Vol.2022-CVIM-230 No.14
2022/5/12

時空間特徴量を用いた手話単語認識における未知単語判定手法の検討

山田 大記[†1]　井上 勝文[†1]　岩村 雅一[†1]　Partha Pratim Roy[†2]　吉岡 理文[†1]

概要：手話単語認識にはラベル付き学習用データが必要となる. しかし, 手話動画のラベル付けには専門知識が必要なため, データには限りがある. そのため, 自動で手話単語動画をラベル付けする手法が求められるが, これを実現するには高精度な認識モデルが必要であるというジレンマがある. 学習済み単語 (既知単語) のデータ数増加に対しては, 既存モデルを活用できるかもしれないが, 学習外単語 (未知単語) に対しては, 既知単語のいずれかに認識しようとする既存モデルの活用は困難である. この問題に対応するためには, 既知単語か未知単語かで処理が異なるため, まずラベルの付いていない手話単語動画の手話動作が既知単語か未知単語かを区別する必要がある. 本研究では, 未知単語動画の有効活用のために, ラベルの付いていない手話単語動画が未知単語かを判定する手法を提案する．具体的には, 3DCNN ベースの手法により手話動作を時空間特徴量列で表し, 特徴量同士の類似度を求めることで，未知単語かを判定する.

1. はじめに

高精度な手話単語認識を実現するためには, ラベル付きの学習用データが必要となる. しかし, 手話単語動画に対してラベル付けを行うためには専門的な知識が必要でありこのようなデータには限りがある. そこで, ラベルのない手話単語動画に対して少ない労力でラベルをつけることでラベル付きデータを増やすことが求められる. ラベルのない手話単語動画は, ラベル付き手話単語動画を収録するデータセットをもとにして考えると, 2 種類に分けることができる. 1 つ目は, ラベルのない手話単語動画が, データセット内に存在するラベルをもつ場合であり, これを既知単語の手話単語動画と定義する. 2 つ目は, ラベルのない手話単語動画が, データセット内に存在しないラベルをもつ場合であり, これを未知単語の手話単語動画と定義する.

また, 自動認識で使用されるようなモデルの多くは, 未知単語の手話単語動画であってもデータセット内のラベルに割り当てようとする. 本研究は, 入力される手話単語動画が未知単語の手話単語動画であるかの判定を行うことを目的とする. このような判定が可能になると, 未知単語の手話単語動画を見つけることができ, 未知単語の手話単語動画のみに集中した効率的なラベル付けが可能になると考えられる. この目的の実現のために, 提案手法では 3DCNN ベースの手法により手話単語の一連の動作を時空間特徴量列で表現し, この特徴量列同士の類似度を求めることで, 未知単語か否かを判定する.

[†1] 現在，大阪公立大学大学院情報学研究科
Presently with Graduate School of Informatics, Osaka Metropolitan University
[†2] 現在，Department of Computer Science and Engineering, Indian Institute of Technology

2. 関連手法

本節では，まず従来の手話単語動画の認識手法について述べる. そして, 本研究の提案手法に用いる I3D という手話単語認識において高い精度を記録した 3DCNN モデルと, Spotting タスクにおいて高い精度を記録した I3D+MLP モデルという 3DCNN モデル, そして手話単語認識において識別器として広く用いられている DTW について述べる.

2.1 手話単語動画の認識手法

初期の手話認識手法では, ハンドクラフト特徴量を用いる手法が主流である. 例えば, Histogram of Oriented Gradients(HOG) 特徴量を用いた手法 [4], Scale Invariant Feature Transform(SIFT) 特徴量を用いた手法 [7] などが存在する. そして, 特徴量をもとに手話の識別を行う識別器が存在している. 例えば, Hidden Markov Models(HMM) を用いた手法 [8], Dynamic Time Warping(DTW)[3] を用いた手法がある.

近年では, Deep Neural Network(DNN) を用いた手話認識手法が多く存在している. 手話認識手法において, 手話動画

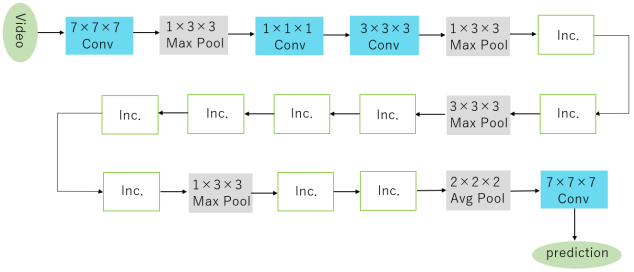

図 1: I3D の概要図

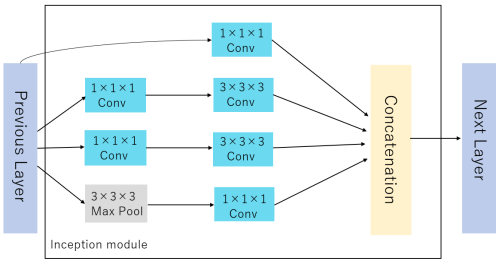

図 2: Inception module の概要図

の外見情報を用いる手法が広く用いられている.

手話動画の外見情報を用いる手法は, 2D Convolutional Neural Network(2DCNN) を用いる手法と 3D Convolutional Neural Network(3DCNN) の 2 種類に分けられる. 2DCNN を用いる手法 [2] は, 2DCNN と Recurrent Neural Network(RNN) を組み合わせた手法が一般的である. 2DCNN は手話動画の各入力画像の空間的な特徴を抽出することができ, RNN は抽出された特徴量をもとに, 各入力画像間の時間的な特徴を捉えることができる. 一方で, 3DCNN は入力する手話動画から時間的な特徴と空間的な特徴の両方を同時に学習できる. そして, 3DCNN の中でも I3D と呼ばれる 3DCNN モデルが各手話データセットにおいて高い認識精度を達成している [2].

本研究では, ラベルのない手話単語動画が未知単語であるかを判定することを目的とした提案手法に, 手話単語認識において高い精度を記録した I3D[1] と, 識別器として DTW[9] を用いる.

2.2 I3D

本節では, I3D(Two-Stream Inflated 3D ConvNet)[1] について述べる. I3D[1] とは, 行動認識に対して高精度を記録した 3DCNN モデルである. I3D のモデル構造は図 1 の通りである. I3D は, 畳み込み層と Max Pooling 層と Inception Module を組み合わせた構造である. I3D の構造において優れた点として, Inception module[6] を導入した点が挙げられる. Inception module とは, 複数の畳み込み層や pooling 層から構成されるネットワークのことである. I3D における Inception module の構造は図 2 の通りである. Inception module では, 複数の畳み込み層を並列につなげ, それぞれの層での計算の結果を最後に連結する. これにより, 層を深

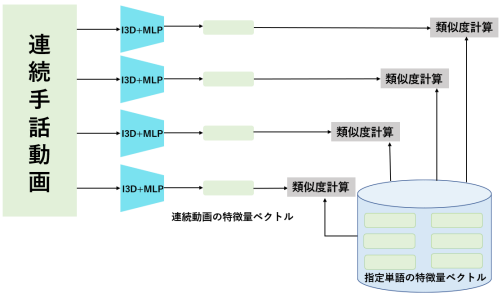

図 3: Spotting 手法の流れ

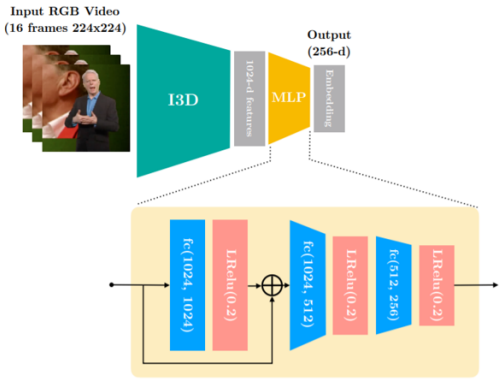

図 4: I3D+MLP モデルの概要図 [5]

くすることが可能となり行動分類精度の向上につながる. 以上のような利点から, I3D を用いることで行動認識に適した時空間特徴量を取り出すことができる. 本研究では, 手話動作における特徴量を抽出するために I3D を用いる.

2.3 spotting タスクに利用されている I3D+MLP モデル

本節では, 連続手話動画の Spotting に関する研究 [5] に用いられている I3D+MLP モデルについて述べる. ここでの Spotting とは, 連続手話動画の中で特定の単語の手話がその連続動画のどこに位置しているのかを推定するタスクのことである. I3D+MLP モデルとは, I3D[1] と Multilayer Perceptron(MLP) を組み合わせたモデルである. この Spotting タスクにおいては, 特定の単語の手話と連続動画内の手話動作の類似性を測る必要があり, 手話動作の特徴量抽出の部分において I3D+MLP モデルが利用されている. つまり, I3D+MLP モデルによって抽出される特徴量は手話動作の特徴を的確に捉えたものであるといえ, 本研究においても活用できるのではないかと考えられる.

続いて, この研究で提案されている手法の流れを述べる. 手法の流れを図 3 に示す. まず, Spotting をする手話単語を指定する. 次に, I3D+MLP モデルを用いて, 指定した手話単語動画の特徴量ベクトルを抽出する. そして, 連続手話動画を数フレームずつずらしながら I3D+MLP モデルに入力し, 時系列順に特徴量ベクトルを抽出する. 最後に, 時系列順に連続手話動画の特徴量ベクトルと 手話単語動画の特徴量ベクトルとのコサイン類似度を測っていく. この類似度

がしきい値より大きい部分が存在すれば, その部分が指定した手話単語を表す部分であると判定する.

この研究で用いている I3D+MLP モデルの概要図を図 4 に示す. I3D は, 前節で説明した通り行動認識に適したモデルである. MLP は, 全結合層と出力層からなるモデルである. I3D から出力される特徴量ベクトルを MLP に入力することで, 手話動作の潜在空間を表現する特徴量ベクトルに変換することができる. このモデルを用いることにより, Spotting タスクにおいて高精度を記録している.

2.4 Dynamic Time Warping

本節では, Dynamic Time Warping(DTW)[9] について述べる. DTW とは, 時系列データ同士の距離や類似度を測る手法である. DTW は, 2 つの時系列データの各点の距離を総当たりで比較し時系列同士の距離が最短となるパスを見つける. 以下では, DTW の処理の流れを述べる.

2 つの時系列データ $S = \{s_1, s_2, ...s_M\}$, $T = \{t_1, t_2, ...t_N\}$ が与えられているとする. なお, M と N はそれぞれの時系列データ S, T の長さである. このとき, 距離関数 d を式 (1) に定義する. なお, L_2 ノルムを用いた場合の定義である. i は M 以下の自然数, j は N 以下の自然数をとる.

$$d(i,j) := \|s_i - t_j\|_2 + \min \begin{cases} d(i, j-1) \\ d(i-1, j) \\ d(i-1, j-1) \end{cases} \quad (1)$$

このとき, $d(0,0) = 0$, $d(i,0) = d(j,0) = \infty$ であるとする. このような距離関数 d を用いて, 時系列データ S, T 間の距離 $D(S, T)$ を式 (2) に表現する.

$$D(S,T) = d(M,N) \quad (2)$$

以上のように動的計画法を用いることによって DTW は時系列データ同士の距離を計算している.

3. 提案手法

本節では提案手法について述べる. 提案手法の流れを, 図 5 に示す. まず, CNN モデルを用いて, 学習用データとテスト用データの時空間特徴量を抽出する. 次に, 時空間特徴量を使用して手話動作の特徴を表す行列 (軌跡と呼ぶ) を作成する. 最後に, 作成したテスト用データの軌跡と学習用データの軌跡との類似度を DTW を用いて求め, 判定する. 以下では, 手話単語動画の時空間特徴量を抽出して軌跡を作成する特徴抽出部と, 軌跡から類似度を求め判定する判定部に分けて詳細に述べる.

3.1 特徴抽出部

本節では, 特徴抽出部について述べる. 特徴抽出部は, CNN モデルを用いた時空間特徴量抽出と, 抽出された特徴量からの軌跡作成という 2 つの処理からなる.

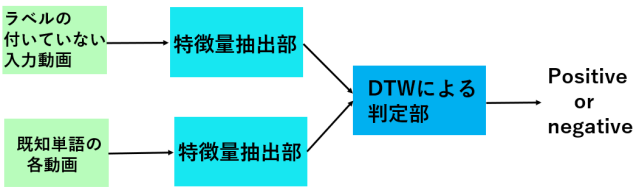

図 5: 提案手法の概要図

まずは, 時空間特徴量抽出の部分について述べる. 手話単語動画の特徴を十分に表現する特徴量を出力させるために, I3D+MLP モデルを用いて, 時空間特徴量抽出を行う. 本研究では, 図 4 に示す I3D+MLP モデルを用いるが, このモデルでは 16 フレームの動画を入力としている. そのため, 入力する手話単語動画に対してフレームをずらしながら 16 フレームごとに順に特徴量ベクトルを抽出する. 動画分割の処理の流れとしては, まず動画における最初のフレーム (0 フレーム目) から 15 フレーム目までの 16 フレームを I3D+MLP モデルに入力する. 次に, 8 フレーム目から 23 フレーム目までの 16 フレームを I3D+MLP モデルに入力する. このとき, 動画内の手話動作をより正確に細かく表現した特徴量を抽出できるように, 16 フレームの半分である 8 フレーム分の重なりをもたせる. このような処理によって, 手話動画を 16 フレームずつに分割する.

続いて, 特徴量を用いた軌跡の作成方法について述べる. 本研究では, 手話動作の特徴を表現するために軌跡を作成する. そして, 軌跡の作成方法として 2 つの方法を提案する.

1 つ目は, 特徴量ベクトルを時系列順に並べて軌跡を作成する方法である. I3D+MLP モデルから抽出される特徴量ベクトルは, 手話認識における重要な手話動作の特徴を的確に表現をしたものである. この特徴量ベクトルを時系列順に並べることにより作成される行列は, 手話単語動画全体の特徴を表現できると考えられる. そこで, 本研究では, I3D+MLP モデルから出力される特徴量ベクトルを時系列順に並べることで作成した行列を “軌跡 1”と呼ぶ. 特徴量ベクトルを $\boldsymbol{v}$, 特徴量ベクトルの個数を N, 軌跡 1 を X_1 とするとき, 軌跡 1 は式 (3) で表現することができる.

$$X = \begin{pmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \\ \vdots \\ \boldsymbol{v_N} \end{pmatrix} \quad (3)$$

図 6 に作成方法の例を示す. まず, 手話単語動画から 16 フレームごとに特徴量ベクトルを時系列順に出力する. 図 6 において説明のため, 7 個の特徴量ベクトル $\boldsymbol{v}$ が時系列順に出力されたと仮定し, 出力された特徴量ベクトルを特徴量ベクトル $\boldsymbol{v_1}$ から特徴量ベクトル $\boldsymbol{v_7}$ とする. そして, この特徴量ベクトルを時系列順に並べることで行列 (軌跡 1) を作

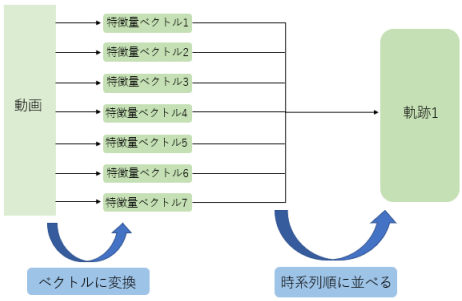

図 6: 軌跡 1 の作成方法の例

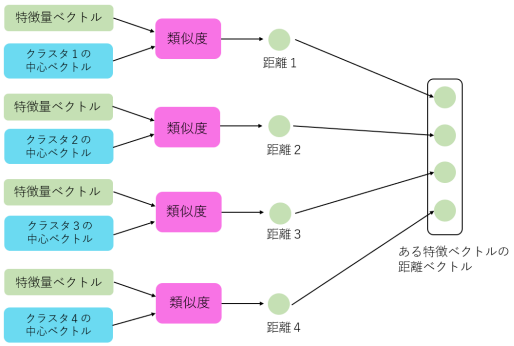

図 7: 距離ベクトルの作成方法

成する.

2 つ目は, k-means 法を用いることで似ている動作ごとのクラスタを作成し, その各クラスタとの距離をもとに軌跡を作成する方法である. k-means 法を I3D+MLP モデルから抽出される手話動画の特徴ベクトルに適用することで, 似ている動作を表現する特徴量ベクトルは同じクラスタに, 似ていない動作を表現する特徴量ベクトルは違うクラスタに分類することが可能であると考える. I3D+MLP モデルから抽出される特徴量ベクトルは手話の動作特徴を捉えたものであるので, k-means 法によって作成される各クラスタ内の特徴量ベクトルは似ている動作を表す特徴ベクトルが集まっていると考えられる. 2 つ目の軌跡は, このような動作ごとのクラスタからの距離を用いて作成する.

作成方法を説明する. まず, 既知単語の手話単語動画からなる学習用データを, フレームをずらしながら特徴量ベクトルに変換する. 次に, 学習用データから抽出されるすべての特徴量ベクトルに対して k-means 法を適用する. これにより, 動作ごとにまとまりのあるクラスタを作成する. 続いて, テスト用データであるラベルの付いていない手話単語動画から, 16 フレームごとに特徴量ベクトルを抽出する. そして, 図 7 のように, ラベルのない手話単語動画の特徴量ベクトルと, k-means 法により作成された各クラスタの中心ベクトルとの距離を順に算出し, 距離を並べたベクトル (距離ベクトル) を作成する. ラベルのない手話単語動画を分割した後の 16 フレーム分動画の特徴量ベクトルを $\boldsymbol{n}$, 各クラスタの中心ベクトルを $\boldsymbol{c_i}$, クラスタ数を k とするとき, i 番目 ($i \in 1, 2, ..., k$) のクラスタとの距離 w_i は式 (4) で表すことができる.

$$w_i = 1.0 - \frac{\boldsymbol{n} \cdot \boldsymbol{c_i}}{\|\boldsymbol{n}\| \, \|\boldsymbol{c_i}\|} \tag{4}$$

なお, この距離 w_i はコサイン距離である. これは, Monemi ら [5] が I3D+MLP モデルから出力したベクトル同士の比較をコサイン距離で行っていたことから, 本研究においてもコサイン距離によってクラスタリングを行う. 距離ベクトル $\boldsymbol{h}$ は式 (5) で表される.

$$\boldsymbol{h} = (w_1, w_2, \ldots, w_k) \tag{5}$$

最後に, 距離ベクトルを時系列順に並べて行列を作成し, “軌跡 2”とする. 距離ベクトルを $\boldsymbol{h}$, クラスタ数を k, ラベルのない手話単語動画を分割した後の 16 フレーム動画の数を N, 軌跡 2 を X_2 とするとき, 軌跡 2 は式 (6) で表す.

$$X_2 = \begin{pmatrix} \boldsymbol{h_1} \\ \boldsymbol{h_2} \\ \vdots \\ \boldsymbol{h_N} \end{pmatrix} \tag{6}$$

このようにして作成される行列を軌跡 2 とする.

3.2 判定部

本節では, 判定部について述べる. 本研究では, 長さの異なる時系列データの類似度を測る一般的な手法の DTW を用いて, 手話動画から得た軌跡の類似度を測定し, 入力単語が既知単語か未知単語かを判定する. 以下, 判定の流れを述べる.

まず, 調べたい未知単語の軌跡とすべての既知単語の軌跡との間の距離を DTW により算出する. これにより, 1 つの未知単語の軌跡に対して既知単語の軌跡の個数分の距離が出力される. このとき, 時系列データ同士の距離の和が DTW により出力される距離であるため, 判定に用いる距離はフレーム数に大きく影響を受ける. そこで, 2 つの動画のフレーム数の平均の平方根で距離を割るという処理 (フレーム正規化) を加える. 2 つの手話単語動画のフレーム数をそれぞれ m_1, m_2 とし, DTW によって求められる距離を D とするときフレーム正規化後の距離 D_{new} を式 (7) のように求める.

$$D_{new} = \frac{D}{\sqrt{\frac{m_1 + m_2}{2}}} \tag{7}$$

つぎに, すべての既知単語の軌跡との距離の中で最小の値を求める. 手話動作というのは, 基本的に動作が類似しているものは大まかに同じような意味を示すといわれている. そのため, 手話動作の類似性から未知単語であるか判定しようと考える. この最小の値がしきい値 t より小さい場合, 既知単語の軌跡と十分近い距離にあり, 手話動作が類似し

ていると考えられるため, その手話単語動画は既知単語の動画であると判定する. また, この最小の値がしきい値 t より大きい場合, 既知単語の軌跡と遠い距離にあり手話動作が類似していないと考えられるため, その手話単語動画は未知単語の動画であると判定する.

4. 実験

本研究では, ラベルの付いていない手話単語動画が未知単語が既知単語のものであるか未知単語のものであるかを判定する実験を行った. 以下に詳細を述べる.

4.1 実験方法

本研究では, イギリス手話単語動画データセットである BSLDict を用いた. BSLDict[5] は, 14062 本の手話単語動画で構成されている. この BSLDict に対して, 学習用データ数が 10550 本, ラベルの付いていないテスト用データ数が 3512 本になるように分割をして実験を行った. この学習用データに含まれている手話単語動画はすべて既知単語のものである. テスト用データに含まれている手話単語動画は, 既知単語のものと未知単語のものの 2 種類があり, それぞれの動画数が 1836 本と 1676 本である. 特徴抽出に用いる I3D+MLP モデルは, Monemi らが使用した重みをそのまま用いる. DTW によって算出された距離に対するしきい値 t を 0.00 から 0.01 ずつ変化させ, このしきい値が 30.00 になるまで実験を行った. 未知単語判定の定量評価は正解率, 真陽性率, 特異率を用いた. 本研究では, 既知単語である動画を positive, 未知単語単語である動画を negative とした. そして, テスト用データが既知単語のものか未知単語のものであるかを判定し, 判定精度を調査する. 正解率は式 (8), 真陽性率は式 (9), 特異率は式 (10) のように計算する.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{10}$$

式中の P/N は positive/negative, T/F はその判定が正しい/誤りであることを表している. TP とは, 既知単語の動画が正しく既知単語と判定できた場合の動画数である. TN とは, 未知単語の動画が正しく未知単語と判定できた場合の動画数である. FP とは, 未知単語の動画が間違って既知単語と判定された場合の動画数である. FN とは, 既知単語の動画が間違って未知単語と判定された場合の動画数である. 正解率は全判定の精度を, 真陽性率は既知単語判定の正確さを, 特異率は未知単語判定の正確さを表現している. 本研究では正解率, 真陽性率, 特異率をそれぞれ%表記したものを評価に用いた.

表 1: 軌跡 1 を用いた判定の実験結果 (%)

フレーム正規化	正解率	真陽性率	特異率
あり	52.53	51.79	53.21
なし	51.91	50.66	53.05

4.2 軌跡 1 に対する未知単語判定の実験

4.2.1 定量評価

本実験では, 軌跡 1 を用いた未知単語判定の精度を評価する. 判定の際に使用する DTW に用いる距離指標は, I3D+MLP モデルから出力される特徴量ベクトルを並べた軌跡同士の距離を求めるため, Monemi ら [5] が I3D+MLP モデルから出力されるベクトルの比較に用いていたコサイン距離を適用する.

軌跡 1 を用いて未知単語判定を行った結果を表 1 に示す. 表 1 より, 軌跡 1 を用いた未知単語判定の精度は 50%をわずかに超える程度であった. このことから, 未知単語判定が難しい例も多く存在していることがわかった. また, 表 1 よりフレーム正規化をすることで精度が上昇したことがわかった. このことから, フレームの正規化が未知単語判定において有効な処理であったことがわかった.

続いて, 横軸にテスト用データの各動画において判定に用いられるコサイン距離の値を, 縦軸に個数をとったときのグラフを図 8 に載せる. このグラフは, 軌跡 1 を用い, フレーム正規化を施した場合のものである. また, 青色は positive, オレンジ色は negative を表している. 図 8 において, 横軸であるコサイン距離が 1.0 と 2.0 の間の部分に注目する. この部分から, positive の手話単語動画が negative の手話単語動画よりも多いことがわかる. また, positive の手話単語動画のコサイン距離の平均値は 1.65 であり, negative の手話単語動画のコサイン距離の平均値は 1.71 である. このことから, positive の手話単語動画のコサイン距離は, negative の手話動画のコサイン距離と比べてわずかに小さいことがわかる. しかし, positive と negative のグラフが完全に分離をしているとは言えず, DTW 以外の時系列データの類似性を測る手法を検討する必要があると考える. また, 似たような軌跡が多くグラフの分離ができていない可能性も考えられる. そのため, 軌跡の作成方法を変更する必要もあると考える.

4.2.2 定性評価

本実験では, 4 つに分けて定性評価を行う.

まず, テスト用データが既知単語のものであり実際に正しく既知単語のものと判定できた場合である. 図 9 にその例を示す. 図 9 に示す 2 つの手話動画の軌跡間の距離は 1.04 であった. 図 9 における上部分と下部分の画像を比較すると, どちらもこぶしを作るような動作が含まれていることがわかる. このような似た動作を類似性が高いものとして扱い, 既知単語のものとして判定できていることがわかる.

次に, テスト用データが未知単語のものであり実際に正しく未知単語のものと判定できた場合である. 図 10 にその

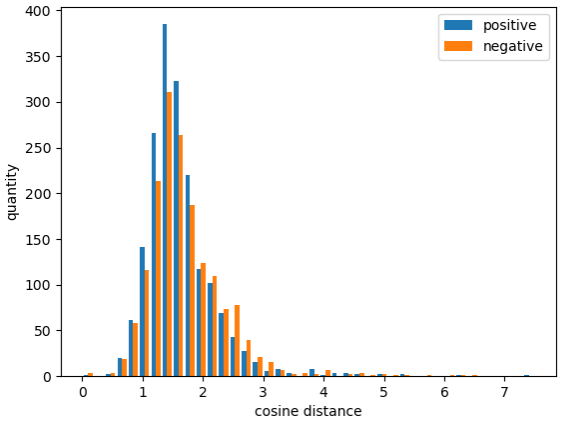

図 8: 軌跡 1 を用いた場合のコサイン距離の分布のグラフ

例を示す. 図 10 に示す 2 つの手話動画の軌跡間の距離は 2.36 であった. 図 10 における上部分と下部分の画像を比較すると, どちらも手をくっつけるような動作を含んでいることがわかる. しかし, これらの動画において, 図 10 の下部分の画像の手話はこの動作のみからなる手話であるのに対し, 図 10 の上部分の方の手話はこの動作以外の動作も組み合わせた手話であった. このような違いを的確に捉え, 未知単語のものであると正しく判定できていることがわかる.

次に, テスト用データが既知単語のものであるが誤って未知単語のものと判定した場合である. 図 11 にその例を示す. 図 11 に示す 2 つの手話動画の軌跡間の距離は 2.18 であった. 図 11 における上部分と下部分の画像を比較すると, どちらも頭の上に腕を上げる動作であるが, 未知単語のものと誤って判定をしている. これは, 上部分の画像の動画が 102 フレーム, 下部分の画像の動画が 111 フレームとかなり長い動画であるため, DTW により得られる距離がかなり大きくなったことが原因であると考えられる.

最後に, テスト用データが未知単語のものであるが誤って既知単語のものと判定した場合である. 図 12 にその例を示す. 図 12 に示す 2 つの手話動画の軌跡間の距離は 0.82 であった. 図 12 における上部分と下部分の画像を比較すると, 手の位置はあまり似ていないように見える. しかし, 既知単語のものと判定している. この原因としては, 上部分の画像の動画が 53 フレーム, 下部分の画像の動画が 53 フレームとかなり短い動画であるため, DTW により得られる距離が小さく出てしまったことが原因であると考えられる.

これらの結果から, DTW やフレームの正規化に改善の余地があり, 時系列データの比較方法の検討や正規化の方法の変更が必要であると考えられる.

4.3 軌跡 2 に対する未知単語判定の実験

4.3.1 定量評価

本実験では, 軌跡 2 を用いた未知単語判定の精度を評価

図 9: 既知単語のものと正しく判定できた場合の手話動作の比較 (上部分: 最小の距離となった手話動画の動作, 下部分: テスト用データの手話動画の動作)

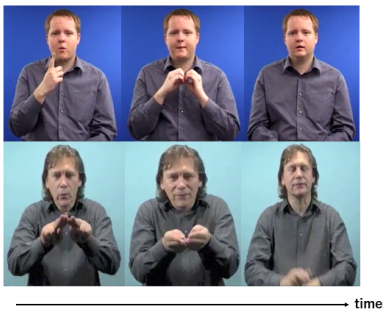

図 10: 未知単語のものと正しく判定できた場合の手話動作の比較 (上部分: 最小の距離となった手話動画の動作, 下部分: テスト用データの手話動画の動作)

図 11: 既知単語のものと正しく判定できなかった場合の手話動作の比較 (上部分: 最小の距離となった手話動画の動作, 下部分: テスト用データの手話動画の動作)

する. 軌跡 2 においては, k-means 法を用いて実験を行う. まず, k-means 法に関する実験条件について述べる. k の値をどのように設定すれば, 同じクラスタに同じ単語の特徴ベクトルが集まるのかを調査した. 具体的には, 各 k の値で k-means 法によるクラスタリングを行い, 各クラスタに特徴ベクトルのもとである手話動画を観察し, 最もクラスタごとの動作の類似性が高いと考えられる k の値を, 手動で動画を確認しながら調べた. そして, そのような k の値が 300 であったため, 軌跡 2 の作成の際に用いる k-means 法において k の値を 300 に設定した. また, Monemi ら [5] はコサイン距離を用いて I3D+MLP モデルから出力されるベクトルの比較を行っていたため, 本実験においてもコサイン距

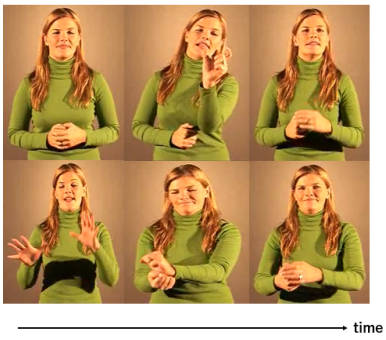

図 12: 未知単語のものと正しく判定できなかった場合の手話動作の比較 (上部分: 最小の距離となった手話動画の動作, 下部分: テスト用データの手話動画の動作)

表 2: 軌跡 2 を用いた判定の実験結果 (%)

フレーム正規化	正解率	真陽性率	特異率
あり	52.31	54.83	50.00
なし	51.77	53.10	50.54

離を用いて k-means 法を行う. そして, 判定の際に使用する DTW に用いる距離指標は, 距離ベクトルを並べた軌跡同士の距離を算出するため, ユークリッド距離を適用する.

軌跡 2 を用いて未知単語判定を行った精度を表 2 に示す. 表 2 より, 軌跡 2 を用いた未知単語判定の精度は 50 %をわずかに超える程度であった. しかし, 表 1 と比較してやや精度が低下していることが分かった. これは, 軌跡 2 は動画を特徴量ベクトルに変換をしてさらに距離ベクトルに変換をしているため, 軌跡 1 と比較して手話動作の情報量が落ちてしまったことが原因であると考えられる. そして, 表 2 よりフレーム正規化をすることでわずかに正解率が向上したこともわかった. このことから, フレーム正規化は軌跡 2 に対しても効果があると考えられる.

続いて, 図 8 と同様に, 横軸にテスト用データの各動画において判定に用いられるユークリッド距離の値を, 縦軸に個数をとったときのグラフを図 13 に載せる. このグラフは, 軌跡 2 を用い, フレーム正規化を施した場合のものである. また, 青色は positive, オレンジ色は negative を表している. 図 13 よりユークリッド距離が 2.5 付近の手話単語動画の数を見ると, positive の手話単語動画のほうが negative の手話単語動画よりも多いことが確認できる. また, positive の手話単語動画のコサイン距離の平均値は 3.52 であり, negative の手話単語動画のコサイン距離の平均値は 3.62 である. このことから, positive の手話単語動画の距離は, negative の手話単語動画の距離と比べてわずかに小さいことがわかる. しかし, 図 8 と同様に positive のグラフと negative のグラフが完全に分離しているとは言えない. これは, positive の手話単語動画と negative の手話単語動画の間で, 距離ベクトルを並べた軌跡 2 がはっきりと識別可能な特徴量ではないため, グラフの分離に失敗していると考えられる. また, 手

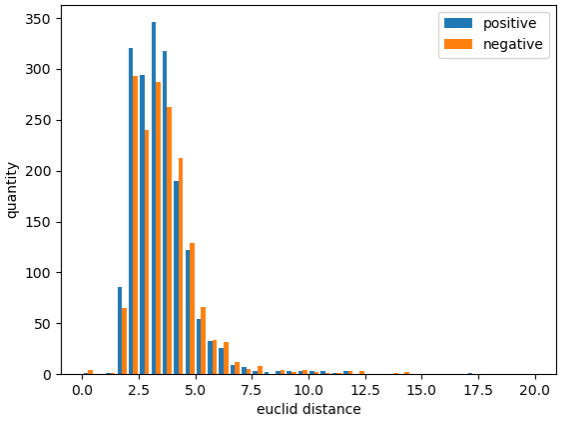

図 13: 軌跡 2 を用いた場合のユークリッド距離の分布のグラフ

話動作自体が類似していることで, グラフ分離自体が難しい可能性も考えられる. 以上のことから, 特徴抽出部の部分の手法を見直すことが必要であると考える.

4.3.2 定性評価

本実験では, 4 つに分けて定性評価を行う. まず, テスト用データが既知単語のものであり実際に正しく既知単語のものと判定できた場合である. 図 14 にその例を示す. 図 14 に示す 2 つの手話動画の軌跡間の距離は 3.25 であった. 図 14 における上部分と下部分の画像を比較すると, どちらも腕を広げる動作をしている. この部分が似ていることで, 既知単語のものと正しく判定できたと考えられる.

次に, テスト用データが未知単語のものであり実際に正しく未知単語のものと判定できた場合である. 図 15 にその例を示す. 図 14 に示す 2 つの手話動画の軌跡間の距離は 6.63 であった. 図 14 における上部分と下部分の画像を比較すると, 動作自体があまり似ていないように見える. この動作の違いを的確に捉えられたことで, 未知単語のものであると正しく判定できていることがわかる.

次に, テスト用データが既知単語のものであるが誤って未知単語のものと判定した場合である. 図 16 にその例を示す. 図 16 に示す 2 つの手話動画の軌跡間の距離は 6.12 であった. 図 16 における上部分と下部分の画像を比較する. このとき, 図 16 の上部分の画像に示す手話単語動画では同じ動作が 2 回繰り返されていた. そのため, よりフレーム数が大きくなりすぎたことで, フレームの正規化の効果が薄くなったと考えられる.

最後に, テスト用データが未知単語のものであるが誤って既知単語のものと判定した場合である. 図 17 にその例を示す. 図 17 に示す 2 つの手話動画の軌跡間の距離は 3.23 であった. 図 17 における上部分と下部分の画像を比較すると, 手の位置がどちらも似たような位置にあり, 動作も似ているように見える. そのため, 類似性が高く, 距離が短く計算されてしまったのではないかと考えられる.

図 14: 既知単語のものと正しく判定できた場合の手話動作の比較 (上部分: 最小の距離となった手話動画の動作, 下部分: テスト用データの手話動画の動作)

図 15: 未知単語のものと正しく判定できた場合の手話動作の比較 (上部分: 最小の距離となった手話動画の動作, 下部分: テスト用データの手話動画の動作)

図 16: 既知単語のものと正しく判定できなかった場合の手話動作の比較 (上部分: 最小の距離となった手話動画の動作, 下部分: テスト用データの手話動画の動作)

以上の結果から, やはりフレーム数の違いの影響がまだ大きい場合が存在することから正規化方法の検討が必要であると考えられる. また, 類似している動作が含まれることで, 未知単語の動画であっても距離が短く出てしまう場合が存在している. そのため, 軌跡の作成方法についても検討する必要があると考える.

5. おわりに

本研究では, 未知単語の手話単語動画を有効活用するその第一歩として, ラベルの付いていない手話単語動画が未知単語のものかを判定するという課題に取り組んだ. 特徴抽出部においては軌跡を作成し, 判定部では, DTW と呼

図 17: 未知単語のものと正しく判定できなかった場合の手話動作の比較 (上部分: 最小の距離となった手話動画の動作, 下部分: テスト用データの手話動画の動作)

ばれる時系列データ同士の類似度を計算する手法を用いてしきい値による判定を行った. そして, その際にフレームの正規化を行った. 実験の結果から, 提案した手法によって判定できるようになった場合も存在する一方で, まだ判定ができていない場合も多々存在することがわかった. そのため, 今後は, 軌跡の作成の方法を見直してより判定に適した特徴量抽出の方法を検討することや, DTW 以外の時系列データ同士の類似度を計算する手法を調査し, 存在判定を行う必要があると考える.

謝辞 本研究は, JSPS 科研費#19K12023 の補助による.

参考文献

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, July 2017.

[2] Dongxu Li, Cristian Rodriguez-Opazo, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. *WACV*, pp. 1448–1458, 2020.

[3] Jeroen F. Lichtenauer, Emile A. Hendriks, and Marcel J.T. Reinders. Sign language recognition by combining statistical dtw and independent classification. *TPAMI*, Vol. 30, No. 11, pp. 2040–2046, 2008.

[4] K Manjushree and Divyashree. Gesture recognition for indian sign language using hog and svm. *IRJET*, pp. 1697–1701, 2019.

[5] Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Watch, read and lookup: learning to spot signs from multiple supervisors. In *ACCV*, 2020.

[6] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[7] Alaa Tharwat, Tarek Gaber, Aboul Ella Hassanien, Mohamed K. Shahin, and Basma Refaat. Sift-based arabic sign language recognition system. In *AECIA*, 2014.

[8] Aliaa Abdel-Halim Youssif, Amal Elsayed Aboutabl, and Heba Hamdy Ali. Arabic sign language (arsl) recognition system using hmm. *IJACSA*, Vol. 2, , 2011.

[9] 櫻井保志, 吉岡正俊. 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, 3 月 2004.