敵対的学習を用いた骨格に基づく手話認識のためのデータ拡張

中村友里也† 荊 雷†

† 会津大学大学院コンピュータ理工学研究科 〒 965-8580 福島県会津若松市一箕町鶴賀
E-mail: †{m5261142,leijing}@u-aizu.ac.jp

あらまし 手話は，手の形や動き，顔の表情，目線などで表現される視覚言語である．近年，深層学習の発展に伴い，視覚ベースの手話認識は活発な研究分野であり，多くの研究が行われている. しかし，手話データは収集が難しく，多くのデータセットにおいてデータ不足や不均衡データの問題があり，機械学習において過学習や精度低下の原因となる. そこで，本研究では，視覚ベースの手話認識に着目し，人物姿勢推定で得られる骨格データを用いたデータ拡張手法を提案する 具体的には，敵対的学習を利用して，データ拡張と機械学習の２つの独立したプロセスを共同で学習する. さらに，アブレーション研究により，提案手法の評価実験を行った.

キーワード 日本手話，手話認識，骨格データ，敵対的学習

1 はじめに

手話は，手や指，顔の表情，口の形など，体全体の動きで表現する視覚言語である. また，言語や聴覚が不自由な人にとって手話は話し言葉に代わる便利なツールであり，コミュニケーションにおける障壁を減らすことができる. 一方，普段手話をコミュニケーションの手段として利用していない人にとって手話を理解することは容易ではない. そのため，様々な手話を認識するアルゴリズムで構築されたプラットフォームを開発することは大きな社会的意義がある.

近年，ディープラーニングの発展に伴い，視覚に基づく手話認識は活発な研究分野であり，多くの研究が行われている [1]. ディープラーニングモデルの成功のための重要な要素は，大量の学習データを利用することにある. しかし，手話のデータ収集とアノテーションにはコストがかかるため，多くの手話データセットにおいてデータ不足や不均衡データの問題がある. そして，これは機械学習において過学習や精度低下の原因となる.

そこで，本研究では，視覚ベースの手話認識に着目し，人物姿勢推定で得られる骨格データを用いたデータ拡張手法を提案する. データ不足や不均衡データの問題に対する一般的な解決策として， 学習データに対してあらかじめ定義された範囲でランダムに変換（例えば，回転，拡大と縮小，ノイズ付加など）を行うデータ拡張がある. しかし，このデータ拡張には問題があり，データの違いを考慮することなく， 学習データ全体に同じデータ拡張を適用する. そのため，機械学習モデルにとって難しすぎる又は簡単すぎるデータを発生する可能性がある.

以上の背景から， 敵対的学習を利用して，データ拡張と機械学習の２つの独立したプロセスを共同で学習する. 具体的には，生成器として機能するデータ拡張モデルは識別器として機能する機械学習モデルの弱点を探るように学習し，同時に識別器は敵対的なデータの特徴を学習する.

2 関連研究

2.1 手話認識

手話認識を行う手法はデバイスベースのアプローチと視覚ベースのアプローチの 2 種類に大別される. デバイスベースは一般にセンサーを備えた電子手袋などのユーザーとシステムを直接つなぐ計測装置を利用する. 視覚ベースは 1 台以上のカメラを用いて得られた手話のデータを利用する. デバイスベースは効率的であることが特徴であるが，ユーザーとシステムをつなぐ際に煩雑なデバイスを装着する必要があるため，現段階での実用性は限定的である. しかし，ビジョンベースでは，この問題は発生せず，実生活に即した実用的なものである.

2.2 人物姿勢推定

人物姿勢推定は人物が映った画像から人体の各関節（腕，頭，胴体など）の座標を識別・分類する方法であり，ヘルストラッキング，手話認識，ジェスチャー制御などの様々なアプリケーションで中心的な役割を果たしている. 本研究では，3 次元における人物姿勢推定手法について述べる．2 次元画像から直接 3 次元関節座標を推定するために，数多くの深層学習技術が提案されてる [2–4]. また，他の手法では，この問題を，まず画像から 2 次元関節座標の推定と，それに続き推定した 2 次元関節座標に基づく 3 次元関節座標の推定に分解している.2 次元関節座標は，CPM [5]，Mask-RCNN [6] などの技術を用いて得ることができる.

2.3 敵対的学習

敵対的生成ネットワーク (GAN) は生成器と識別器の間でミニマックスゲームを行うように設計されている.Yu and Graum らは画像比較学習時の監視の疎密を克服するために，GAN を使用して画像ペアを合成する手法を提案した [17].A-Fast-RCNN は，物体検出のための変形を生成するために GAN を用いる [18]. また，人間の姿勢推定における GAN の応用として，姿勢推定ネットワークを生成器として扱い，識別器を用いて追加の監視

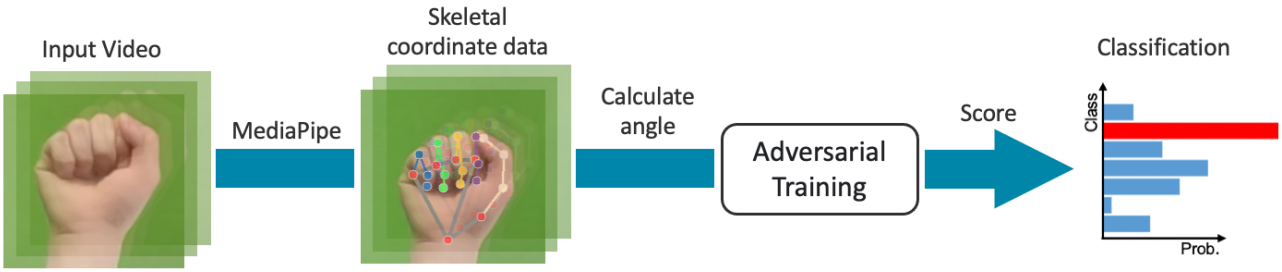

図 1　提案手法の流れ. 人物姿勢推定による得られた骨格データを用いて敵対的学習を行い，評価を行う.

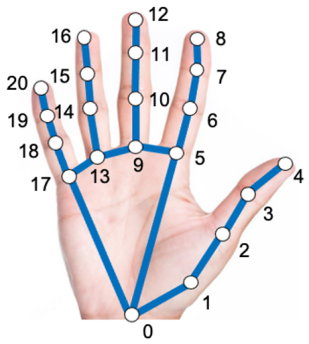

図 2　MediaPipe Hands のキーポイント

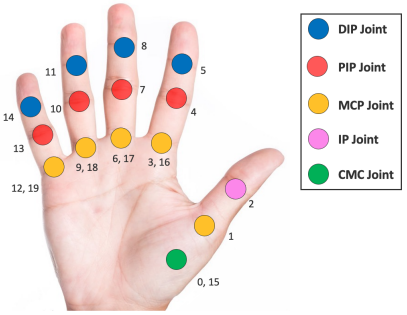

図 3　関節角度のキーポイント

を行う [19].

2.4　RNN を用いた骨格データに基づく行動認識

骨格データは，照明の変化や背景の変化に対してロバストであり，高精度の深度センサーや人物姿勢推定を用いることで容易に取得できる [7–9]. また，人間の関節の位置を 2 次元または 3 次元座標の形で時系列に表現することができるため，その動きのパターンを分析することで，人間の行動を高精度で認識することが可能になる. 近年，時系列データの時間的パターンを学習するという RNN の利点により，いくつかの RNN ベースの行動認識方法が提案された.Temporal Segment LSTM(TS-LSTM) では，短期，中期，および長期のネットワークで構成されたモデルを用いて，骨格データを別の座標系に変換し，静止姿勢の特徴を抽出した [10]. また，人間の関節の共起を観測するために 3 層の Bidirectional LSTM(Bi-LSTM) モデルを利用した手法が提案された [11].Spatio-Temporal LSTM(ST-LSTM) では関節を木構造として，トラバーサルすることで，関節の空間的な関係を構築した [12].

3　提 案 手 法

提案手法の流れを図 1 に示す. 本研究では，人物姿勢推定を用いて手話動画から生成した骨格データを利用して敵対的学習を行う. 以下では，MediaPipe による人物姿勢推定，関節角度データの作成，提案する敵対的学習手法ついて説明する.

3.1　MediaPipe による人物姿勢推定

MediaPipe [13] は，オープンソースでカスタマイズ可能な ML ソリューションである. 本研究では，トップダウン・アプローチに基づいた 3 次元の人物姿勢推定モデルである MediaPipe Hands [14] を使用する．この手法は，RGB 画像に映っている人物の手の 3 次元骨格座標を推定することができる．キーポイントは図 2 に示す．MediaPipe Hands から得られる推定結果は，21 キーポイントの 3 次元骨格座標データ，推論された手の種類のスコア (右手か左手か) の 2 種類である.

3.2　関節角度データの作成

本研究では，MediaPipe によって得られた 3 次元骨格座標データを角度データに変換して学習に用いる. 角度データへの変換には，以下の式を用いた.

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|a\|\|b\|} \tag{1}$$

関節角度のキーポイントは図 3 に示す. 0 から 1 4 の関節角度は垂直角度で構成される.15 から 19 の関節角度では，親指以外の MCP 関節と CMC 関節は，他の関節と違って左右に動かせるため関節の水平角度で構成される. したがって合計 20 個の関節角度が取得される.

3.3　敵対的学習

敵対的学習手法の流れを図 4 に示す. まず，生成器にノイズを入力し，角度ノイズを生成する. この角度ノイズを生の角度データに合成する. 次に，生データと敵対的データを識別器に入力し，それぞれのデータの損失と精度を出力する. 生成器 G は以下の式に従ってパラメータの更新を行う.

$$\max_G \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\mathbb{E}_{z\sim p_z(z)}\mathcal{L}[D(G(z)),y] - \mathcal{L}[D(x),y] \tag{2}$$

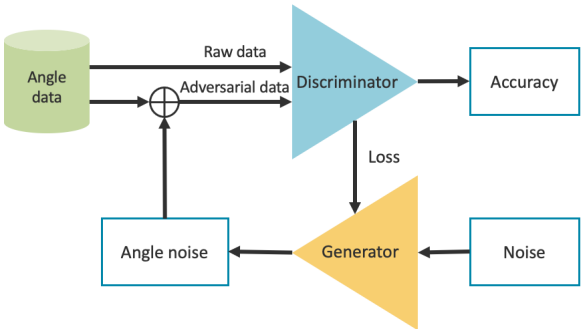

図 4　提案した敵対的学習の流れ

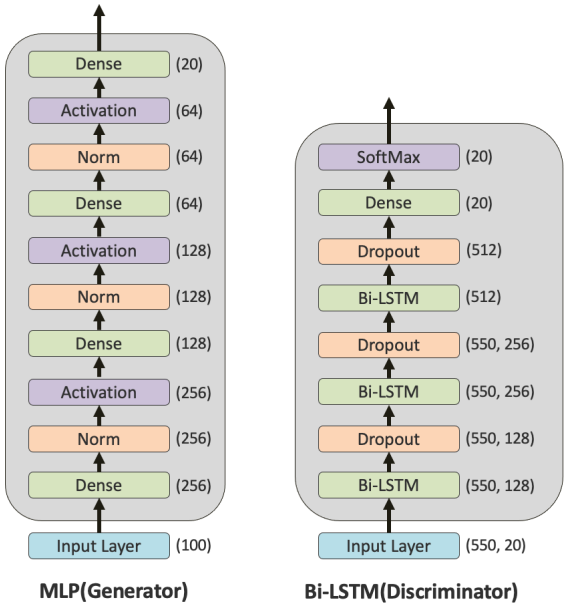

図 5　生成器，識別器のアーキテクチャ

G は期待値を最大化することで，生データ x と比較して D の損失を増加させる可能性のある敵対的データ $G(z)$ を出力する.z は乱数ベクトルである.$\mathcal{L}(-,-)$ は予め定義された損失関数であり，y は正解ラベルである.

識別器 D は以下の式に従ってパラメータの更新を行う.

$$\min_D \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\mathbb{E}_{z\sim p_z(z)}\mathcal{L}[D(G(z)),y]+\mathcal{L}[D(x),y] \quad (3)$$

D は期待値を最小化することで，より良い性能を得るために敵対データから学習する. 従って，敵対的データ $G(z)$ は生データ x よりも D の弱点を反映することができ，その結果，より効果的な学習が可能となる.

本研究では生成器に多層パーセプトロン (MLP)，識別器に Bi-LSTM を用いる. それぞれのアーキテクチャは図 5 に示す.MLP の活性化関数には LeakyReLu を用いる. そして最適化アルゴリズムには共に Adam を用いる.

4　実　　験

我々の提案するデータ拡張手法の有効性を評価するため，様々な角度から撮影された手話動画を用いてアブレーション研究によって評価を行った.

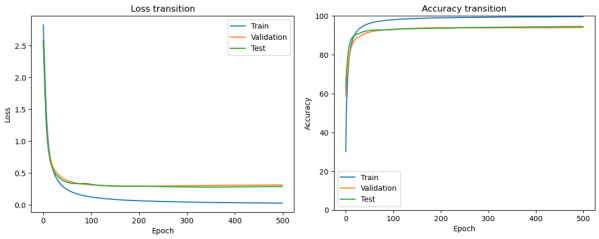

図 6　損失と精度の推移（敵対的学習）

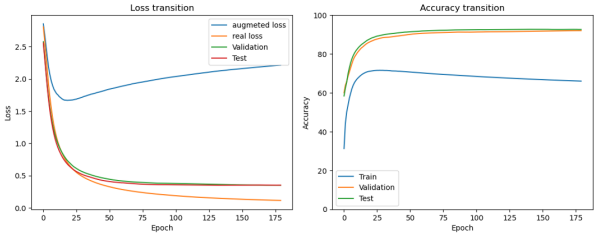

図 7　損失と精度の推移（敵対的学習あり）

4.1　データセット

使用する手話動画データでの被験者は 5 人の非母国語手話者であり，3 台のカメラを同時に使用して，同じ手話を 3 つの異なる角度から 5 回撮影している. データセットは無作為に選択した 20 種類の日本語指手話 1500 個の手話動画データ (MP4) から構成される. 本研究では，MediaPipe による人物姿勢推定における信頼スコアが 50%以下となったものは失敗フレームとして排除し，推定後に残ったフレームが全体の 20%以下となった動画は削除した. その結果，骨格データセットは 1494 個の手話動画データから取得された骨格データによって構成される. データセットは訓練用に 3 人，検証用に 1 人，テスト用に一人と分割する.

4.2　アブレーション研究

図 6 より，敵対的学習を行わない場合，validationd データの精度において Epoch300 から少し過学習の傾向がある. しかし，敵対的学習を行わないモデルと比較して敵対的学習を行ったモデルの精度は約 1.4%減少した. これは，7 より，敵対的データの損失が増加し続けているのに対し，訓練データ全体の精度が減少し続けていることから，生成器が識別器にとって難しすぎるデータを生成していまっていることを示している. また，図 8 より，生成器が生成したノイズの値は大きく，生データに大きな影響を与える可能性がある.

5　お わ り に

本研究では，骨格データに着目した敵対的学習手法を提案した. アブレーション研究によって本手法の性能評価を行った. その結果，提案手法を用いた場合，精度が減少することが示された. これは，生成器が識別器にとって難しすぎるデータを生成していることが原因だと考えられる. また，生成したノイズの値が大きかったことから，生データに大きな影響を与えた可能性がある.

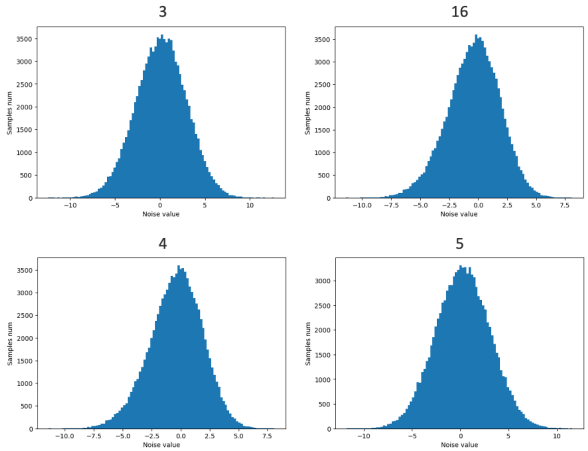

図 8　敵対的データの分布. 図は人差し指の場合に生成するノイズの分布である.

今後の課題として，生成するノイズの値の大きさに制限を加えることが挙げられる.

文　　献

[1] N. Mohamed, M. B. Mustafa and N. Jomhari, "A Review of the Hand Gesture Recognition System: Current Progress and Future Directions, " in IEEE Access, vol. 9, pp. 157422-157436, 2021.

[2] Mehta, D., "Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision", arXiv preprint arXiv:1611.09813, 2016.

[3] Park, S., Hwang, J., and Kwak, N., "3D Human Pose Estimation Using Convolutional Neural Networks with 2D Pose Information", arXiv preprint arXiv:1608.03075, 2016.

[4] Luo, C., Chu, X., and Yuille, A., "OriNet: A Fully Convolutional Network for 3D Human Pose Estimation", arXiv preprint arXiv:1811.04989, 2018.

[5] Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y., "Convolutional Pose Machines", arXiv preprint arXiv:1602.00134, 2016.

[6] He, K., Gkioxari, G., Dollár, P., and Girshick, R., "Mask R-CNN", arXiv preprint arXiv:1703.06870, 2017.

[7] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y., "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", arXiv preprint arXiv:1812.08008, 2018.

[8] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y., "OpenPose: Realtime Multi-Person 2D Pose Estimation us- ing Part Affinity Fields ", arXiv preprint arXiv:1812.08008,2018.

[9] Zhang, F., "MediaPipe Hands: On-device Real-time Hand Tracking", arXiv preprint arXiv:2006.10214, 2020.

[10] Ma, C.Y., Chen, M.H., Kira, Z., et al., "TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition", Signal Process., Image Commun., 2019.

[11] Zhu, W., Lan, C., Xing, J., et al., "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks", AAAI, 2016.

[12] Liu, J., Shahroudy, A., Xu, D., and Wang, G., "Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition", arXiv preprint arXiv:1607.07043, 2016.

[13] Lugaresi, C., "MediaPipe: A Framework for Building Perception Pipelines", arXiv preprint arXiv:1906.08172, 2019.

[14] Zhang, F., "MediaPipe Hands: On-device Real-time Hand Tracking", arXiv preprint arXiv:2006.10214, 2020.

[15] Kipf, T. N. and Welling, M., "Semi-Supervised Classification with Graph Convolutional Networks", arXiv preprint arXiv:1609.02907, 2016.

[16] Nagasima, Y., "Kogakuin University Japanese Sign Language Multi-Dimensional Database (KoSign)", Informatics Research Data Repository, National Institute of Informatics, 2021.

[17] Yu, A. and Grauman, K., "Semantic Jitter: Dense Supervision for Visual Comparisons via Synthetic Images", arXiv preprint arXiv:1612.06341, 2016.

[18] Wang, X., Shrivastava, A., and Gupta, A., "A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection", arXiv preprint arXiv:1704.03414, 2017.

[19] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation", In ICCV, 2017.