識別性と物理的な類似性を両立する日本手話構成要素の表現学習

井上純大　原大介　三輪誠
豊田工業大学大学院
{sd24410,daisuke,makoto-miwa}@toyota-ti.ac.jp

概要

本研究は，手話単語を構成する位置・動き・手型の構成要素について，カテゴリの識別性を維持しつつ，物理的な類似性も反映した埋め込み表現の獲得と，その有効性の検証を目的とする．この目的に向けて，PU-AUC 最適化を導入し，物理的に類似した構成要素を埋め込み空間上で近接させる手法を提案する．さらに，物理的な類似性の表現を評価するため，日本手話順序付きトリプレットデータセットを新たに構築する．構築したデータセットで評価した結果，提案手法は識別性を保ちつつ物理的な類似性をより正確に反映していることを確認した．また，下流タスクである適格性解析においても性能向上を示し，獲得した埋め込み表現の有効性を示した．

1　はじめに

手話は書き言葉を持たない自然言語であり，その基本単位である単語は手の位置・動き・手型の構成要素[1)]から構成される [1]．手話言語学では，単語を体系的に記述し分析するため，構成要素をカテゴリ[2)]として整理し，記号化してきた [3, 4]．

近年，手話動画コーパスの整備と深層学習の進展に伴い，カテゴリ体系をデータに基づいて大規模に運用する基盤として，動画から構成要素のカテゴリを自動識別する手法が提案されている [5, 6, 7]．これらの手法では，入力動画から特徴を抽出し，それを埋め込みとして表現した後，正解ラベルが付与されている各構成要素に対してカテゴリ識別を行う．

しかし，従来手法はカテゴリ識別のみを学習の目的とするため，カテゴリ識別に関係のない物理的な情報が捨象されてしまう（図 1 上部）．例えば，二つの動画で指の曲がり具合がわずかに異なっている場合でも，カテゴリが同じである場合は区別する必要はなく，異なる場合は大きく区別する必要がある．このような情報の捨象は，手話認識のような意味的な解析以外の手話処理（動作の似た例を探す類似例検索やろう者による手話の自然さの判断を扱う適格性解析など）に利用する上で問題となる．

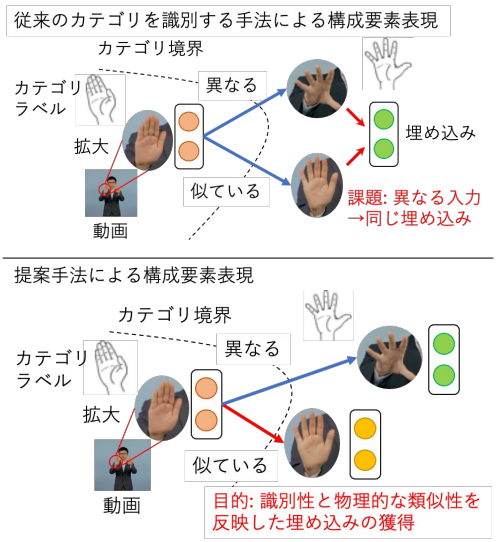


図 1: 手話単語の構成要素（手の位置・動き・手型）の表現における従来手法と提案手法の比較．（上）従来のカテゴリ識別では同じカテゴリ内の手型や位置の違いなどの物理的な情報が捨象される．（下）提案手法では物理的な類似性も反映した埋め込み表現を実現する．

そこで，本研究は日本手話を対象に，カテゴリ間の識別性を維持しつつ，構成要素間の物理的な類似性（どの事例同士がどれほど似ているか）を反映した埋め込み表現の獲得を目的とする．このために，構成要素のカテゴリラベルのみから学習でき，かつ別カテゴリでも物理的に似ている事例

1)　これらの構成要素は音声言語における音素に相当する．
2)　本研究で使用する日本手話音節データベース [2] の定義では，手型カテゴリは手の形状（握り拳の形，親指だけ伸ばした形など）を表す．位置カテゴリは手が手話空間または身体部位（目・口・身体など）に対する相対位置を表す．動きは軌跡運動（上・下・右・左方向の動きなど），手首の動作，手指動作に大別される．

を一律に遠ざけないための表現学習手法として，Positive-Unlabeled (PU) 学習の枠組みに基づく AUC 最適化 [8] を採用し，同一カテゴリの事例（正例）がそれ以外のランダムな事例（ラベルなし例）より埋め込み空間上で近くなるように学習する（図 1 下部）．PU 学習は正例（P）とそれ以外（U）の順序関係のみを考慮するため，U の中に含まれる類似の事例が P に近づくことを許容する余地が生まれる．さらに，物理的な類似性の表現を定量的に評価する新たな言語資源として日本手話順序付きトリプレットデータセットを構築する．既存の手話コーパスは主として意味的な情報に着目したカテゴリを付与しており，物理的な類似性に基づく評価ができない．そこで，各構成要素の物理的な順序関係を表す順序付きトリプレットに基づくデータセットを構築し，埋め込み空間上でこの順序関係をどの程度再現できるかにより，物理的な類似性の表現を評価する．

本研究の貢献を以下にまとめる．

- 物理的な類似性を埋め込み空間上で評価するための新たな言語資源として，日本手話順序付きトリプレットデータセットを構築した．
- 手話構成要素のカテゴリ識別性と物理的な類似性を両立する埋め込み表現を獲得するために，PU-AUC 最適化に基づく学習手法を提案し，構築したデータセットにおいて，従来手法と比べて，物理的な類似性を埋め込み表現により正確に反映できることを示した．
- 提案手法を下流タスクである適格性解析で評価し，専門家による記号記述との併用により，記号記述単体を上回る性能を達成し，記号では捉えにくい物理的な類似性が有用であることを明らかにした．

2 関連研究

手話は日本語や英語などと同様に自然言語であるが書き言葉が存在しない．この手話の構造を客観的に分析するため，手話の表記法が広く研究されてきた．Stokoe [1] は手話を言語学的に研究し，手話単語が手の位置・動き・手型の 3 つの構成要素からなると分析し，これらの構成要素に対して意味を弁別する単位で区切り，記号を割り当てた．その後，The Hamburg Sign Language Notation System (HamNoSys) [3] や SignWriting [4] など，構成要素をカテゴリとして整理し，単語を記述するための表記法が提案されてきた．

近年は手話動画コーパスの整備と深層学習の進展により，動画から構成要素のカテゴリを自動識別する研究が進められている [5, 6, 7]．また，手話単語の意味を認識する手話認識においても，構成要素のカテゴリの同時識別が有効であり，認識性能の向上につながることが示されている．

3 提案手法

本研究では，識別性を維持しつつ，従来捨象されていた物理的な類似性を埋め込み表現に反映する手法を提案する．具体的には，Positive-Unlabeled (PU) 学習の枠組みを用いて，AUC を最大化する損失を最適化する．この損失では，同一カテゴリの事例（正例）がそれ以外（ラベルなし例）より高いスコアを得るように学習することでカテゴリの識別性を確保すると同時に，別カテゴリの事例を負例として確定せず，物理的に似た事例が埋め込み空間上で近くに配置されることを許容する．提案手法の全体像を図 2 に示す．

3.1 アーキテクチャ

モデルは，Video Vision Transformer (Video ViT) [9] エンコーダと，全結合層（分類器）を組み合わせたものである．Video ViT エンコーダにより時空間特徴を抽出し，その後，全結合層により各カテゴリのロジットを算出し，カテゴリごとのロジットに基づいて PU-AUC 損失を計算する．

3.2 目的関数: PU-AUC 最適化

PU-AUC 最適化 [8] を One-vs-rest で多カテゴリに適用する．カテゴリ集合を $\mathcal{C}$ とし，各カテゴリ $c \in \mathcal{C}$ について，c の事例を正例（P），それ以外をラベルなし（U）として扱い，別カテゴリの事例を負例として一律に遠ざけるのではなく，物理的に似た事例が埋め込み空間上で近くに配置されることを許容する．入力動画 x に対し，分類器が出力する c におけるロジットをスコア $g_c(x)$ とする．ラベルなし分布はカテゴリの事前確率 π_c を用いて，$D_U = \pi_c D_c + (1 - \pi_c) D_{\neg c}$ で表される．本研究では π_c を学習データにおける c の出現比率で与える．各 c の PU リスクと PP リスクはロジスティック損失 $\ell(z) = \log(1 + \exp(-z))$ を用いて，式 (1)，式 (2) で表

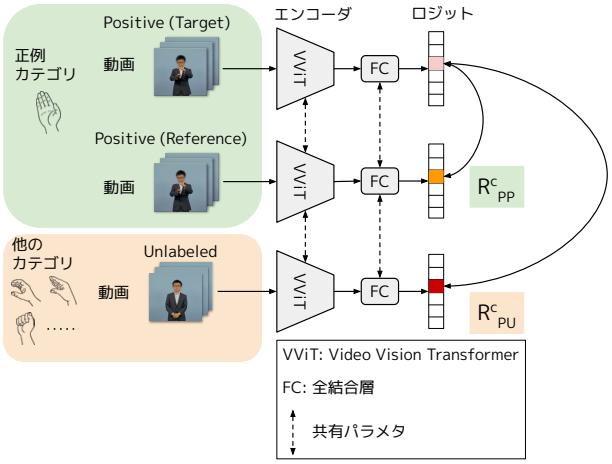


図 2: 提案手法の全体像．手話動画を Video ViT エンコーダに入力し，全結合層ヘッドを経てカテゴリごとのロジットを出力する．図は手型の処理を示しているが，位置および動きについても同様の処理を独立に適用する．

される．

$$\mathcal{R}^c_{\mathrm{PU}}(g) = \mathbb{E}_{x_P \sim D_c,\, x_U \sim D_U}[\ell(g_c(x_P) - g_c(x_U))], \quad (1)$$

$$\mathcal{R}^c_{\mathrm{PP}}(g) = \mathbb{E}_{x_P, x_{P'} \sim D_c}[\ell(g_c(x_P) - g_c(x_{P'}))] \quad (2)$$

また，AUC リスク（PN リスク）は式 (3) となり，これを全カテゴリ $\mathscr{C}$ において最小化する．

$$\mathcal{R}^c_{\mathrm{PUAUC}}(g) = \frac{\mathcal{R}^c_{\mathrm{PU}}(g) - \pi_c \mathcal{R}^c_{\mathrm{PP}}(g)}{1 - \pi_c} \quad (3)$$

4 日本手話順序付きトリプレットデータセット

学習された埋め込み表現が構成要素の物理的な類似性を表現できているかを定量的に評価する新たな評価用データセットとして，構成要素の物理的な順序関係に基づく日本手話順序付きトリプレットデータセットを構築する．本節では，データセットの構築手順とその統計情報について述べる．なお，本データセットの構築においてはアノテーションの難易度と品質を考慮し，利き手に関する構成要素のみを対象とし，非利き手の構成要素は対象としない．

4.1 サンプルセットの作成

日本手話音節データベース [10]（動画・構成要素カテゴリ）を用いてサンプルセットを多数作成する．各サンプルセットは，ある構成要素における 2 つの異なるカテゴリから 3 本ずつサンプリングした計 6 本の手話単語動画からなる．本研究では，手型の動的な変化を考慮し，動作開始時（変化前）と終了時（変化後）の手型を区別して扱う．

4.2 トリプレットの作成と判断基準

アノテータは，各サンプルセットについて，一つのカテゴリの事例 v_1 からもう一つのカテゴリの事例 v_3 への段階的な差異を示す順序付きトリプレット (v_1, v_2, v_3) を特定する．中間的な事例 v_2 が定められない場合は，そのセットを除外する．トリプレットの判断は，構成要素に応じて以下の基準に基づいて行う．判断基準の詳細と，アノテーションしたトリプレットの例を付録 A に示す．

手型　(i) 手の開閉度（開手 → 握り拳），(ii) 指の屈曲度，(iii) 指の開き具合（開 → 閉）．

位置　(i) 奥行き（身体へ近づく／遠ざかる），(ii) 水平・垂直位置．

動き　(i) 方向・軌道の連続性（運動方向が段階的に遷移するか）．

4.3 統計と一致率

157 トリプレットを特定し，これをデータセットとした．その構成要素ごとの統計を付録 B に示した．アノテーションは修士学生 2 名が独立に行い，順序が一致した割合は 84.21%，スピアマンの順位相関係数は 0.88 であった．

5 実験

実験では，物理的な類似性の再現性，構成要素カテゴリの識別性，下流タスク（適格性解析）への有効性の 3 つの観点から，埋め込み表現を評価した．

5.1 実験設定

構成要素の学習と識別性の評価には，日本手話音節データベース [10] の動画 1,078 件（訓練/開発/テスト: 755/161/162 件）を用いた．比較手法は，同一アーキテクチャを交差エントロピーを用いて学習した識別モデル（CE）とした．主要なハイパーパラメタは付録 C に示した．

5.2 物理的な類似性の評価方法

順序付きトリプレットに対し，各動画の埋め込み表現を e_1, e_2, e_3 とする．両端 e_1, e_3 が張る軸 $V = e_3 - e_1$ に対する中間点の射影係数 $\mathrm{Score} = \frac{(e_2 - e_1)\cdot V}{\|V\|^2}$ を計算し，$0 < \mathrm{Score} < 1$ を満たすとき，e_2 が e_1 と e_3 の間に配置されているとみな

表 1: 開発データにおける順序保存率 [%]

	順序保存率	
構成要素	CE	PU-AUC
手型（変化前）	84.00	**94.00**
手型（変化後）	84.00	**94.00**
位置	76.74	**81.40**
動き	84.00	**96.00**

表 2: 開発データにおける識別正解率 [%]

構成要素	手法	利き手	非利き手
手型 (変化前)	CE	46.58	66.46
	PU-AUC	**47.83**	**71.43**
手型 (変化後)	CE	47.83	65.22
	PU-AUC	**50.93**	**70.81**
位置	CE	78.88	87.58
	PU-AUC	**85.09**	**88.82**
動き	CE	**62.73**	83.85
	PU-AUC	59.01	**90.68**

し，順序関係を再現できた（正解）と判定し，全トリプレットに対する正解率である順序保存率を計算した．

5.3 物理的な類似性と識別性の評価

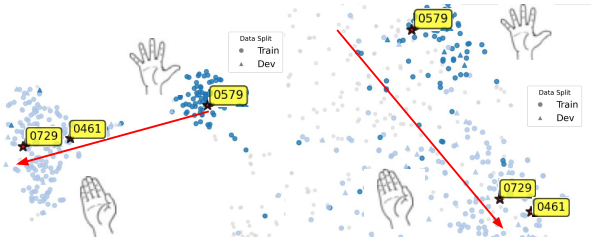
表 1 の結果より，提案手法は全構成要素で順序保存率を改善し，物理的な類似性を反映した埋め込み表現を獲得できた．また，表 2 を見ると，カテゴリ識別の正解率も多くの構成要素で同等以上であり，識別性は損なわれないことが確認された．さらに t-SNE [11] 可視化（図 3）では，中間的な事例がカテゴリ間に配置される連続的な配置が確認できた．

5.4 適格性解析への有効性検証

提案手法で獲得した埋め込み表現が下流タスクに有用であるかを検証するため，ろう者の言語直観に照らして手話として自然に成立するか否かを判定する適格性解析 [10] の 2 値分類を行った．PU-AUC および CE で学習した動画エンコーダのパラメタは凍結し，各動画から得た埋め込み表現（以下，動画埋め込み）を特徴量として用いる．また，専門家による記号記述は，位置・動き・手型に加えて掌の向きなどの詳細な識別情報を含む一方，動画由来の埋め込み表現は物理的な類似性を捉えることが期待される．この両者を組み合わせることで，記号的な識別情報と物理的な類似性の情報を同時に利用でき，適格性解析の性能向上が見込まれる．そこで，記号記述（460 次元）のみ，動画埋め込みのみ，および両者の連結（記号記述+動画埋め込み）を入力として，ロジスティック回帰モデルを学習し評価した．

表 3: 適格性解析の正解率 [%].

	正解率	
モデル	開発	テスト
記号記述	74.53	70.37
PU 埋め込み	71.43	70.37
CE 埋め込み	68.94	59.26
記号記述 + CE 埋め込み	69.57	63.58
記号記述 + PU 埋め込み	**76.40**	**70.99**



(a) 従来手法　　(b) 提案手法

図 3: 手型に関する埋め込み空間の t-SNE 可視化．「0579」から中間形（「0729」）を経て「0461」へと連続的に遷移する例を示す．(a) 従来の分類学習はカテゴリ分離はできるが連続的な変化を保持しない．(b) 提案手法は中間形を 2 つの手型の間に配置し，連続的な変化も捉えられている．

結果を表 3 に示す．動画埋め込み単体では PU-AUC が CE を大きく上回り，より有用な表現を与えることが分かる．さらに記号記述と併用すると，PU-AUC 埋め込みとの組み合わせが最良となり，この結果は，適格性解析において，記号的な弁別情報に加え，物理的類似性の情報も重要な役割を果たしていることを示唆する．

6 おわりに

本研究は，手話構成要素のカテゴリ識別性を維持しながら，構成要素間の物理的な類似性を反映した埋め込み表現の獲得を目的に，PU 学習に基づく AUC 最適化に基づく埋め込み表現学習手法を提案した．さらに，物理的な類似性を定量評価するため，日本手話順序付きトリプレットデータセットを構築した．実験では，提案手法は従来の分類学習より順序保存率を一貫して改善し，識別性も大きく損なわないことがわかった．また，適格性解析でも記号記述との併用で性能が向上し，ろう者の判断に物理的な類似性が寄与する可能性を示唆した．今後は，トリプレット損失や対照学習などの PU-AUC 損失以外の損失関数の評価を行うとともに，適格性解析における説明可能性の解析を進める予定である．

謝辞

本研究は JSPS 科研費 JP23H00626 の助成を受けたものです．産総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」の支援を受けて利用した．

参考文献

[1] William C. Stokoe. **Sign Language Structure: An Outline of the Visual Communication System of the American Deaf**, Vol. 8 of **Studies in Linguistics, Occasional Papers**. University of Buffalo, Buffalo, NY, 1960.

[2] 原大介. 新日本手話コーディングマニュアル. 2019.

[3] Thomas Hanke. Hamnosys–representing sign language data in language resources and language processing contexts. In **sign-lang@ LREC 2004**, pp. 1–6. European Language Resources Association (ELRA), 2004.

[4] Valerie Sutton. Lessons in sign writing. sign-writing., 1990.

[5] Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata, and Angelo Cangelosi. WLASL-LEX: a dataset for recognising phonological properties in American Sign Language. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 453–463, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[6] Lee Kezar, Jesse Thomason, and Zed Sehyr. Improving sign recognition with phonology. In Andreas Vlachos and Isabelle Augenstein, editors, **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 2732–2737, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[7] Jundai Inoue, Makoto Miwa, Yutaka Sasaki, and Daisuke Hara. Enhancing syllabic component classification in Japanese Sign Language by pre-training on non-Japanese Sign Language data. Torino, Italia, May 2024. ELRA and ICCL.

[8] Tomoya Sakai, Gang Niu, and Masashi Sugiyama. Semi-supervised auc optimization based on positive-unlabeled learning. **Machine Learning**, Vol. 107, No. 4, pp. 767–794, 2018.

[9] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 14549–14560, June 2023.

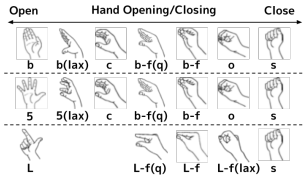
[10] Satoshi Yawata, Makoto Miwa, Yutaka Sasaki, and Daisuke Hara. Analyzing well-formedness of syllables in Japanese Sign Language. In Greg Kondrak and Taro Watanabe, editors, **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 26–30, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[11] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. **Journal of machine learning research**, Vol. 9, No. Nov, pp. 2579–2605, 2008.

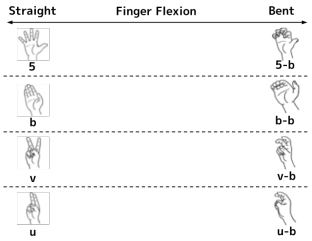
[12] Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled, 2024.

A　日本手話順序付きトリプレットデータセットの構築詳細

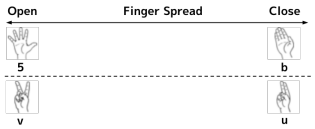
本節では，本データセットを構築する際に用いたアノテーション判断基準の例を図 4 から 6 に，順序付きトリプレットの具体例を図 7 に示す．



(a) 開閉度



(b) 屈曲度



(c) 開き具合

図 4: 手型の判断基準．

B　日本手話順序付きトリプレットデータセットの統計

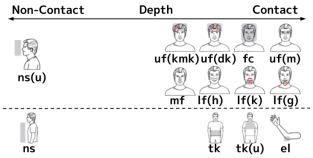
表 4 に構築した 157 トリプレットの内訳を示す．

表 4: 日本手話順序付きトリプレットデータセットの内訳．

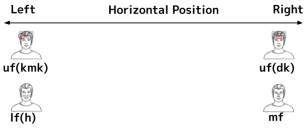
構成要素	数
手型（変化前）	50
手型（変化後）	39
位置	43
動き	25
合計	157

C　実装詳細（ハイパーパラメタ）

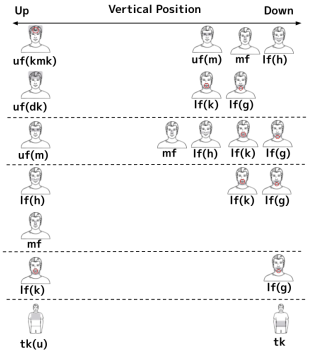
主要なハイパーパラメタを表 5 に示す．



(a) 奥行き

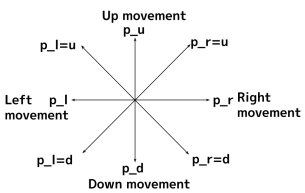


(b) 水平

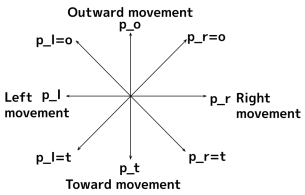


(c) 垂直

図 5: 位置の判断基準．



(a) 動きの方向における連続的な変化（視点 1）．



(b) 動きの方向における連続的な変化（視点 2）．

図 6: 動きの判断基準．

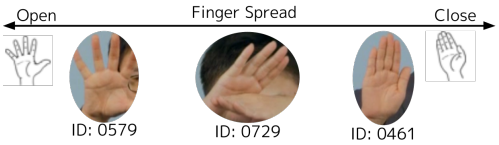


図 7: 順序付きトリプレットの具体例

表 5: 主要ハイパーパラメタ

エンコーダ	Video ViT-B/16-224
埋め込み次元	128
入力解像度	224×224
フレーム数	16
最適化	SF-AdamW [12]
学習率	1.0×10^{-4}
バッチサイズ	2（勾配蓄積 16）
エポック数	300