K-022

手話認識における動きの個人差を吸収するためのデータ拡張手法

Data Augmentation Methods for Accommodating Individual Movement Variations in Sign Language Recognition

本郷 望実† 後藤 啓介† 和田 直哉† 村上 文雄†
Nozomi Hongo Keisuke Goto Naoya Wada Fumio Murakami

1. はじめに

手話は聴覚障害を持つ人々にとって重要なコミュニケーション手段であり，手や指，顔の表情，身体の動きなどを組み合わせて，視覚的に情報を伝達する言語である．手話話者と手話を話せない人との円滑なコミュニケーションを支援することを目的とした，手話の自動認識（以下，手話認識）と機械翻訳に関する研究が盛んに行われている．これまでに，機械学習や深層学習を用いた様々な手話認識モデルが多数提案されてきたが，依然として高精度な手話認識の実現には多くの課題が残されている[1].

手話認識モデルの性能を向上させるためには，発話者の服装や体格，文脈や個人差による動作のばらつき，撮影環境の違いといった，認識の障害となる要素が十分に盛り込まれた，大規模で高品質な手話動画のデータセットが不可欠である．しかしながら，公開されている手話動画のデータセットの多くは，単一の手話単語を撮影した動画を収録したものであり，実際の手話で用いられる連続の手話単語（以下，連続手話単語）を撮影した動画データの数は非常に限られている．また，新たに手話動画のデータセットを作成するには，撮影環境の整備や手話話者の確保，アノテーション作業など多大なコストと労力が必要となる．このようなデータ不足の問題は，特に日本手話のようなリソースが限られている言語において顕著であり，研究開発の大きな障壁となっている．

我々は，手話動画から手話に関する身体部位のランドマーク座標の時系列データ（以下，ランドマーク座標データ）を入力とした連続手話単語の認識システムの構築に取り組んでいる．手話表現は，発話者による動作の速度や大きさの違い，利き手の違い，さらには個人の癖など，様々な個人差が存在するため，これらに対応したモデルの構築が求められる．

そこで本研究では，手話表現の速度変化や動作の大きさ，利き手の違いといった個人差に対応可能なランドマーク座標データの拡張手法を新たに提案する．さらに，我々が独自に作成した日本手話データセットに対してこれらのデータ拡張手法を適用し，連続手話単語認識の既存手法である Conformer [2]を用いた手話認識を行い，提案手法の有効性を検証した．

本研究は，限られたデータ資源下においても手話認識モデルの性能を向上させるための新たなアプローチを示すものである．

2. 関連研究

本章では，手話認識における代表的な認識手法と，特に本研究と関連の深いデータ拡張技術について述べる．

2.1 手話認識手法

手話認識の研究は，入力データの種類や認識対象によって多様なアプローチが試みられている[3]．これらは主に，RGB 画像やデプス画像などの手話動画から直接特徴を学習し End-to-End で認識を行う手法と，動画から手指や身体のランドマーク座標を抽出し，その座標系列を時系列モデルに入力して認識する手法の 2 つに大別される．深層学習技術の発展は，手話認識の分野に大きな影響を与え，いずれの手法においても性能向上に貢献している．

前者の End-to-End 手法では，CNN (Convolutional Neural Network) が RGB 画像やデプス画像から空間的な特徴を抽出するために広く用いられている．CNN は手指の形状や手の動きの局所的なパターンを捉えるのに有効であり，抽出された特徴は後段の時系列モデルへの入力となることが多い．

一方，本研究で主に焦点を当てるランドマーク座標に基づく手法では，抽出されたランドマーク座標の系列が，手話の時系列的な特徴を捉えるモデルへの入力となる．従来，このような手話のダイナミクスをモデル化するために，RNN (Recurrent Neural Network)，特に LSTM (Long Short-Term Memory) や GRU (Gated Recurrent Unit) が利用されてきた．これらのネットワークは，ビデオフレームから抽出された CNN 特徴や，ランドマーク座標の系列を入力として訓練される．

近年では，自然言語処理の分野のモデルである Transformer が，Attention 機構により系列データ内の長距離依存関係を効果的に捉えられるため，手話認識にも応用されている[4]．稲田ら[2]は音声認識の分野で提案された Conformer [5]を用いた手話単語分類手法を提案している．Conformer は Transformer と CNN を組み合わせた深層学習のネットワークであり，時系列データの全体に対する特徴抽出および局所的な特徴抽出が可能な点を特徴とする．

2.2 手話認識のデータ拡張手法

MediaPipe [6]のような高精度な人体ランドマーク推定ツールの登場により，手話動画から抽出された手や身体のランドマーク座標を入力とする手話認識手法が活発に研究されている．ランドマーク座標データは，背景や照明の変化に対する頑健性が比較的高く，またデータ量が画像そのものよりも小さいため，計算コストを抑えられる利点がある．

† 京セラ株式会社　KYOCERA Corporation

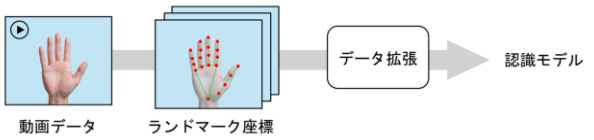

図 1　手話認識の流れ

図 2　ノイズを付加する関節

ランドマーク座標に対するデータ拡張手法としては，座標全体の平行移動，回転，スケーリングといった幾何学的な変換や，座標値へのガウスノイズなどのランダムノイズ付加が基本的なアプローチである[7]．これらの変換は，カメラや話者の位置の違い，センサーノイズなどに対するモデルの頑健性を高めることを目的とする．

より高度な手法として，中村ら[8]は，日本手話データセットである KoSign を用いたデータ拡張手法を提案している．MediaPipe を用いて取得された 3 次元ランドマーク座標から指の関節角度を求め，ガウス分布に従ったノイズを付加した後，順運動学により角度データを座標データに変換することによってデータ拡張を実現している．本研究では関節角度への変換に極座標変換を用い，ノイズ付加後の角度を関節可動域内に調整する独自のステップにより，過度に不自然な姿勢の生成を防いだデータ拡張を目指す．

3. 提案手法

手話認識の流れを図 1 に示す．提案手法は，MediaPipe 等の人体ランドマーク推定ツールによって動画データから取得された 3 次元ランドマーク座標データに対して左右反転，オクルージョンマスク，関節角度ノイズ，速度変換，動作スケーリングを適用するデータ拡張手法である．各手法は独立した処理として設計されているため，手法の任意の組み合わせが可能である．

3.1 左右反転

手話表現には左右対称なものと左右非対称なものが存在する．特に，左右非対称な表現では話者の利き手が左右どちらであるかによって，動作を主に行う手が異なる場合がある．データセットにおいて，特定の片方の手を利き手としたデータが偏って多く含まれる場合，少数派の利き手パターンの表現の学習が不十分となり，認識精度の低下を招く恐れがある．そこで，3 次元ランドマーク座標$P = \{p_j = (x_j, y_j, z_j) \mid j = 1,2,\dots,m\}$を$x$軸に関して反転させ，$P' = \{p'_j = (-x_j, y_j, z_j) \mid j = 1,2,\dots,m\}$とすることで，利き手に依存しないデータを作成でき，モデルの認識精度の向上が期待できる．

ただし，「右」「左」などの方向そのものに意味を持つ単語集合については，左右反転を適用すると意味が変わってしまう．このため，単語ごとに反転可能かどうかを考慮する必要がある．

3.2 オクルージョンマスク

オクルージョンマスクは，オクルージョンと呼ばれる，手の位置や身体の向きによって手指が隠れる状況を模擬するため，部分的なマスク処理を行う手法である．各フレーム t の 3 次元ランドマーク座標集合$P_t = \{p_i = (x_i, y_i, z_i) \mid i = 1,2,\dots,m\}$について，正面から見て$xy$平面上での距離$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$が閾値$D$以下となる点対集合$S = \{(p_i, p_j) \mid d_{ij} \le D, i \ne j\}$を抽出する．その中からランダムに$n$組$\{(p_{a_1}, p_{b_1}), (p_{a_2}, p_{b_2}), \dots, (p_{a_n}, p_{b_n})\} \subset S$を選び，各組$(p_{a_k}, p_{b_k})$において$z_{a_k} > z_{b_k}$であれば$p_{b_k} \leftarrow (0,0,0)$，そうでなければ$p_{a_k} \leftarrow (0,0,0)$と置換することで，重なりによる隠れを擬似的に再現する．この処理により，実際の手話認識において生じる部分的なオクルージョンへの頑健性を高めることが期待される．

3.3 関節角度ノイズ

関節角度ノイズは，手指の関節角度に対してノイズを付加する手法である．ノイズを加える関節を図 2 に示す．まず，3 次元ランドマーク座標データ$P = \{p_i = (x_i, y_i, z_i) \mid i = 1,2,\dots,m\}$を極座標系を用いて関節角度データ$\Theta = \{\theta_j \mid j = 1,2,\dots,n\}$へ変換する．次に，変換後の各関節角度$\theta_j$に対して平均$\mu = 0$，標準偏差$\sigma$のガウシアンノイズを加え，$\theta'_j = \theta_j + \mathcal{N}(0, \sigma^2)$とする．ノイズ付加後は，関節$j$の可動域$[\theta_{j,min}, \theta_{j,max}]$を逸脱しないように角度を調整し，$\theta'_j = \max(\min(\theta'_j, \theta_{j,mqx}), \theta_{j,min})$とする．最後に，ノイズ付加後の角度データ$\Theta' = \{\theta'_j \mid j = 1,2,\dots,n\}$を再度座標データへ逆変換することで，ノイズを反映させたランドマーク座標$P' = \{p'_i = (x'_i, y'_i, z'_i) \mid i = 1,2,\dots,m\}$を得る．これにより，手指の形状に自然な揺らぎを持たせたデータ拡張が実現でき，認識モデルの汎化性能向上が期待される．

3.4 速度変換

速度変換は，手話動作の時間的特性を変化させることで，実環境における話者間の速度変動を模擬するデータ拡張手法である．手話は話者によって表現速度に顕著な個人差が生じることが知られており，同一の意味内容であっても，熟練者は素早く流暢に，初心者はやや緩やかに表現する傾向がある．このような速度変動に対して頑健な認識モデルを構築するためには，訓練データにおいても同様の変動を含有させることが効果的である．

本研究では，速度変換について，フレーム補間とフレーム間引きの手法を用いて実現する．フレーム補間は動画の連続するフレーム間に新たなフレームを線形または非線形補間によって生成する．フレーム間引きでは一定間隔でフレームを抽出することにより生成する．これらの手法を組み合わせることで，様々な速度変動を持つデータセットを生成し，モデルの時間的変動への頑健性を向上させることが可能となる．

表 1　作成したデータセットの概要

言語	日本手話
FPS	30
データ解像度	フル HD (1920×1080px)
話者数	10
文種類数	87
平均文長	5.1
語彙数	43

3.5 動作スケーリング

手話動作を表現する際の動作の大きさは話者によって異なる．動作スケーリングは動作の大きさの変動を模したデータ拡張手法である．左右それぞれの肩の座標と対応する手首の座標を結ぶベクトルを基準とし，各手指のランドマーク座標を$p'_{\text{hand}} = p_{\text{shoulder}} + \alpha \cdot (p_{\text{hand}} - p_{\text{shoulder}})$と変換する．ここで，αはスケーリング係数であり，左右で同一の値を用いることで動作の一貫性を保持する．この変換により，各肩を起点として左右の手の動きの大きさを制御的に変化させることができる．

4. 実験

連続手話単語認識をタスクとして評価実験を行った．本章では，評価実験の詳細について述べる．

4.1 使用したデータ

日本手話を日常言語として用いている人を話者として独自にデータセットの作成を行い，実験に使用した．表 1 に作成したデータセットの概要を示す．話者は男女 5 名ずつで，20 代から 70 代まで偏りなく含まれている．話者には 1 語以上の手話単語で構成される手話文を 1 文につき 5 回繰り返し表現してもらい，RGB カメラで正面から撮影を行った．データ撮影後，3 名のアノテータが各手話文および各手話単語を対象に，それらが示されている時間帯をフレーム単位でラベル付けした．

4.2 認識モデル

本研究では，稲田らが提案した手話分類手法で用いられている ConformerBlock (CBI) および ConformerEncoder (CEn) をエンコーダとして評価実験に使用する．これらのモデルは手話の映像に対して MediaPipe によって検出した姿勢および手のランドマーク座標を入力データとし，手話単語の判別に用いる特徴量を抽出する．

入力データには手話の映像を構成するTフレームそれぞれについて，MediaPipe で取得した 3 次元座標を肩幅で正規化した値を用いる．MediaPipe によって取得する 3 次元座標は手首，手指の 42 箇所に姿勢の手首 2 箇所を加えた 44 箇所の座標情報を 132 次元の値として扱う．正規化は，両肩の中心が原点となるよう座標を平行移動させた後，x座標上で両肩の距離が 1 となる係数を各次元に対して乗じることで行う．モデルの出力は，連続手話文中の各単語ラベルの系列である．

表 2　データ拡張手法のパラメータと拡張倍数

拡張手法	パラメータ設定	拡張倍数
オクルージョンマスク	$D = 0.05$，$n = 2$（シード 2 種）	6
関節角度ノイズ	$\sigma = 1$（シード 2 種）	6
速度変換	2/3 倍，3/2 倍	6
動作スケーリング	$\alpha = 0.9$，1.1	6
全手法組み合わせ	上記すべて	18

4.3 実験条件

評価実験における各データ拡張手法の設定および拡張倍数は，表 2 に示す通りである．左右反転の処理を事前に適用し，元データおよび左右反転データの両方に各手法を適用した．全手法組み合わせは，元データと左右反転データに対し，4 種類の拡張手法を適用した．各手法はデータに対し 2 つのバリエーションを生成する．これにより，元データ，左右反転データ，およびこれら 2 種類のデータそれぞれに 4 手法を 2 バリエーションずつ適用して得られる 16 種類のデータが作成され，データ数は元データの 18 倍になった．なお，使用したデータには左右反転すると意味の変わる単語は含まれていなかった．

データセットに含まれる 10 名の話者のうち，8 名を学習用データ，1 名を検証用データ，1 名を評価用データに分け，10 分割交差検証で評価を行った．データ拡張手法は学習用データのみに適用し，評価指標にはレーベンシュタイン距離に基づく Word Error Rate（WER）を用いた．

認識モデルのデコーダ方式には，RNN Transducer (RNNT) を採用した．エンコーダへの入力系列に対しては，畳み込み処理によるサブサンプリングを行い，系列長を短縮した．このサブサンプリングの間隔は 4 とした．エンコーダおよびデコーダにおける隠れ層の次元数は，それぞれ 36 次元に設定した．また，エンコーダ内部で使用する Attention 機構の層数は 1 とした．ドロップアウト率は，(0.2, 0.1, 0.0) の値を適用した．

モデルの学習においては，学習エポック数を 1,000，バッチサイズを 64 とした．最適化手法には AdamW を用い，そのハイパーパラメータは$\beta_1 = 0.9$，$\beta_2 = 0.98$，weight decay 係数を1×10^{-3}に設定した．学習率は初期値を2×10^{-4}とし，最初の 1,000 ステップで学習率を線形に増加させるウォームアップ期間を設けた学習率スケジューラを適用した．

4.4 実験結果

実験結果を表 3 に示す．WER の数値は交差検証 10 回分の平均である．いずれの条件も拡張前と比較し提案手法を適用した拡張後は WER の値が下がっており，精度向上していた．また，拡張前と拡張後でt検定を行い，すべての条件において 5%水準での有意差があることが確認された．

4.5 考察

本研究では，手話認識モデルの精度向上を目的として，複数のデータ拡張手法を提案し，その有効性を検証した．実験結果は，提案したいずれのデータ拡張手法も，単一ま

表 3　実験結果

条件	WER[%]
拡張前	4.92
オクルージョンマスク	1.59*
関節角度ノイズ	1.55*
速度変換	1.81*
動作スケーリング	1.88*
全手法組み合わせ	1.16*

* 拡張前と比較して 5%水準で有意差あり

たは組み合わせることで，拡張前と比較して WER を有意に改善することを示した．これは，データ拡張が手話認識モデルの頑健性および汎化性能の向上に寄与することを示唆している．

提案手法の一つである左右反転は，話者の利き手に起因するデータセット内の偏りを緩和し，モデルが利き手に依存しない普遍的な動作特徴を学習することを意図した拡張である．実験の全拡張条件に本処理を加えたが，これは事前の予備実験において，左右反転処理を適用することで左利きの表現を含む話者の精度が大きく改善されたためである．このことから，本処理は特に利き手が少数派である話者の認識精度改善に寄与し，モデルの汎化能力向上に有効であったと考えられる．

オクルージョンマスクは，手話動作中に頻繁に発生しうる部分的な手の隠れを模擬することを目的としている．このような拡張データを学習することで，モデルは部分的な情報欠損が生じた場合でも，残された情報から全体を効果的に推測する能力を獲得し，実際の認識シーンにおける頑健性の向上に繋がったと考えられる．

関節角度ノイズは，話者ごとの微妙な手の形状や動きの癖，あるいは同一話者による表現の揺らぎといった，手話の自然な変動を再現することを目的としている．関節角度空間でのノイズ付加と可動域制限を通じて生成されたデータは，モデルがこれらの微細な差異への過学習を防ぎ，より本質的な動作パターンに注目するよう促したと考えられる．これにより，未知の話者やわずかに異なる表現スタイルに対する認識の安定性向上が期待される．

速度変換は，手話表現に見られる話者間や発話状況による速度の変動に対応するため，動作の時間的スケールを変化させることを目指した．フレーム補間と間引きによって生成された多様な速度のデータを学習することで，Conformer が持つ時間的コンテキスト理解能力をさらに強化し，幅広い速度の入力に対する認識の安定性を高めたと推察される．

動作スケーリングは，話者の体格や表現空間の使い方の違いなどから生じる，手話動作の大きさのバリエーションを模擬するために適用された．肩を基準とした手の動きのスケールを調整することにより，モデルは動作の絶対的な大きさではなく，相対的な動きのパターンや形状といった，より本質的な特徴に焦点を当てて学習することが可能となり，異なるスケールの手話表現に対する頑健性が向上したと考えられる．

全手法組み合わせでは，上記で述べた個々の変動要因が複合的に発生する，より現実世界に近い複雑な状況をシミュレートした．各拡張手法がそれぞれ異なる側面からモデルの頑健性を高めるため，これらを組み合わせることで相乗効果が生まれ，入力データに対して安定した性能を発揮できるようになったと推察される．実験結果において全手法組み合わせが実験の条件の中で改善幅が最も大きかった．これは複数手法の組み合わせにより各手法がモデルの汎化能力を多角的に強化した結果と言える．

5. おわりに

本研究では，手話表現の動きの個人差に着目し，ランドマーク座標データに対する 5 種類のデータ拡張手法—左右反転，オクルージョンマスク，関節角度ノイズ，速度変換，動作スケーリング—を提案した．これらの手法は，手話表現における自然な変動を模擬し，認識モデルの汎化性能を向上させることを目的としている．提案手法を我々が作成した日本手話データセットに適用し，既存の手話単語認識モデルの訓練に使用した結果，すべての拡張手法についてモデルの認識精度の向上が確認された．また，複数の拡張手法を組み合わせることで，単一の手法を適用した場合よりもさらに高い精度向上が得られることも明らかになった．

今後の課題として，他の手話データセットへの適用による一般性の検証が挙げられる．また，本研究ではランドマーク座標データに焦点を当てたが，画像ベースの手話認識にも応用可能なデータ拡張手法の開発も重要である．

謝辞

本研究を遂行するにあたり，貴重な助言をいただきました豊田工業高等専門学校木村教授に謝意を表します．

参考文献

[1] M. De Sisto, V. Vandeghinste, S. E. Gomez, M. De Coster, D. Shterionov, and H. Saggion, “Challenges with Sign Language Datasets for Sign Language Recognition and Translation,” Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 2478–2487 (2022).

[2] 稲田 渉, 木村 勉, “Conformer を用いた手話単語認識に関する研究,” 手話コミュニケーション研究会論文集, Vol.4, pp. 1-8 (2021).

[3] T. Tao, Y. Zhao, T. Liu, and J. Zhu, “Sign Language Recognition: A Comprehensive Review of Traditional and Deep Learning Approaches, Datasets, and Challenges,” in IEEE Access, vol. 12, pp. 75034-75060 (2024).

[4] N.C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, “Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10023-10033 (2020).

[5] A. Gulati, J. Qin, C. C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented Transformer for Speech Recognition,” arXiv preprint arXiv:2005.08100 (2020).

[6] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann, “MediaPipe: A Framework for Building Perception Pipelines, ” arXiv preprint arXiv:1906.08172 (2019).

[7] A. R. Verma, G. Singh, K. Meghwal, B. Ramji, and P. K. Dadheech, “Enhancing Sign Language Detection through Mediapipe and Convolutional Neural Networks (CNN) ,” arXiv preprint arXiv: 2406.03729 (2024).

[8] 中村 友里也, 荊 雷, “KoSign データを拡張するための Data Augmentation 手法の検討, ” NII-IDR ユーザフォーラム 2022, P20 (2022).