4C-05

骨格座標推定による日本語対応手話の分類

上野　友裕†
金沢工業大学大学院 工学研究科 情報工学専攻†

1. はじめに

ろう者と聴者の間には少なからずコミュニケーションの隔たりが存在する．それらを解決するための手話翻訳の手法が従来から研究されている．従来から，特別な装置を用いて RGB 画像と深度情報を取得し翻訳する手法や，カラーグローブを用いる手法などが研究されてきた．しかしこれらの手法は日常での使用に適さないという問題点や，機材が高価な場合があるという問題点がある．また，日本語対応手話の動画を翻訳する研究は実用レベルに達していない．本研究では，最終的に手話をスマートフォンなどで動画として撮影することで手軽に，かつリアルタイムに翻訳することができる手法を開発することを目指している．そのための試みとして，RGB カメラと機械学習の手法のみを用いて，静止画として撮影された動きを伴わない指文字の分類を行い，それによって得られた知見をもとに手話の単語を表す動画を分類した．

2. 手話画像の分類

単眼 RGB カメラで日本語対応手話のかな指文字の静止画を撮影し，機械学習を用いて分類を行った研究[1]では，動きを伴わない 41 文字のかな指文字に対して 76.9%の正解率で分類を行えることが示された．また，その研究で得られた知見を，手話翻訳や手話動画の分類などにも用いることができる可能性があることを示した．

3. 手話動画分類の提案手法

この章では，手話動画の分類に関する提案手法について述べる．提案手法では，26 単語の手話を RGB カメラでそれぞれ撮影した動画の分類を行った．骨格座標などの特徴量の抽出には OpenPose[2]を用い，分類器にはニューラルネットワークを用いた．

3.1 データセットの作成

手話動画に関するデータセットを作成した．協力者の数は 14 名で，それぞれの協力者に対し 26 単語分の動画を一単語ずつ撮影した．単語は手話の教材[3]に掲載されている，スポーツなどの得意なことについて自己紹介する場面を想定した 10 文以内の例文を参考に選択した．撮影した単語を以下の図 1 に示す．

けれども，する，どちら，ゴルフ，サッカー， スポーツ，テレビ，一緒，上手，休み，会社，場所， 好き，学校，得意，教わる，朝，父親，疲れる，私， 練習，行く，見る，車，遠い

図 1 撮影した単語のリスト

また，ある協力者を撮影する際，11 名に対しては正面からと左からの視点で，残りの 3 名に対しては正面と右からの視点で撮影を行った．ゆえに，動画の総本数は 14(名)×26(単語)×2(方向)= 728(本)となる．動画には様々な物体が映るなど背景のノイズ情報も存在するが，そのまま学習や推論に用いることにした．

3.2 手話動画分類に用いる特徴量

手話動画の分類では，先行研究[1]で示された指の骨格に関する特徴量に加え，クロッピングされた領域内における指の骨格の相対的な長さに関する特徴量と，全身の骨格に関する特徴量を追加した．全身の骨格に関する特徴量は 2 種類ある．一つ目の特徴量としては，全身の座標の最大値と最小値でクロッピングし正規化を行ない，骨格座標のクロッピングされた領域における相対的な位置を求めたものを使用した．この特徴量は合計で 18 個である．二つ目の特徴量としては，全身の骨格における関節角を用いた．関節角の特徴量の数は合計で 16 個である．それらのイメージ図を以下の図 2 に示す．

Classification of Japanese Sign Language using Pose Estimation.
†Tomohiro UENO
†Graduate School of information and Computer Engineering, Kanazawa Institute of Technology

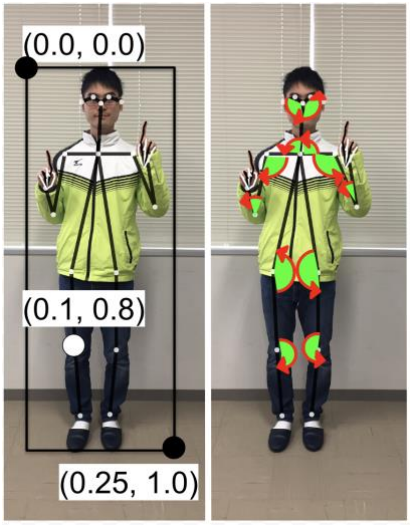

図 2 全身の骨格のクロッピングされた領域における相対的な座標（左）と全身の骨格の関節角（右）に関する特徴量のイメージ

3.3 手話動画分類に用いるモデルの構成

手話動画の分類にはBidirectional LSTMもしくはLSTMを用いた．そのネットワーク構成を以下の図3に示す．LSTMとBidirectional LSTMのどちらを用いるかや，学習率，エポック数，LSTMの隠れ層の内部で用いるユニット数と活性化関数の種類，最適化手法などはハイパーパラメータサーチを行うことで選択した．

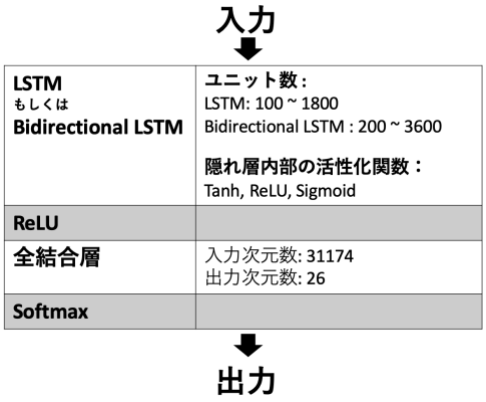

図 3 モデルのネットワーク構成

なお，入力となる手話動画のフレーム数は一つの動画につき 286 フレームで固定となっている．

4. 手話動画分類の実験

この章では，提案手法の性能を評価するための実験について述べる．

4.1 実験の概要

手話動画の本数が少ないため，Nested-Cross-Validationを用いて評価する．各Foldには2名分の動画を全て割り当て，合計 7 Fold に分けた上で 7 個のそれぞれのモデルのハイパーパラメータを求めた．最終的なモデルの性能は，求められたハイパーパラメータを用いて学習を行った 7 個のモデルの精度を平均することや，混同行列などにより評価した．

4.2 実験結果

ネットワーク構成は LSTM が三つの Fold で最適とされ，Bidirectional LSTM が四つの Fold で最適とされた．テストデータにおける正解率の全ての Fold に対する平均値は，正面から撮影された動画に対して 80.1%，左から撮影された動画に対して 77.7%，右から撮影された動画に対して 60.2%，三方向の全ての動画に対しては 78.2%の正解率となった．左から撮影したデータは正面から撮影されたデータに比べ 2.3%程度の劣化に留まっている．右から撮影したデータは３名のみから構成されており，そのことが原因で正解率が下がったものと思われる．分類に失敗した動画の特徴として，暗い場所で撮影されていることや，遠くから撮影されていることなどが挙げられる．

5. 考察

手話動画の分類に，LSTM とあらかじめ多様かつ大量のデータセットで学習を行なった OpenPose を用いることで，背景に物体が写るなど多少のノイズ情報がある場合でも分類可能であることがわかった．また，OpenPose を用いることで入力データの次元を削減し，少ないサンプル数のデータを用いた場合に起こりやすい過学習の問題を回避することができたと思われる．

6. まとめ

本研究では，手話翻訳を実現させるための試みとして，RGB カメラと骨格座標推定の手法を用いることにより，手話動画の分類を行なった．

手話動画の分類では，26 単語を表す手話動画を，78.2%の正解率で分類することができた．

今後，実用化する場合は大量の手話データを様々な角度から撮影する必要がある．

参考文献

[1] 上野友裕，“OpenPose による日本語対応手話のカナ指文字の分類，” in 2019 年度電気・情報関係学会北陸支部連合大会，F2-31，2019.
[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” ArXiv181208008 Cs, May 2019.
[3] 全国手話研修センター，DVD で学ぶ手話の本 5 級．東京：中央法規出版，2016.