1R-03

# 手話認識のための手話動作の音素切り出し手法の検討

内田佳佑 † 澤野弘明 † 土屋 健 †† 小柳恵一 ‡
† 愛知工業大学 †† 公立諏訪東京理科大学 ‡ 早稲田大学

## 1 はじめに

手話の学習方法に，動画や書籍を用いた独学がある．独学では，手話学習者が空いた時間で気軽に学習できるが，学習の際に有識者がいないため，手話学習者は有識者からのフィードバックが得られない．すなわち手話学習者は学習した手話を，正しい動作で使用できているのか，手話動作の客観的な評価が困難である．そこで本研究では，独学でも手話動作の客観的な評価を得るために，機械による手話認識手法に着目する．

手話認識によって手話学習者は期待した認識結果が出力されるかどうかで，手話動作の客観的な評価が得られる．手話認識のためにニューラルネットワークを用いた手法 [1] を松田らが提案している．ニューラルネットワークを用いて高い精度を出すためには大量の学習データが必要となり，さらに様々な単語に対応するには膨大な学習データが必要になる．ニューラルネットワークを用いない手法として，単眼カメラで撮影した映像を用いて画像処理による手話認識手法 [2] を松尾らが提案している．この手法では単眼カメラで撮影した映像を使用しており，三次元の動きの認識が行えず，手を前後に動かす手話に対応できない．

そこで本研究では，Kinect を用いて話者の三次元骨格座標を検出し，手話認識を行う．手話動作判定において，手話単語だけでなく手話文章にも対応していることが望ましい．手話文章は複数の手話単語で構成されており，また手話単語は手話音素 [3] と呼ばれる動作の最小単位の組み合わせにより成り立つ．手話音素の切り出しは重要課題であるため，本稿では手話音素を話者の骨格情報から切り出す手法を検討する．

## 2 提案手法

本節では，Kinect から得られる映像を使用した，手話音素切り出し手法について述べる．音素切り出しには松尾らの画像系列の状態分割手法 [2] を拡張し，二次元座標に対して行う処理を三次元座標に対応させる．使用する骨格を図 1 に示す．図 1 における上の円の位置の頭，下の円の位置の手首を，手話音素の切り出しに使用する．Kinect を用いて検出した，頭と手首の骨格座標の時系列データから手話音素に該当する区間を求める．手が動いている手話動作中は，手首の速度が閾値 $t_v$ pixel 以上の区間が手話音素に該当する．手話音素の組み合わせによっては，音素と音素の間に手が静止せずにつぎの音素へ移行してしまい，複数の音素を一つの音素と判定してしまう場合がある．そこで複数の音素を切り分けるために手首の加速度を使用する．


図 1: 使用する骨格

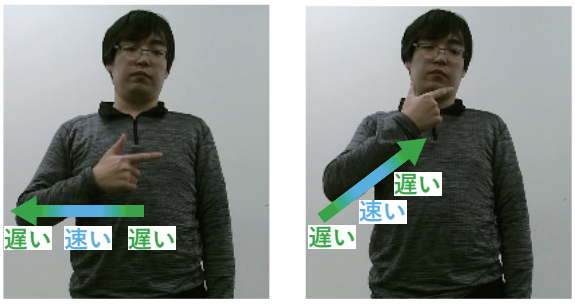
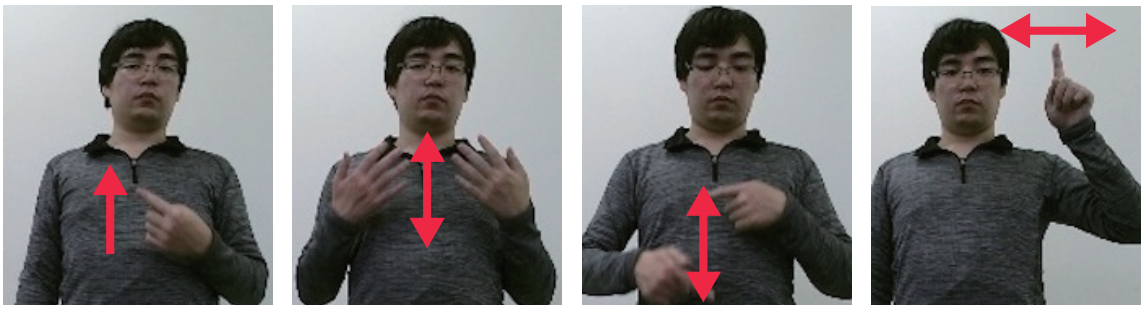



図 2: 手話音素間の速度変化

図 2 のように手話動作は手話音素開始から徐々に加速し，その後減速する．手話音素区間中に手首が減速中で，かつ減速の始点と終点の速度の差が閾値 $t_d$ pixel 以上であれば，減速の終点を手話音素区間の終端とする．ただし，顔付近で行う手話については手の動きが小さいため，手首の速度だけでは手話音素区間を検出できない．頭と手首の距離が閾値 $t_r$ pixel 以下，かつ前フレームと現在フレームの頭と手首の辺の角度の差が閾値 $t_\theta$ rad 以上の区間も手話音素と判定する．

---

A Study on a Phoneme Segmentation Method for Sign Language Recognition
†Keisuke UCHIDA †Hiroaki SAWANO
Aichi Institute of Technology
††Takeshi TSUCHIYA
Suwa University of Science
‡Keiichi KOYANAGI
Waseda University



手話音素: 私　手話音素: 勉強　手話音素: 手話　手話音素: 一人

図 3: 実験に使用した手話文章動作の例 (私は一人で手話を勉強する)

表 1: 実験結果

| | 正検出 | 過検出 | 未検出 |
|---|---|---|---|
| 検出数 | 33 | 9 | 4 |

## 3　実験と考察

1 音素の単語で構成される，手話文章映像 10 映像 (37 単語) に対して提案手法を用いて，手話音素の切り出しを行う精度評価実験を行った．実験に使用したコンピュータの環境は Windows 10 Home 64bit, GeForce GTX960 GDDR5 2.0 GB, Intel Core i7-4790 3.60 GHz, DDR3-1600 16 GB Memory である．実験では Kinect v2 による，フレームレート 30 fps の手話文章映像を使用した．10 フレームごとの移動平均を算出し，平滑化したデータを使用した．使用した閾値はそれぞれ $t_v = 31$, $t_d = 6$, $t_r = 70$, $t_\theta = 0.4$ である．一つの手話音素を複数に分割した場合を過検出，二つ以上の手話音素を一つの手話音素とした場合を未検出とした．分割結果の例を図 4 に示す．図 4 は手の速度の 10 フレームごとの移動平均グラフである．グラフの塗られている箇所が手話音素と判定された区間である．図 4 の例では，すべての手話音素の切り出しに成功している．

実験の結果を表 1 に示す．手話文章映像 10 映像中の手話単語 37 単語に対して手話音素の切り出しを行った結果，正検出 33，過検出 9，未検出 4 という結果が得られた．過検出や未検出の原因として，手の交差や顔と重なる手話動作に関して，Kinect による骨格検出の精度の低さが原因である．過検出や未検出を防ぐために，交差や顔と重なりづらい肘の骨格条法から音素を切り出す手法の検討が今後の課題として挙げられる．

## 4　おわりに

本稿では話者の骨格座標の取得に Kinect を使用し，取得した骨格座標を用いて手話文章映像から，手話動作の最小単位である手話音素の切り出しを行った．提案手法では手話音素間の話者の手の速度や進行方向の変化に着目することで手話音素の切り出しを行った．提案手法を用いて 1 音素の単語で構成される，手話文章映像 10 映像 (37 単語) に対し精度評価実験を行った．実験の結果，正検出 33，過検出 9，未検出 4 という結果が得られた．過検出，未検出の原因は，両手の交差や，顔と手が重なる動作が対象であることが明らかとなった．今後の課題として交差や顔と重なりづらい肘の骨格を使用した音素切り出し手法の検討が挙げられる．

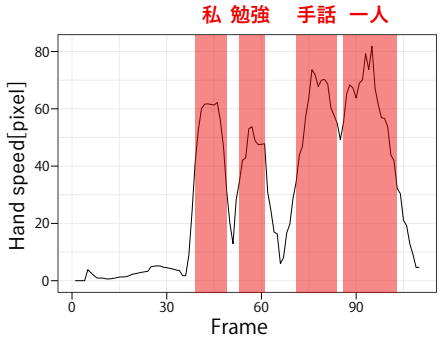



図 4: 分割結果 (手話文章: 私は一人で手話を勉強する)

## 謝辞

本研究は科研費 17K01149 の助成を受けたものです．

## 参考文献

[1] 松田啓佑, 飯塚博幸, 山本雅人: “手話分類における RCNN モデルの性能評価と内部状態解析”, 人工知能学会全国大会論文集, Vol. JSAI2018 (2018)

[2] 松尾直志, 山田寛, 白井良明, 島田伸敬: “HMM を利用した画像処理による手話単語の認識のための特徴抽出および状態分割”, ヒューマンインタフェース学会論文誌, Vol. 15, No. 1, pp. 86–94 (2013)

[3] 米川明彦: “手話言語の記述的研究”, 明治書院 (1984)