# 人認識AIを活用した手指標識を中心とした手話言語データの蓄積

田中 省作[1,4,a)]　本田 久平[2]　長谷川 由美[3]　バイティガ ザカリ[4]

**概要**：言語研究において言語データの蓄積は非常に重要である．視覚言語である手話でもデータの蓄積が試みられているものの，手話の媒体は視覚情報で記述や記録が統一的ではないなど，容易ではない．一方で，近年の人認識 AI 技術の進歩に，手話情報処理への展開も期待される．本研究は，手話映像から，とくに手指標識を中心とした手話言語データの蓄積，動作が似た手話表現の自動検出などのパイロット研究に言及しつつ，人認識 AI の手話研究への有用性を示す．

**キーワード**：手話, 言語データ, 人認識 AI, 手指標識

# Processable Digitization of Manual Markers in Sign Language Video Data Using Personal Recognition AI

***Abstract:*** Digitizing linguistic data is essential for language research. However, collecting data for sign languages, which are inherently visual, is challenging owing to the lack of standardized methods for describing and recording visual information. Recent advancements in personal recognition AI have expanded new possibilities for sign language processing. This study explores the potential of such AI technologies for use in sign language research, with a focus on pilot studies that involve digitizing data related to manual markers in sign language videos and automatically detecting sign expressions similar to manual markers' motion .

***Keywords:*** Sign language, linguistic data, personal recognition AI, manual marker

## 1. はじめに

本研究の対象である手話は，たしかに言語であり[*1]聴覚障害者にとっては，きわめて重要な言語である．そんな手話は，手指標識とよばれる腕や手型にかかわる言語要素と，非手指要素とよばれる表情や口型，姿勢など，それ以外の言語要素を駆使し，意思疎通する．本研究では，我が国の聴覚障害者にとって，もっとも一般的な日本手話を対象とする．以降，単に手話と記した場合は日本手話を指す．

さて，言語の用例等を蓄積し，適切に検索，参照できる言語データのデータベースは，言語研究に不可欠な研究基盤のひとつである．日本語や英語といった音声言語では，コーパス言語学と総称されるような概念も確立され，その重要性はますます増している．言語である手話でも，質の高い手話の会話データなどの蓄積が進められている [5], [6]．ただ，人手によるデータベース化には，一般に大変な労力を要する．

そこで本研究は, 近年, 急速に進歩した人認識 AI(Artificial Intelligence) を活用した，手指標識を中心とした手話映像の言語データ化を示す．手話映像に人認識 AI を適用し，手指標識に関する情報を取得，事前に付与しておく．それらは手話を検索したり，比較するような枠組みへの足がかり

[1] 立命館大学
Ritsumeikan University, 56-1 Toji'in Kitamachi, Kitaku, Kyoto 603-8577, Japan
[2] 大分工業高等専門学校
National Institute of Technology, Oita College, 1666 Maki, Oita 870-0152, Japan
[3] 近畿大学
Kindai University, 930 Nishimitani, Kinokawa, Wakayama 649-6493, Japan
[4] 沖縄工業高等専門学校
National Institute of Technology, Okinawa College, 905 Henoko, Nago, Okinawa 905-2192, Japan
[a)] sho@ritsumei.ac.jp

[*1] 手話言語ともよばれる．手話は，音声言語と同様に，国や地域によって異なっており，方言もある．手話固有の文法体系もあり，内的に外的にも言語としての特性を有している．

となる．本研究は，このような手話言語データの蓄積のための形式や具体的な手続きと，パイロット研究での活用事例を紹介しつつ，人認識 AI の手話研究への有用性を示す．

## 2. 手話言語データと人認識 AI

### 2.1 視覚情報の言語データ化

主要な音声言語には書記体系があり[*2]，発話などの言語産出物が文字などの，なにか処理できる単位に対応付けられていることは，データベース化の要件のひとつである．一方，手話は腕，手型，表情などの視覚情報で意味ある表現を構成している．手話映像を蓄積するだけでも有意義ではあるものの，言語データとしての蓄積，そしてデータベース化という視座から見直すと，効率的な検索などは難しく，不十分であることがわかる[*3]．適切な単位への対応付けが求められる．

そこで，手話言語データの蓄積に関して，いくつかの方式や成果が提示されている [4]．たとえば，手話表現を文字列で表す記法を活用し [7], [8]，手話映像に現れる手話表現に併記したり，手話表現に相応する日本語などの音声言語で注記したりする．前者の手話表現の文字列としての表記法は，一般の手話話者や学習者へ十分に浸透しておらず，また後者の音声言語での訳出は負荷も高い．そのほかにも，[5], [6] では，さまざまな分野での活用を念頭に置き，緻密にデザインされ，物理的にも詳細な，非常に質の高い手話言語データも開発されている．これらのデータの大規模化に際しては，専門家や機器や要する時間など，高コストであることが難点である．

### 2.2 人認識 AI

深層学習などの機械学習や，インターネット上の超大規模データや計算資源の高度化で，近年，AI 技術は大きく進展している．画像や映像のなかの物体や人を自動認識する AI 技術の進歩も著しい．ここでは，本研究で実際に利用している人認識 AI を概括する[*4]．

**YOLO**

YOLO は，物体認識のためのライブラリである．人だけでなく，さまざまな物体を認識できる．個々人やチームによっていくつかのバージョンや改良版がリリースされている．安定的に参照できる YOLO のひとつは，Ultralytics 社によるもので，11 が最新のバージョンである [12]．

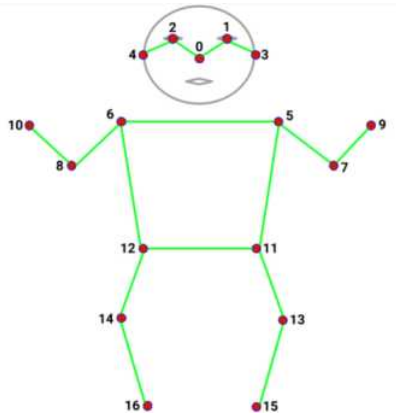


**図 1** YOLO が認識する人の部位 (キーポイント)
**Fig. 1** Human body parts(keypoints) recognized by YOLO

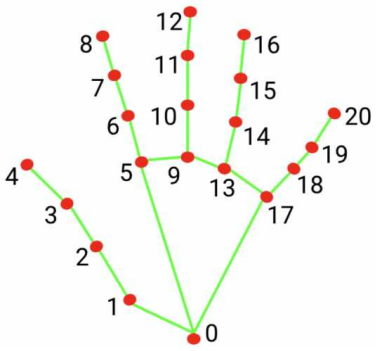


**図 2** MediaPipe(Hand Landmaker タスク) が認識する手の部位 (キーポイント)
**Fig. 2** Hand parts(keypoints) recognized by MediaPipe(Hand Landmarker task)

YOLO による人認識で，とくに本研究にかかわる機能は次のとおりである．

- 映像のフレームごとに，人を認識する．そして，認識した人ごとに映っている矩形領域の情報 (左上・右下の座標) を得る
- 認識した人ごとに，肩や肘，手首など主要部位 17 点の位置情報を得る (**図 1**)
- 認識した人ごとに，人追跡アルゴリズムに基づき，当該時点までのフレームに現れている，同一と推定される人を，ID を通してひもづける

**MediaPipe**

Google による画像からテキストまで，さまざまなメディアを自動処理するためのライブラリである．YOLO 同様，映像内の物体認識もできる．YOLO に比べ，局所的で詳細な情報が得られる．とくに本研究に関わる機能は，次のとおりである．

- 人の主要部位の認識は，より細かく，33 点の位置情報を得る
- 位置情報はフレーム内の縦・横に関する情報だけではなく，奥行も得る．つまり，3 次元座標として与える

---

[*2] 固有の文字や書記体系をもたない音声言語であっても，これまでの長い言語学研究の下で，他の書記体系を転用したり，独自に拡張することで，代替する体系が与えられることが多い．
[*3] 手話映像の単なる記録は，音声言語でいえば，各自の自筆や発話音声記録に相応する．たとえば，書かれた文字でいえば「字形」である．同じ文字，たとえば「あ」でも書き手によって，それぞれ差異がある．私たちは「あ」の文字としての枠組み (字体) に基づいて日本語の「あ」と理解する．データベースでは，少なくとも字体のレベルに引き上げる必要がある．ひるがえって，手話映像の単なる記録は，さきの例でいえば字形の記録にはなっても，字体での記録にはなっていない．
[*4] バージョンや URL は，2025 年 1 月 25 日確認のものである．

- 人の手に専念した処理では，指関節など 21 点の位置情報 (3 次元) を得る (**図 2**)．手の左右に関する情報も推定する
- 人の顔に専念した処理では，目・鼻・口の主要部位にくわえ，顔メッシュのための 478 点の位置情報 (3 次元) も得る[*5]

物体認識の下位処理として人認識も可能ではあるものの，YOLO のように映像内での人のひもづけまでは行わない．奥行は魅力的な情報だが，残念ながら精度は高くはない．手の左右の推定精度もあまり高くはないようである．

**YOLO と MediaPipe の併用**

このように，人の認識だけを取り上げてみても，YOLO, MediaPipe それぞれがもつ機能や，焦点を当てる範囲，精細度，推定精度なども異なる．YOLO は人の認識に使いやすく，MediaPipe は手話の手指標識には欠かすことができない，詳密な情報が得られる．そこで，まず YOLO で人の認識を行い，MediaPipe で手に対する詳細な認識を行うこととした．

## 3. 手話映像のデータ化

### 3.1 手話映像に対する手続きと得られる情報

前節で示した人認識 AI を活用すると，次のような手続きで，映像に映っている手話話者の手指標識に関する情報を取得できる．

**【A】手続き**

(1) 映像 (フレーム列) のフレームごとに, 次の処理を行う．
  (a) YOLO のポーズ推定 (Pose Estimation) タスクを適用し，フレーム内の人の矩形領域に関する情報 (相対座標と画像)，人 ID，肩・肘・手首に関する位置情報を得る
  (b) (a) で切り出した画像を，人 ID ごとに，フレーム番号に順じて結合することで，人 ID の映像を得る
(2) (1) で得られた人 ID ごとの映像に対して，次の処理を行う．
  (a) MediaPipe の手のランドマーク検出 (Hand Landmark Detection) タスクを適用し，手首・5 指の指先・関節の位置情報を得る
  (b) YOLO が認識した左右の手首の座標と, MediaPipe が認識した手の手首の座標を比較し，MediaPipe の手の左右を決定する[*6]
(3) (1),(2) で得られた情報をフレーム番号・人 ID ごとに記録する

[*5] 本研究は，非手指標識は扱わないが，このような顔に関する詳細な情報があれば，非手指標識の表情などもいずれ取り入れていくことができるだろう．

[*6] MediaPipe も手の左右を推定はするものの，手のランドマーク検出タスクにおいては，あまり高くないように見受けられた．そのため，YOLO の手首の左右の情報を継承している．

映像内には映っているものの，(1) の YOLO で認識されなかった話者は，そもそも対象から外れてしまう．さらに，(2) の MediaPipe の手の認識でも，もれは起こる．このように 2 段階で認識もれが起こる可能性には留意しておく必要がある．

このような手続きを経た結果，映像のフレームと人 ID ごとに得られる情報は，次のようなものである．

**【B】得られる情報**

(1) フレーム番号
(2) 人 ID
(3) 人の矩形情報 (左上の相対座標，右下の相対座標)
(4) 肩の相対座標 (左・右)
(5) 肘の相対座標 (左・右)
(6) 手首の相対座標 A(左・右)
(7) 手首の相対座標 B(左・右)
(8) 親指の指先・3 関節の相対座標 (左・右)
(9) 人差指の指先・3 関節の相対座標 (左・右)
(10) 中指の指先・3 関節の相対座標 (左・右)
(11) 薬指の指先・3 関節の相対座標 (左・右)
(12) 小指の指先・3 関節の相対座標 (左・右)

(2)-(6) は YOLO による認識結果で，相対座標は 2 次元である．(7)-(12) は MediaPipe による認識結果で，相対座標は奥行まで加わった 3 次元である．元の映像の基本情報 (フレーム幅・高，フレーム長，フレームレート) もメタ情報として，別途，記録している．よって，ひとつの【B】は物理的にも，元映像の特定時間の，特定の人，その人の手指標識の場所と対応づいている．この【B】の束が，本研究の手話言語データで，核となる．

### 3.2 手指標識のモデル化

前項の，手話映像に対する情報に基づいた手指標識のモデル化の一例を示す．これらは [2], [3], [9] で実際に活用したもので，日本手話の指文字[*7]をこのモデルで特徴づけた，SVM による自動認識では，9 割を超える精度となっている [3].

特定の人の手話映像 $X$ のフレーム列を $\boldsymbol{X}$ とする．$\boldsymbol{X}$ が $m$ つのフレームから成るとき，$\boldsymbol{X} = [x_1, x_2, \cdots, x_m]$ と表すこととする．このような映像であるフレーム列を構成する 1 フレーム $x \in \boldsymbol{X}$ のなかの手指標識を，次のように特徴づける．なお，L,R は手の左・右を表す．

(1) 腕の位置
  肩を原点とした肘・手首の 2 つの 2 次元位置ベクトルとして与える．

$$v_x^{(1,h)}, v_x^{(2,h)} \ (h \in \{\mathrm{L}, \mathrm{R}\}) \qquad (1)$$

(2) 指の位置

[*7] 日本語の五十音に対応する，日本手話の表現.

手首を原点とした 5 指それぞれの指先の 5 つの 2 次元位置ベクトルとして与える．

$$v_x^{(3,h)}, v_x^{(4,h)}, \cdots, v_x^{(7,h)} \ (h \in \{\mathrm{L}, \mathrm{R}\}) \tag{2}$$

(3) 掌の向き
人差指の根本 (MP) を原点とした小指の根本への 2 次元位置ベクトルとして与える．

$$v_x^{(8,h)} \ (h \in \{\mathrm{L}, \mathrm{R}\}) \tag{3}$$

フレーム $x$ は (1)-(3) の 8 特徴，左右で延べ 16 特徴のベクトル組 $\boldsymbol{v}_x$ で表現される．簡単化のため $i$ 組目のベクトルを $v_{x,i}$ で記述する．

$$\begin{aligned}\boldsymbol{v}_x &= \langle v_x^{(1,\mathrm{L})}, v_x^{(2,\mathrm{L})}, \cdots, v_x^{(8,\mathrm{L})}, v_x^{(1,\mathrm{R})}, v_x^{(2,\mathrm{R})}, \cdots, v_x^{(8,\mathrm{R})} \rangle \\ &= \langle v_{x,1}, v_{x,2}, \cdots, v_{x,16} \rangle \end{aligned} \tag{4}$$

フレーム内にもともと部位が映っていなかったり，認識もれしていたりして，$\boldsymbol{v}_x$ には値が不定となる $v_{x,i}$ が含まれることがある．

そして，フレーム列 $\boldsymbol{X}$ のベクトル表現は，次のような各フレームのベクトル組の列である．

$$\boldsymbol{v}_{\boldsymbol{X}} = [\boldsymbol{v}_{x_1}, \boldsymbol{v}_{x_2}, \cdots, \boldsymbol{v}_{x_m}] \tag{5}$$

$\boldsymbol{v}_{\boldsymbol{X}}$ を適宜，手指標識のベクトル組列とよぶ．この手指標識のベクトル組列を活用した応用研究を 4.3 節で示す．

## 4. パイロット研究

### 4.1 手話映像のデータ化プログラム

MP4 や MOV などの手話映像の動画ファイルが与えられたとき，3.1 の【A】の手続きを経て，【B】のデータを得るプログラムの公開の準備を進めている．認識自体は YOLO, MediaPipe に依るので，情報学関係者のようなプログラミングに慣れた人であれば簡単に実装できる，技術的な独自性は低いプログラムである．ただ，手話話者，手話研究者など手話の関係者は必ずしもプログラミングに通じているわけではない．このような素朴なプログラムでも公開していくことには意味があるだろう[*8]．

そして，これらのプログラムを活用し，現在，次のような手指標識を中心とした手話言語データの蓄積を進めている．

- 手話の授業における学習者の手話会話映像
- 高齢聴覚障害者らの手話会話映像
- 手話教材やインターネット上の手話会話映像

### 4.2 手話言語データ作成における示唆

映像内の人の認識を YOLO が，同定された人の手の認識を MediaPipe が担う．YOLO の人認識に影響するのは，個々の話者が映っている角度や大きさ以外に服の色やその場にあるもののほか，訓練モデルと映像の解像度などが考えられる．体系的で一般性のある評価とはまったくならないものの，手話言語データを蓄積していく過程で得られた示唆を，次の 3 つの映像の一部を例に説明する．

(1) 手話授業における多人数話者の映像 (**図 3**)
約 90 分の手話授業での映像で，幅 $w$ と高 $h$ の解像度を $(w, h)$ と表すと，もともとオリジナルの映像は $(1920, 1080)$ で，受講者 (手話話者) の斜めから映したものである．会話の際はペアやグループを作るため，受講生はさまざまな方向を向く．カメラに対する角度の分散が高いだけでなく，人も手指標識も物理的に隠れてしまうことがある．

(2) 少人数の手話会話の映像 (**図 4**)
5 名の高齢の聴覚障害者らがカメラへ緩やかな扇形で対して，自由に手話で会話をしている，$(1920, 1080)$ の高解像度の映像である．ときおり手をポケットに入れたり，後ろに回したりして，手が隠れることがある．現場には話者 5 名のほかに，スタッフやファシリテータの手話通訳者が 5 名おり，映像内に断続的に入り込み，人の交差もよく起こる．

(3) 手話テキストの正準的な映像 (**図 5,6**)
手話学習のテキストにも使用される [13], [14] の，辞書パートの約 110 分の映像で，解像度は $(720, 480)$ である．基本的な手話表現を 2 名の講師が表出している．テキスト用であることから，話者以外の情報もなく，真正面から，話者，手指標識が十分に確認できるような映像である．タイトルが事項ごとに入るため，話者は不連続に現れる．

まず，人の認識についてみていく．なお，(1)-(3) に適用した YOLO の訓練モデルはいずれも共通で，最小の `YOLO11n` である．(1) の**図 3** は，上下段ともにある 1 フレームの認識結果で，枠がついている部分が人を認識している箇所である．上段は，元は高解像度だった映像を YOLO が $(640, 448)$ の低解像度にして認識したもので[*9]，6 名しか認識できていない．下段はオリジナルのままで 13 名認識される．オリジナル (下段) でも認識されない 3 名は，他の人にかなりの割合で重なり，隠れており，仮に認識できたとしても手指標識はほとんどみることができないものである．**図 4,5,6** の (2),(3) は，角度が良好で，人の認識もれはほぼない．このようなことから，解像度に加え，角度が重要である．

次に，人のひもづけを，(2),(3) を例に説明する．ここに挙げた映像以外でも，人が映像内に出て続けている間は，きちんとひもづけがなされる．しかし，人が不連続に映し

[*8] 自然言語解析の入口である形態素解析ツールが長らく，コーパスを活用した英語学研究の黎明期を支えた．

[*9] 当然，解像度を落とすことで，YOLO の処理時間は速くなる．

出される場合は，不安定になるようである．(2) では話者のほか，スタッフなど 5 名が行きかい，人の交差や話者が映像から消えてしまうと，ひもづけがうまくいかず，断続的になる．現場には 10 名しかいないにもかかわらず，90 分の映像を通して，延べ 1,315 名分もの人 ID が振られてしまう．YOLO のひもづけをそのまま活用することは難しく，人 ID に対する後処理が求められることがわかる．(3) では話者が 2 名で，人の交差はない映像である．映像では頻繁にタイトルが挟まり，数秒，話者が画面から消え，不連続に映し出され，110 分を通して，人 ID は 33 名分となる．人のひもづけは，手話映像に特化した問題ではないので，関連する研究や事例を精査する必要がある．ただ，不連続に映し出される場合でも，人の交差によるようなとき，ひもづけは切れやすいようである．

最後に，手指標識の認識について，人の認識もれがほぼない (2),(3) を例に考えてみる．そもそも MediaPipe の手の認識は人の認識に比べると，もれが多い．そのうえ，握りこぶしのように手の部位が重なっている，関節が映りにくい，一部が隠れているような場合には著しく認識が難しくなる．(2) で人認識がされたフレームは延べ約 74 万で，そのうち手が認識され【B】のデータが完備されるのは 68.1%の約 50.3 万フレームである．一方，(3) は手が隠れるような場面はほとんどなく，(2) に比べ良好な条件で映り続けているにもかかわらず，人認識がされたフレームの約 7.2 万のうち，手まで認識され，【B】のデータが完備されるのは 30.7%の約 2.2 万フレームである．(3) のこの低さは，手が隠れたりするのではなく，MediaPipe による認識もれが主因で，手の認識は，解像度の影響が大きいことがわかる．

今後の YOLO や MediaPipe の精度向上に期待すべきところもあるものの，このようなことを鑑みると，本方式による手話映像に対する手指標識に関する情報は，不定値を含む不完全なデータであることを前提としてしなければならない．次項では，このようなデータであっても，十分に利活用の可能性があることを，具体的な事例で示す．

### 4.3 動作類語の検出

音声言語で綴りが似ていても意味が全く異なるような語があるように，手話にも腕や手指の動作がよく似ているものの，表す内容は異なるような語 (表現) がある (**図 5**)．本研究では，このような語を，ある語に対する動作類語とよぶ[*10]．このような動作類語に関する網羅的な情報は，手話学習などで非常に重要である [1]．ここで示すのは，本研究のデータや 3.2 で示した手指標識のモデル化を，このような動作類語の検出に活用した事例である．

動作類語の検出のためには，手話表現から変換されたベクトル組列に対する距離を与えればよい．本研究では，手話表現間の距離の算出に，時系列データ間の比較の際に時間軸を柔軟に伸縮し，対応づけることができるダイナミックタイムワーピング (Dynamic Time Warping; DTW) を採用する [11]．

それぞれ長さ $m,n$ の列 $\boldsymbol{X} = [x_1, x_2, \cdots, x_m]$, $\boldsymbol{Y} = [y_1, y_2, \cdots, y_n]$ に対する DTW 距離 $d_{\mathrm{DTW}}(\boldsymbol{X}, \boldsymbol{Y})$ は，次のように与えられる．

$$\begin{aligned} &d_{\mathrm{DTW}}(\boldsymbol{X}, \boldsymbol{Y}) = f(m, n) \\ &f(i,j) = \delta(x_i, y_j) + \min \begin{cases} f(i, j-1) \\ f(i-1, j) \\ f(i-1, j-1) \end{cases} \end{aligned} \tag{6}$$

ただし，$f(0,0) = 0$, $f(i,0) = f(0,j) = \infty$ $(i = 1, 2, \cdots, m; j = 1, 2, \cdots, n)$ である．ここで，$\delta(x_i, y_j)$ は $x_i, y_j$ 間の距離で，課題に応じて設定すればよい．本研究では，手話表現のフレーム同士の距離にあたる．

手話表現ベクトル組列には，不定値を含むベクトル組があり，その扱いが問題となる．この動作類語の検出では，予備実験の結果から，不定値を含むベクトル組は除いた，すなわち認識もれがないフレームだけを対象とし，DTW 距離を与えることとした．

手話表現 $X, Y$ 間の DTW 距離を計算する際，まず，フレーム列 $\boldsymbol{X}$ から，$\boldsymbol{v_X}$ のなかで不定値を含むベクトル組に対応するフレームを除く．そんな処理を施したフレーム列を，次のように表す．

$$\boldsymbol{X}' = [x'_1, x'_2, \cdots, x'_{m'}] \tag{7}$$

当然，$\boldsymbol{X}'$ は $\boldsymbol{X}$ と同じか，短くなり，$m' \leq m$ で，$m' = 0$ の場合，$X$ は計算対象とはならない．$Y$ でも同様の手続きを行い，$\boldsymbol{Y}'$ を得る．たとえば，実験データの「若い」は $m$ が 120 のフレーム列で，不定値をふくむベクトル組は 48 で，それらを除くと $m'$ が 72 のフレーム列となる．

DTW 距離を計算する際の $\delta$ は，不定値を含まないベクトル組同士を考えればよく，次のように与える．

$$\delta(x, y) = 1 - \sum_{i=1}^{16} w_i \frac{v_{x,i} \circ v_{y,i}}{\|v_{x,i}\| \|v_{y,i}\|} \tag{8}$$

$a \circ b$ は $a$ と $b$ の内積，$\|a\|$ は $a$ のノルムを表す．$w_i$ は $w_i \geq 0$, $\sum_{i=1}^{16} w_i = 1$ で，それぞれの特徴に対する重みである．

DTW 距離はフレーム長が大きいもの同士の方が大きくなる傾向はある．$\delta$ が (8) のとき，DTW 距離の最大値は $m' + n' - 1$ なので，手話表現 $X, Y$ の距離を次のように与える．

$$d(\boldsymbol{X}, \boldsymbol{Y}) = \frac{d_{\mathrm{DTW}}(\boldsymbol{X}', \boldsymbol{Y}')}{m' + n' - 1} \tag{9}$$

実験に使用したデータは，4.2 の (2)，すなわち全国手話検定 4,5 級テキスト [13], [14] の 938 語を【A】でデータ化

[*10] 語と表現を厳密に区別しないので，動作が類似した手話の表現同士も適宜，動作類語で包摂する．


**図 3** 手話授業における人の認識 (上段 低解像度・下段 高解像度)
**Fig. 3** Person recognition in a sign language class (top: low resolution, bottom: high resolution)

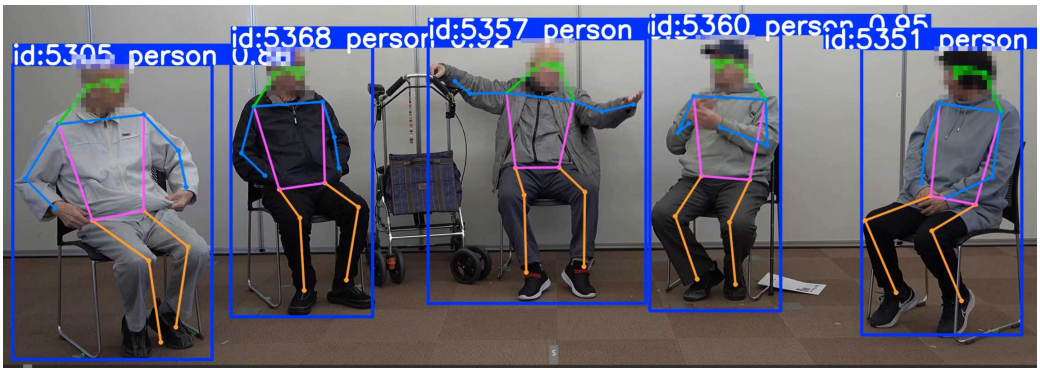


**図 4** 複数名の手話会話の映像の 1 シーン (高解像度)
**Fig. 4** A scene from a multi-person sign language conversation video (high resolution)

**図 5** 「若い・黒い・高校」の手話表現
**Fig. 5** Sign language expressions for “young, black and high school”

したもので，組み合わせ数は約 44 万である．ただし，語 $X, Y$ 間で，$m \leq n$ のとき，$n/m < 1.2$ の 404,268 組を対象とした．(8) の $w_i$ は均等に 1/16 とした．

$d$ が最も小さなものは，0.006 の「うらやましい」と「もう 1 度」である (**図 6** 上段)．$d$ が小さい 50 組に対して動作類語と判断されたものは 20 組だった．類似していない場合の原因のひとつは，一方もしくは両方の語に，不定値を含むベクトル組が多いことである．

語 $X$ を固定し，$X$ に対して $d$ が小さな語 $Y$ をみていくと，より効率的に動作類語を見出すことができる．人による語間の類似性判定では，とくに類似していない場合，短時間，容易につくことが多い．**図 5** の 3 つの表現 (若い・黒い・高校) も，それぞれの $d$ が小さな語を個別にみていくことで容易に得られる．最終的に人による判断を経るならば，許容できる適合性と考えられるかもしれない．

そのほか，「手話」と「車いす」といった，左右の腕のそれぞれ動く方向，手指形は同一だが，動きの位相が異なるようなものも，とらえられる (**図 6** 下段)．「車いす」に対して「手話」は 6 番目，「手話」に対して「車いす」は 3 番目に近い語として列挙される．

このように，手指標識に関する情報が不完全なデータでも，手話表現間の類似性をとらえることができていた．

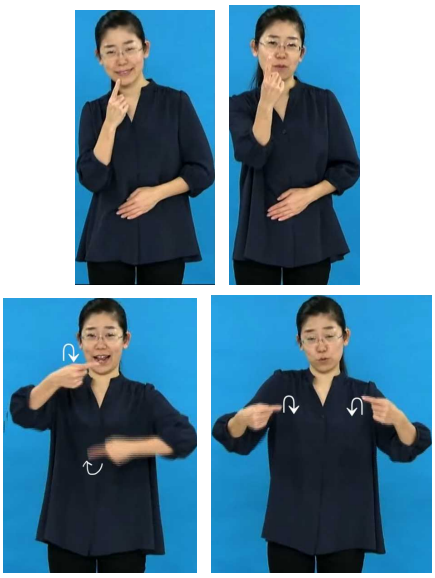

**図 6** 上段「うらやましい・もう 1 度」 下段「手話・車いす」
**Fig. 6** Top: “envy, again”, bottom: “sign language, wheelchair”

## 5. おわりに

本研究では，人認識 AI を通すことで，手話映像を低コストに言語データとして蓄積し，活用していく方向性を示した．[5] などの精密に作り上げられた質の高い手話言語データと異なり，現在の人認識 AI をもってしても，手指標識でさえ，完全には記述しきれない．具体的な課題やデータの取り扱い方によっては不完全性に頑健で，実際に動作類語の検出では従来とは異なる特徴をとらえることもできている．また，低コストに構築できるため，用例検索を指向するような，一定の量が求められるようなことにも，拡張していくことができるだろう．今後，こういった手話言語データの確実な利活用法や，不完全な手指標識データ下での検索，データの補完を検討する．

### 参考文献

[1] 広間陽, 平塚茂幸, 池田尚志: 手話における手指動作記述文の解析と手話単語の動作類似性について, 言語処理学会第 8 回年次大会, pp. 144-147 (2002).
[2] Honda, K., Yano, M., Hasegawa, Y., Tanaka, S.: Hand Gesture Recognition for Robot Control Using the Leap Motion Controller, *ICISIP2017*, pp. 83-88 (2017).
[3] 本田久平, 田中省作, 長谷川由美, 宮崎佳典: 手関節の AI 認識を用いた指文字学習支援 Web システム, 日本教育工学会 2023 年春季全国大会, 3-S04D2, pp 251-252 (2023).
[4] 菊池浩平: 手話言語の言語管理研究へ向けて, 千葉大学社会文化科学研究所研究プロジェクト報告書, 129, pp. 79-90 (2006).
[5] 工学院大学多用途型日本手話言語データベース（KoSign）, 情報学研究データリポジトリ, `https://www.nii.ac.jp/dsc/idr/rdata/KoSign/` (Last access: 2025/1/5).
[6] 長嶋慎二, 原大介, 堀内靖雄, 酒向慎司, 渡辺桂子, 菊澤律子, 加藤直人, 市川熹: 多様な研究分野に利用可能な超高精細・高精度手話言語データベースの開発, 言語資源活用ワークショップ 2018, pp. 148-155 (2018).
[7] Stokoe, W.: Sign Language Structure, Studies in Linguistics Occasional Papers 8, Buffalo: University of Buffalo Press (1960).
[8] Sutton, V.: Textbook and Workbook, 3rd ed. The deaf action comittee for SignWriting (2002).
[9] 田中省作, 本田久平, 長谷川由美: ネットワーク分析に基づいた日本手話初学者の指文字読取の誤り分析, 統計数理研究共同研究リポート 2021, 450, pp. 23-44 (2021).
[10] Google MediaPipe Team: MediaPipe, `https://google.github.io/mediapipe/` (Last access: 2025/1/5). 2022.6.17).
[11] 櫻井保志, 吉川正俊: ダイナミックタイムワーピングのための類似探索手法, 情報処理学会論文誌, 45(4(TOD21)), pp. 23-36 (2004).
[12] ultralytics YOLO Vision, `https://docs.ultralytics.com/` (Last access: 2025.1.5).
[13] 全国手話研修センター (編): DVD で学ぶ手話の本 全国手話検定試験 4 級対応 三訂, 中央法規出版 (2016).
[14] 全国手話研修センター (編): DVD で学ぶ手話の本 全国手話検定試験 5 級対応 三訂, 中央法規出版 (2016).