M-012

単語学習させた LSTM を用いた手話の短文動作識別に関する基礎検討
Basic study on short sentence motion classification of sign language using word-learned LSTM

栗原 泰晴† 黒尾 陽太† 若尾 吏†
Yasuharu Kurihara Youta Kuroo Tsukasa Wakao
川喜田 佑介† 西村 広光† 田中 博† 三次 仁‡
Yuusuke Kawakita Hiromitsu Nishimura Hiroshi Tanaka Jin Mitsugi

1. はじめに

既に人工知能の技術によって音声認識や自動翻訳などが広く普及しつつある．しかし，聴覚障碍者の方が使用している手話に関しては，健聴者との間では筆談などに頼ることがほとんどであり，コミュニケーションの上で大きな障壁となっている．手話の自動翻訳という観点では，単語レベルの手話動作の識別[1]から，現在は文章を前提とした連続手話単語の認識に研究対象が移行しているようである[2].

動きの検出と推定に関しては，加速度センサを身体に装着しその動きを検出し，機械学習を適用してその動作や作業を推定する研究は従来より行われてきた[3]．しかし，センサデバイスの寸法等の制約から多数の装着は現実的ではなく，複雑かつ細かな動きの検出は難しかった．しかし，センサ信号の高速かつ多チャネル受信をバッテリレスかつワイヤレスで可能とするバックスキャッタ通信の研究開発が進められ[4]，加速度センサのモーションキャプチャへの適用が現実的になりつつある．

一方，機械学習という観点では，深層学習が広い領域で適用され多数の実績が報告されている．この中で時系列データに関する予測や識別問題に LSTM(Long Short-Term Memory)が広く用いられ，手話単語の動作識別にも適用されている[5]．文献[2]は手話認識に関する先導的な論文と考えられるが，その中では個別の手話単語動作を識別するための DTW(Dynamic Time Warping)とそれらを連続した手話単語として認識するための隠れマルコフモデルが適用されている．

筆者らは文献[2]とは異なる手法を提案し，その可能性を検討する．手話を構成する各単語を LSTM モデルに学習させる．そのモデルに対して学習した単語を含む 3 単語からなる短文の連続した手話動作を与え，その動作内の各単語の識別性能を調べた．本報告では提案手法の手話文章翻訳への適用の可能性とともに，認識精度向上のための今後の課題を明らかにした結果を述べる．

2. 対象手話動作と手話動作データ取得

2.1 対象手話動作

本検討では，株式会社ケイ・シー・シーが開発，販売している手話学習者向けの動画辞典である SmartDeaf[6]に含まれる手話動作の中から日常生活において身近に使用する 11 単語の手話動作を選択した．そして，選択した単語動作を含む 3 単語で構成される連続した短文動作 5 つを識別対象とした．対象とした単語・短文手話動作を表 1 に示す．これらの単語・短文手話動作はともに加速度センサで取得することとした．

表 1　使用する手話短文・短文に含まれる単語

短文	1. 私/走る/汗だく	2. 私/ダイエット/中	3. 私/眼鏡/買う	
	4. 私/眼鏡/忘れる	5. 息子/顔色/悪い		
単語	1. 汗だく	2. 最中(中)	3. ダイエット	4. 走る
	5. 顔色	6. 買う	7. 眼鏡	8. 息子
	9. 悪い	10. 忘れる	11. 私	

2.2 手話動作データ取得

動作データを取得するための同期バックスキャッタセンサを使用した構成を図 1 に示す．質問器から UHF 帯の電波を発射し，バックスキャッタセンサ（本構成では加速度センサ）に電波給電するとともにセンサデータで変調した電波を受信する．今回は手話者の左右の手首，肘の 4 か所に加速度センサを取り付けた．それぞれ 4 つの無線チャネルからの受信である．ここで，受信データに対して外れ値除去，損失パケット補填，リサンプリングのポスト処理を行いデータの信頼性を確保している[7]．なお，用いた加速度センサは Analog Devices ADXL 362 であり，サンプリング間隔は 10ms である．

手話動作のデータ取得は，株式会社ケイ・シー・シー Smart Deaf の開発に携わり手話を使用している方の監督・指導のもと実施した．手話者は Smart Deaf を表示したモニタで手話動作を確認し監督者から手話動作に問題がないと

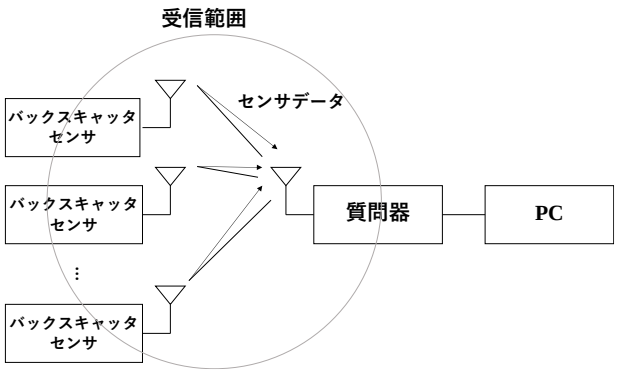

図 1　同期バックスキャッタセンサを使用したデータ取得の構成[4]

†神奈川工科大学　Kanagawa Institute of Technology
‡慶應義塾大学　Keio University

図 2　手話データ取得時の模様[8]

の判断後，両手を下した状態から動作を開始した．データ数の確保と効率的なデータ取得のために短文・単語手話動作ともに，1 短文・単語を 20 動作連続で行った．このとき，データとして 1 動作を切り出しやすくするため各動作間に約 2 秒の静止状態を設けた．また，カメラ画像を加速度データとの比較対象として同時に取得した．手話動作データ取得時の模様を図 2 に示す．

3. 使用データ

取得した加速度データの一例として，同じ手話者の右肘の箇所の単語“走る”のデータと，短文“私 / 走る / 汗だく”の動作データを図 3, 4 に示す．単語“走る”の動作データが短文“私 / 走る / 汗だく”の動作データの 1s~2s の動作部分に同じ動きとして含まれていることが確認できる．

本検討に当たり取得した手話動作データのサンプル総数は単語の場合は，手話者 4 名，手話動作 11 単語，1 動作のサンプル数 20，合計 880 サンプルである．短文の場合は，手話者 4 名，手話動作 5 単語，1 動作のサンプル数 12，合計 240 サンプルである．本報告では，学習する単語数による識別精度への影響も確認するため，7 単語として合計 420 サンプルを学習，11 単語では合計 660 サンプルを学習させ

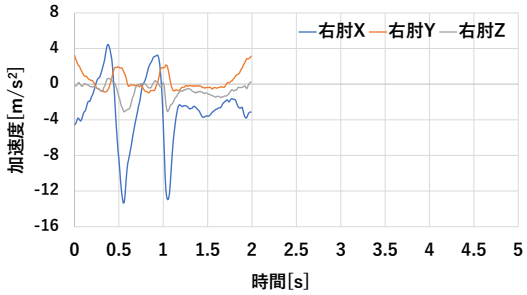

図 3　“走る”動作データ(右肘)

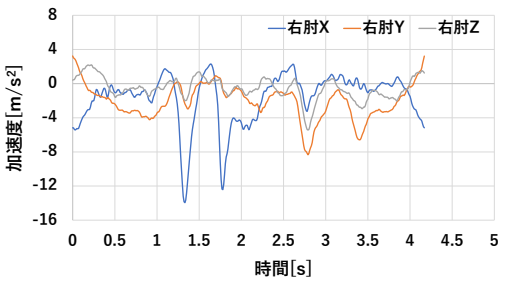

図 4　“私 / 走る / 汗だく”動作データ(右肘)

表 2　取得データセット

	手話者	手話動作	1動作サンプル数 (1手話者)	合計サンプル数
単語	4	11	20	880(4*11*20)
短文	4	5	12	240(4*5*12)

表 3　使用データセット

	手話者	手話動作	1動作サンプル数 (1手話者)	合計サンプル数
学習データ(1)	3	7	20	420 (3*7*20)
学習データ(2)	3	11	20	660 (3*11*20)
評価データ	1	5	1	5 (1*5*1)

た識別モデルを作成することにした．この 7，11 単語学習した識別モデルに対し，評価データは今回は手法の可能性の確認として合計 5 サンプルの短文手話動作データを使用した．取得した単語・短文手話動作データセットを表 2，本検討で使用したデータセットを表 3 に示す．なお，学習で用いるデータと評価で用いるデータの手話者は異なる人物としている．

4. 識別モデルの作成

本検討では，時系列データの予測や識別に優れた LSTM を使用し識別モデルを作成した．隠れユニット数 200 の LSTM 層とし，最終段にサイズが 7 あるいは 11 の全結合層を含めることによって分類のクラス数を指定し，その後にソフトマックス層と分類層を配置した．入力層のサイズ（特徴の数）は，3 人×3 次元の 9 である．今回は MATLAB で実装した．なお，各パラメータを表 4 のように設定した．

学習で用いた環境は，OS：Windows10，CPU：Intel(R) Core(TM) i5-4460 CPU@ 3.2GHz 24GB メモリ，GPU：NVIDIA GeForce GTX 1050 2GB メモリで実施し，学習には 20 分程度要した．学習時の Accuracy 曲線は，7 単語，11 単語の場合でそれぞれ 70%，50%程度で収束した．識別モデルの作成という観点からは課題があると思われる．

表 4　学習パラメータ

パラメータ	設定
隠れユニット	200
ソルバー	adam
最大エポック数	60
学習率	0.001

5. 識別実験の結果と考察

3 節で述べた学習データ(1), (2)を用いて作成した識別モデルを使用し，短文動作データを入力してそのデータ要素ごとに出力される識別結果をプロットした結果の一例を図 5, 6, 7 に示す．“私 / 走る / 汗だく”，“私 / ダイエット / 中”，“私 / 眼鏡 / 買う”の 3 種類の短文に対する 7，11 単語で学習させた識別モデルを用いた結果である．“Activity”のラベルに実際の動作内に含まれる単語を四角形の赤枠で示し，グラフ上にはビデオ映像からの目視と短文動作の加速度のグ

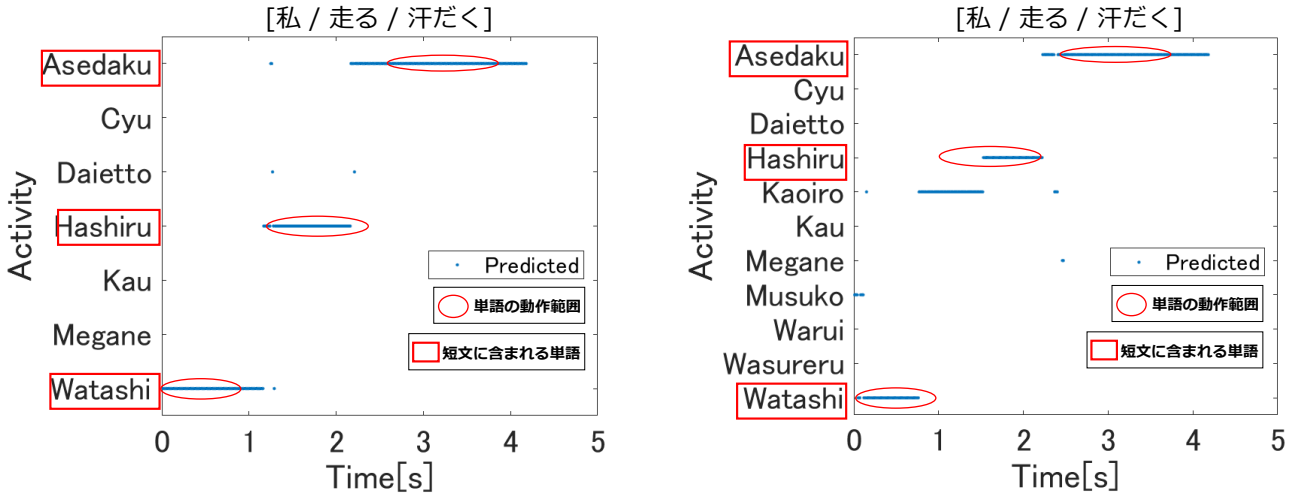

図 5　識別結果 1(私 / 走る / 汗だく)

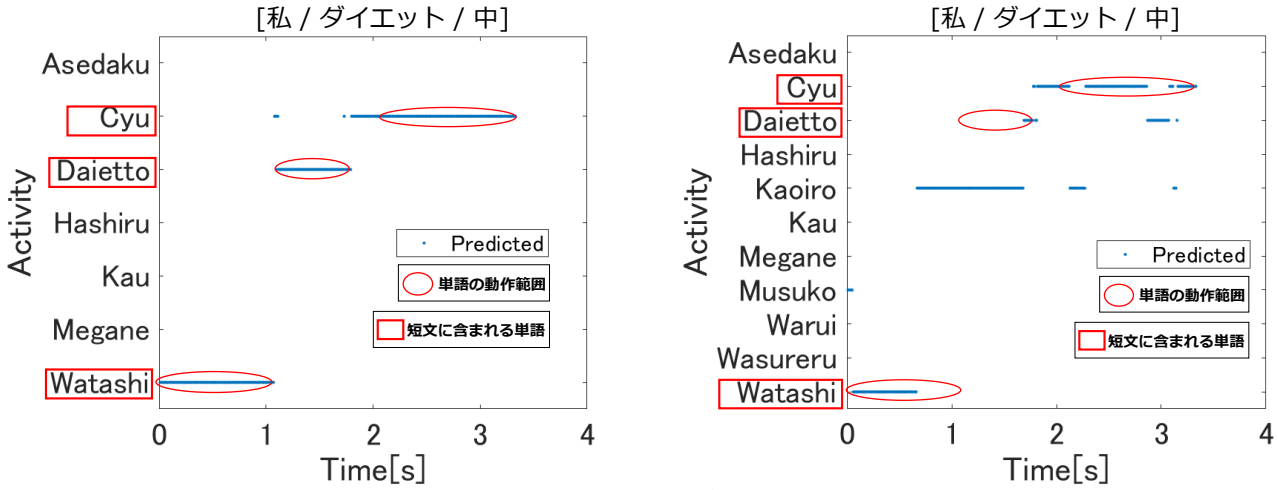

図 6　識別結果 2(私 / ダイエット / 中)

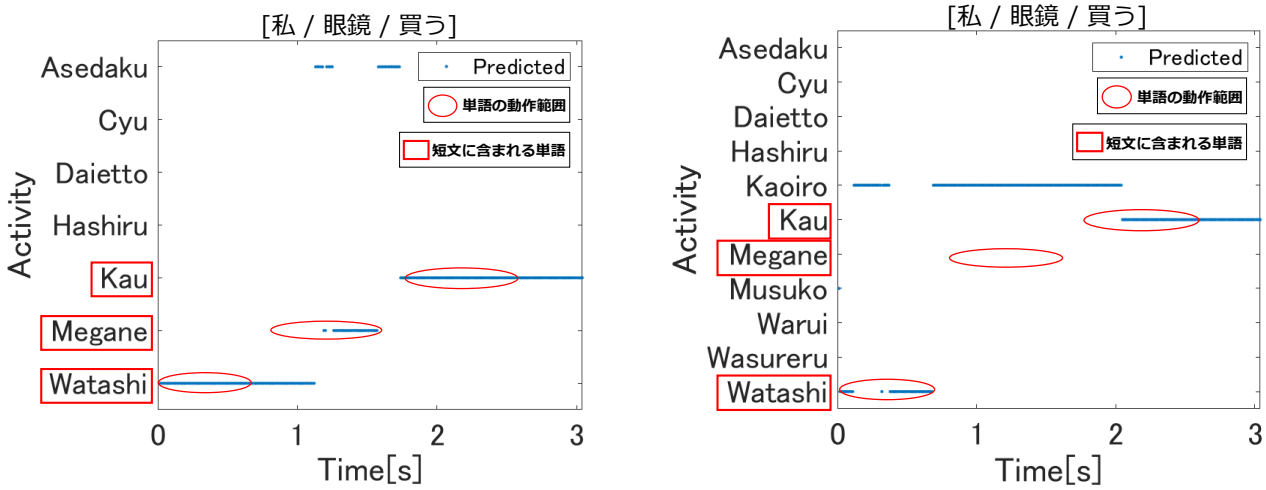

図 7　識別結果 3(私 / 眼鏡 / 買う)

ラフから確認した動作内に含まれる各単語のおおよその動作区間を楕円の赤枠で示している．

実験結果より，学習単語が 7 種である学習データ(1)を用いた識別モデルでは短文動作内の単語をおおよそ識別できていることが分かる．一方，11 種である学習データ(2)を用いた識別モデルでは学習データ(1)の場合と比べ，誤識別している部分が増加したことが分かる．特に，顕著に誤識別が見られるのは短文動作内の 2 つ目の単語の動作であり，

図 8　短文手話動作動画の一部

主に単語“顔色”に誤識別されている．この誤識別について直接の原因は断定できないが，1つ目の単語動作，3つ目の単語動作と接続する動きであり単語動作ではない “遷移”部分 の動きの影響が大きくなった可能性がある．手話者の短文“私 / 走る / 汗だく”の加速度データ取得時の動画の一部を図 8 に示す.

6. まとめと今後の課題

本稿では，LSTM を用いて短文動作を構成する単語動作の加速度センサデータを用いて識別モデルを作成した．識別モデルを用いて手話の短文動作内に含まれる単語の識別を行った．7 単語学習の識別モデルでは短文動作内の単語を動作の継続時間を考慮すると誤りなく識別できた．11 単語学習の識別モデルでは短文動作内の 2 単語目にあたる部分が単語“顔色”という特定の単語に識別が偏る傾向によって識別精度が低下したものの一定の精度は確保できた．この結果から手話動作の識別に関して，手話単語を学習させることで文章の手話動作の識別に展開できる可能性を示した.

今後は識別精度向上のため学習データを含め，識別モデルの最適化を図っていく．並行して，識別の際の各動作単語の尤度を考慮した単語判定の導入や，手話単語動作の時間長および手話の語順など文法的観点を考慮した識別を行っていく予定である.

謝辞

手話動作をご指導いただいた株式会社ケイ・シー・シーの関係各位に感謝いたします．本研究開発は総務省の「電波資源拡大のための研究開発(JSJ000254)」によって実施した成果を含みます.

参考文献

[1] 小澤 辰典他, “光学式カメラを用いた色領域の抽出に基づく手話認識手法, “画像電子学会論文誌, Vol.49, No.1, pp.12-24, 2020.
[2] N. Takayama et al., “Weakly-Supervised Learning for Continuous Sign Language Word Recognition Using DTW-Based Forced Alignment and Isolated Word HMM Adjustment,” IIEEJ Transactions on Image Electronics and Visual Computing, Vol.7, No.2, pp.88-96, 2019.
[3] 市川 峻大他, “判別分析手法を用いた研究室内の行動推定法の提案と評価実験,” HCG シンポジウム 2010, B4-2, pp.181-186, 2010.
[4] J. Mitsugi, et al., “Perfectly Synchronized Streaming from Multiple Digitally Modulated Backscatter Sensor Tags,” IEEE J. RFID, vol.3, no.3, pp.149-156, Sept. 2019.
[5] K. Kawaguchi et al., “Basic Investigation of Sign Language Motion Classification by Feature Extraction using Pre-trained Network Models,” IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 4 pages, 2019.
[6] 株式会社ケイ・シー・シー，手話学習者向けスマートデバイス動画辞典, SmartDeaf, http://www.smartdeaf.com/.
[7] J. Mitsugi, et al., “Wireless Modal Analysis With Backscatter Sensors Subjected to Clock Instability,” IEEE Journal of RFID, vol. 5, no. 3, pp. 269-277, Sept. 2021.
[8] 佐藤 辰也他, “同期バックスキャッタセンサの手話動作識別への適用,” 2022 信学総大, B-15-54, p.492, 2022.