7

軌跡画像では線形補完を行うため，手話動作の枚数が結果に影響しにくいと考えられる．

６．今後の課題

課題として，手話単語を十分な精度で翻訳できることを示すにはクラスの数が少ないこと，指文字など動作以外の手話を分類できるモデルではないことが挙げられる．また撮影時間よりも推論時間が長く，処理時間の短縮が課題となっている．

７．本手法の有効性

本研究において有効だと判断した軌跡画像による手話翻訳の精度について，先行研究[2]の最高正答率 94%に対し本研究では 98%となっている．また撮影条件に影響されない点で有効だといえる．

８　参考文献

[1] 北海道大学大学院情報科学研究科　松田　啓佑:　“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）.

[2] 埼玉大学大学院　西川友弘:　“畳み込みニューラルネットワークを用いたフレーム間差分画像による手話動作識別の検討”（2017）.

[3] 映像情報メディア学会誌 Vol.56，No.2　小渡　悟:“類似動作で意味が異なる手話単語の所要時間が弁別に及ぼす影響”（2002）.