2M-07

OpenPose を用いる簡便な手話認識手法

青野　由崇†　　花泉　弘†

法政大学情報科学部†

1. まえがき

手話はろう者が意思疎通を行う上で重要な手段であるが、手話の習得は難しく、健聴者の多くはその意味が理解できない．こうした背景から、手話の自動翻訳が望まれており、様々な取り組みが行なわれてきた[4][5]．文献[1]では、Kinect から得られた骨格ポーズデータの手の座標を取得し、それらをディープラーニングによって各手話単語に対応付けることで認識を行っていた．しかし、認識を行っているのは単体の手話単語のみであり、手話の文章を認識させるには再びその手話の文の映像からディープラーニングで学習させる必要があった．本研究は、特殊なカメラやディープラーニングによる学習を用いないより簡便な手話認識システムの実現を目的としている．ここでは OpenPose[6]を使用して簡単に手首や指の関節点群の動きを得た．パターンマッチングに基づいて認識する手法を提案する．

2. 原理と処理手順

使用する手話の映像は CG による手話単語の映像を教師データとし、手話者が行っている手話の映像を判別・翻訳させる．映像は 60FPS の環境で撮影したものである．まず、手話の動作を OpenPose で取得できる関節座標のとして抽出し、体格の違いを補正して各手話単語のデータベースを構築する。次に、CG と手話者が行っている手話に含まれている単語間の分割するための処理が必要になる．手首の動作速度から次の手話単語の動作に移るときに生じる極小点、極大点を用いて分割を行う[2]．分割したデータをマッチングさせる際に、人によって動作速度の違いがあるので Dynamic Time Warping(DTW)[3]を用いてマッチング処理を施す．分割した各手話単語の手首の移動量からお手本となる CG の手話単語のデータベースよ DTW によるパターンマッチングの類似度を求め手話の判別を行うという手法で手話の判別を行う．

2.1 関節点位置

OpenPose で取得できる関節座標は体 25 点、手指 42 点顔 70 点であるが本研究では、上半身の首、右の肩・右ひじ、右手首、左肩、左ひじ、左手首の計 7 点と手指の座標を用いる．上半身のそれぞれ部位を p_1、p_2、p_3、p_4、p_5、p_6、p_7 とし，この中で最も動きの少ない首 p_1 を基準にし、各部位までのベクトル $\overrightarrow{p_{12}}$、$\overrightarrow{p_{13}}$、$\overrightarrow{p_{14}}$、$\overrightarrow{p_{15}}$、$\overrightarrow{p_{16}}$、$\overrightarrow{p_{17}}$ を求める．手首の動きが体格の影響をなくすためこれらのベクトルを，首から両肩までの距離の平均値

$$l = \frac{|\overrightarrow{p_{12}}| + |\overrightarrow{p_{15}}|}{2} \quad (1)$$

で割って正規化して体格による大きさの違いを吸収する．各ベクトルに k 点移動平均をかけ、ノイズを低減させる．

次に、単語の分割のため、文献[2]の手法より手話の次単語への手の移動に生じる速度の極小点、極大点を検出のため手話の映像の映像から得られた手首の x,y 座標の移動量を求め、各移動量データに Δt 秒で割ってジェスチャーの速度を算出

$$\Delta t = t_{n+1} - t_n \quad (2)$$

n はフレーム番号であり、

$$P(t_n) = (x_n, y_n) \quad (3)$$

の n フレーム目の手首の位置を用いて

$$V = \frac{|P(t_{n+1}) - P(t_n)|}{\Delta t} \quad (4)$$

と図 1 (a)で示している手首の速度を算出する．手首の速度データ V から手話単語間に分割する極大値、極小値を求める．手話の単語分割に極大値、極小値より、最大となる極大値の位置を区切りの開始地点とし 0 近い値となる極小値、区切りの終了地点とする．手話に複数の手話単語が含まれている場合、最初の単語の分割が終了時、次の最大となる極大値の位置を区切りの開始地点とし 0 に近い値となる極小値を区切りの終了地点と繰り返し行う．

2.4 パターンマッチング

DTW は時間のずれを考慮して，二つの時系列データの 距離を算出する方法である．以下に概要を説明する．比較する時系列データ X、Y をそれぞれ N 個と M 個のデータ量を要素として持っている．

$$X= (x[1],x[2],\cdots ,x[N]),$$
$$Y= (y[1],y[2],\cdots ,y[M]) \quad (5)$$

まず，X の要素**x**[n]と Y の要素**y**[m]の類似を表す局コスト値をすべての時刻の組合せで計算する．

$$C(x[n],\ y[m]) = f(x[n] - y[m])$$
$$(n =1,\ldots ,N,\ m = 1,\ldots ,M) \quad (6)$$

ここで，f は要素間の相違を計算する評価関数である．

$$f(x[n] - y[m]) = \|x[n] - y[m]\|_2 \quad (7)$$

次に，再帰的に蓄積コスト値を計算し距離を求める．

$$D(n,m) = min\begin{Bmatrix} D(n-1,m-1) \\ D(n-1,m) \\ D(n,m-1) \end{Bmatrix} + c(x[n],y[m]) \quad (8)$$

4. 実験および考察

一般のカメラで撮影した CG の手話と人が行っている手話の映像に OpenPose により上半身の関節点座標を取得し、DTW による手話判別を行った．本研究で、使用した手話の映像はテンプレートの CG 手話単語①「私」、②「自転車」、③「壊す」、④「怒られる」、⑤「パソ

A Simplified Method for Recognizing Sign Language by Using
†Yoshitaka Aono, †Hiroshi Hanaizumi
Facully of Computer and Information Sciences, Hosei University

コン」、⑥「ハワイ」、⑦「行く」、⑧「先生」、⑨「妹」の教師データで構築したデータベースを用いた. 判別させる手話は被験者 A,B の 2 人で行った「私は自転車を壊す」、「私の妹がハワイに行く」、「私が先生に怒られる」を用いた. まずこれらのデータから手話の単語部分を分割するため文献[2]の手法を用いて、手首の速度から極小点、極大点より単語間分割を行った. 図 1 より被験者 A「私は自転車を壊す」の結果例を示す.

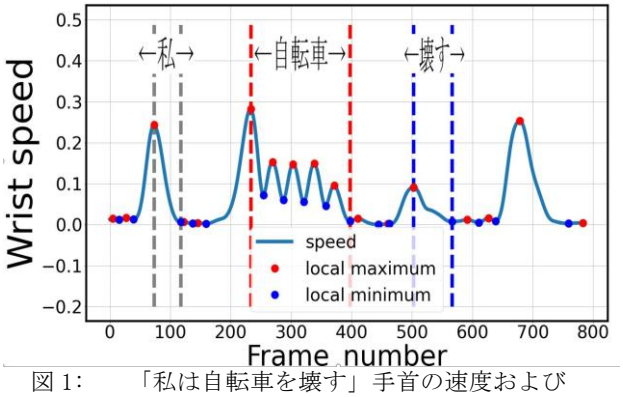

図 1: 「私は自転車を壊す」手首の速度および「私」、「自転車」「壊す」部分の分割領域

図 1 の赤点線の範囲は「私」、黒点線の範囲は「自転車」、青点線の範囲は「壊す」の手話を行っている部分の分割範囲を表している.「自転車」の範囲内に極大点４つ入っているがこの範囲に 0 近い値となる極小値がないため手話を続けて行っていると判別して範囲内の極大点 4 つは区切り位置の場所ではないとしている,

教師データの CG 手話単語データと被験者 A,B が行った手話の分割後、これらのデータをパターンマッチングさせる. 評価方法として分割した手話単語データと各テンプレートの CG 手話単語データを DTW による類似度で評価した.表 1 から表 3 より、同じような手首の動作する①「私」や⑨「妹」の場合でも、それぞれ対応した手話単語に対して倍以上の高い類似度が出た. これは、①「私」や⑨「妹」は手の形に違いがあり、手と指の関節点を用いたことでそれぞれ明確に区別できたのではないかと考えられる. その他の手話も表 1 から表 3 より太字箇所で対応した手話に対して他の手話よりも倍以上の高い類似度を示している. これら結果から手話に手と指の関節点を用いたことでそれぞれより明確に区別しマッチングできたのではないかと考えられる

表 1 「私は自転車を壊す」の類似度

	①	②	③	④	⑤	⑥	⑦	⑧	⑨
A①	**$4*10^{-1}$**	$1*10^{-3}$	$8*10^{-4}$	$8*10^{-4}$	$3*10^{-4}$	$1*10^{-3}$	$8*10^{-4}$	$1*10^{-3}$	$6*10^{-1}$
A②	$8*10^{-3}$	**$7*10^{-1}$**	$1*10^{-4}$	$2*10^{-4}$	$2*10^{-4}$	$8*10^{-4}$	$1*10^{-3}$	$1*10^{-3}$	$3*10^{-4}$
A③	$2*10^{-3}$	$3*10^{-3}$	**$3*10^{-1}$**	$1*10^{-4}$	$1*10^{-4}$	$2*10^{-4}$	$8*10^{-4}$	$9*10^{-4}$	$3*10^{-4}$
B①	**$6*10^{-1}$**	$8*10^{-3}$	$5*10^{-8}$	$7*10^{-4}$	$2*10^{-4}$	$1*10^{-3}$	$1*10^{-4}$	$1*10^{-3}$	$4*10^{-7}$
B②	$3*10^{-3}$	**$5*10^{-1}$**	$1*10^{-4}$	$1*10^{-4}$	$1*10^{-4}$	$8*10^{-4}$	$1*10^{-3}$	$1*10^{-3}$	$2*10^{-4}$
B③	$1*10^{-3}$	$2*10^{-3}$	**$4*10^{-1}$**	$1*10^{-4}$	$1*10^{-4}$	$1*10^{-4}$	$1*10^{-4}$	$7*10^{-4}$	$1*10^{-4}$

表 2 「私の妹がハワイに行く」の類似度

	①	②	③	④	⑤	⑥	⑦	⑧	⑨
A①	**$8*10^{-1}$**	$3*10^{-3}$	$8*10^{-4}$	$7*10^{-4}$	$3*10^{-4}$	$1*10^{-3}$	$1*10^{-4}$	$3*10^{-3}$	$4*10^{-1}$
A⑨	$5*10^{-7}$	$3*10^{-4}$	$2*10^{-4}$	$4*10^{-4}$	$6*10^{-4}$	$5*10^{-3}$	$7*10^{-3}$	$1*10^{-3}$	**$7*10^{-1}$**
A⑥	$1*10^{-3}$	$4*10^{-4}$	$4*10^{-4}$	$3*10^{-4}$	$4*10^{-3}$	**$5*10^{-1}$**	$3*10^{-4}$	$1*10^{-3}$	$1*10^{-4}$
A⑦	$1*10^{-4}$	$4*10^{-4}$	$2*10^{-4}$	$4*10^{-4}$	$2*10^{-3}$	$4*10^{-3}$	**$8*10^{-1}$**	$1*10^{-3}$	$4*10^{-3}$
B①	**$1*10^{-1}$**	$3*10^{-3}$	$9*10^{-4}$	$8*10^{-4}$	$2*10^{-4}$	$1*10^{-3}$	$1*10^{-4}$	$3*10^{-3}$	$3*10^{-7}$
B⑨	$4*10^{-7}$	$3*10^{-4}$	$2*10^{-4}$	$4*10^{-4}$	$7*10^{-4}$	$6*10^{-3}$	$1*10^{-3}$	$1*10^{-3}$	**$5*10^{-1}$**
B⑥	$1*10^{-3}$	$4*10^{-4}$	$4*10^{-4}$	$2*10^{-4}$	$4*10^{-3}$	**$6*10^{-1}$**	$4*10^{-4}$	$1*10^{-3}$	$1*10^{-4}$
B⑦	$1*10^{-4}$	$3*10^{-4}$	$2*10^{-4}$	$4*10^{-4}$	$2*10^{-3}$	$4*10^{-3}$	**$9*10^{-1}$**	$1*10^{-3}$	$4*10^{-3}$

表 3 「私は先生に怒られる」のの類似度

	①	②	③	④	⑤	⑥	⑦	⑧	⑨
A①	**$9*10^{-1}$**	$3*10^{-3}$	$9*10^{-4}$	$8*10^{-4}$	$3*10^{-4}$	$1*10^{-3}$	$8*10^{-4}$	$2*10^{-3}$	$4*10^{-1}$
A⑧	$3*10^{-4}$	$2*10^{-4}$	$4*10^{-4}$	$6*10^{-4}$	$5*10^{-3}$	$8*10^{-3}$	$3*10^{-3}$	**$8*10^{-1}$**	$8*10^{-4}$
A④	$3*10^{-4}$	$1*10^{-3}$	$1*10^{-4}$	**$8*10^{-1}$**	$1*10^{-3}$	$4*10^{-4}$	$4*10^{-4}$	$3*10^{-4}$	$4*10^{-3}$
B①	**$1*10^{-1}$**	$3*10^{-3}$	$9*10^{-4}$	$8*10^{-4}$	$3*10^{-4}$	$1*10^{-3}$	$7*10^{-4}$	$2*10^{-3}$	$4*10^{-7}$
B⑧	$3*10^{-4}$	$2*10^{-4}$	$4*10^{-4}$	$6*10^{-4}$	$5*10^{-3}$	$8*10^{-3}$	$3*10^{-3}$	**$7*10^{-1}$**	$8*10^{-4}$
B④	$3*10^{-4}$	$2*10^{-3}$	$1*10^{-4}$	**$9*10^{-1}$**	$1*10^{-3}$	$4*10^{-4}$	$4*10^{-4}$	$3*10^{-4}$	$4*10^{-3}$

4. まとめ

ここでは、画像から人物の関節点と手話の手首の動作速度による単語間の分割、DTW を用いて手話判別手法を提案した. 実験では、手首の動作速度の極小点、極大点を用いることで単語部分の分割を行えた. 分割した教師データの CG 手話データと被験者が行った判別させる手話データとのマッチングは DTW を用いた. 今回、使用した判別させる手話は教師データの 9 つの CG 手話単語のデータベース内にある手話単語データに高い類似度を示した. しかし、日常の会話で扱う手話単語数はおよそ 500 語が目安とされている. そのため手話単語数を増やす必要があるがパターンマッチングの処理時間大きくなることが問題となる. 今後の課題としては、お手本の手話のデータベースを増やしたとき、似ている手話をまとめてグループ化し、マッチングの処理時間の対策することである.

文 献

[1] Al Amin Hosain, Panneer Selvam Santhalingam, Parth Pathak, Huzefa Rangwala, Jana Košecká, “Sign Language Recognition Analysis using Multimodal Data”, Proc. DSAA-2019, pp. 1-8, 2019

[2] 喜安千弥, 藤村貞夫, “動画像による連続手話の認識方式”, 計測自動制御学会論文集, Vol.35, No.8, pp.1-6, 1999

[3] Benjamin Johnen, Bernd Kuhlenkoetter, “A Dynamic Time Warping Algorithm for Industrial Robot Motion Analysis”, Proc. CISS-2016, pp. 1-6, 2016

[4] Shogo Okada; Osamu Hasegawa, “Motion recognition based on Dynamic-Time Warping method with Self-Organizing Incremental Neural Network”, Proc. ICPR-2008, pp. 1-4, 2008

[5] J. Huang, W. Zhou, H. Li, and W. Li, “Sign language recognition using 3d convolutional neural networks,” in 2015 IEEE International Conference on Multimedia and Expo (ICME), June 2015, pp. 1–6.

[6] Z. Cao, T. Simon, S-E Wei, Y.Sheikh, “Realtim e Multi-Person 2D Pose Estimation using Part Affinity Fields”, In 2017 IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), No.121, pp.1302-1310, July 2017