CNN-LSTMを用いた手話認識システムの開発

Japanese Sign Language Recognition System Using CNN-LSTM

土井ゆりか[1*]　八木拓真[2]　水口智仁[3]
Yurika Doi[1]　Takuma Yagi[2]　Tomohito Minakuchi[3]

東京大学工学部機械情報工学科[1]
[1] The University of Tokyo, Department of Mechanoinformatics
東京工業大学工学部情報工学科[2]
[2] Tokyo Institute of Technology, Department of Computer Science
慶應義塾大学医学部医学科[3]
[3] Keio University, School of Medicine

Abstract: We propose a Japanese sign language recognition system combining Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM). Existing research has had two problems. First, it has assumed that sign language could be recognized by extracting hand/arm positions and directions as features although non-manual signals play an important role in sign language. Second, it has divided temporal structure by using velocity of the hands or the movement section of the hands. However, this assumption might have left out the complex temporal structure of sign language. In this research, we created a dataset of movies of the upper bodies of sign language signers by using Kinect version2. In order to extract the effective features that include non-manual signals, we put the visible images and depth images of the dataset into CNN by frames. Then the extracted features were put into LSTM frame by frame to capture the complex temporal structure of sign language. We trained our whole network by using the backpropagation algorithm. Comparing this CNN-LSTM model to control models, we suggest that this model is more effective for sign language recognition.

1　はじめに

ろう者と健聴者のコミュニケーション支援を目的とした手話認識研究は, これまでに, 様々な手法が提案されてきたが、未だ広く実用化されている手法は存在しない. 既存の手話認識手法の多くは, 手話映像から手指の形や角度を特徴量として取り出して, それらを基に分類を行っている. しかし, 実際の手話では手指の形だけではなく, 表情, 頷き, 目の動きなどの非手指信号が重要な役割を果たしている. 既存の手話認識研究ではこれらの非手指信号が考慮されていないことが, 認識精度向上の妨げとなっている可能性がある. そこで筆者らは, Kinect を用いて手話を撮影し, 指や手の位置情報だけでなく動画全体の深度画像と可視画像を入力として, 特徴量抽出を畳み込みニューラルネットワーク (Convolutional Neural Network, 以下 CNN) によって行うことで, 非手指動作も含めた特徴量を抽出できると考えた. また, 時系列構造の学習については, 既存研究では研究者が恣意的にフレームを分割を行っていることが多かった. しかし, 恣意的なフレーム分割によって動作のつなぎや手話の複雑な時系列構造が抜け落ちている可能性があった. そこで, 筆者らは、Long-Short Term Memory(以下 LSTM) を用いて時系列相関の学習を行うことによって, 中長期的な時系列を考慮した単語予測が可能となると予想し, CNN-LSTM を採用した手話認識システムを構築した上で実際の手話動画を用いて認識実験を行った. 結果, 本稿で提案する CNN-LSTM モデルが手話認識に有用であることが示された.

2　関連研究

既存の手話研究には, Enable Talk[1] のように被験者の手指に直接センサを取り付け, 特徴点の動きを計測する方法と, カメラで動画像を取得する方法がある. 1990 年代から 2000 年代の動画像を利用した手話認識では, 手話者に特別な色の手袋を装着してもらう手法 [2] や

*@E-mail: 5986703900@mail.ecc.u-tokyo.ac.jp

赤外線カメラを用いた温度画像を用いた手法 [3] によって手指の特徴量抽出を行っていたが, これらは特殊な設備や装具を必要とし，限られた状況でしか使用できないという問題点がある．また, 特殊な装置を使わず画像処理によって特徴量抽出を行った研究では, オプティカルフローを利用した手法 [4] や動きベクトルを用いた手法 [5] などがあるが, これらの手法は大局的な手や腕の動きを検出することはできるものの, 手の形状については十分考慮されていない. 川東ら [6]，柳ら [7] は手の色情報に基づいて手領域を分離し, 手形状特徴を求めている．しかし, これらの手法は手領域が他の肌色領域と重複する場合は正確な手領域を得られず, 手形状の特徴量抽出が困難となる．これらの問題点に加えて以上の既存研究には, 手話を手指の形・動きのみの表現と捉え, 手領域の抽出や手指の特徴量抽出のみを問題として設定したため, 非手指信号が考慮されていないという弱点がある．非手指信号とは表情, 頷き, 目の動きなどのことで, 手話においてこれらの非手指信号は意味の伝達に重要な役割を果たしている [8][9]. また, 手話の時系列についても, 既存研究においては白井ら [10] や松尾ら [11] のように, 動素を定める, もしくは速度や運動区間によって時系列を分割すると言ったように, 研究者が恣意的にフレームを分割しており, 一つの動作から次の動作に移るときの動作のつなぎが考慮されていない可能性があった. そこで筆者らは, 畳み込みニューラルネットワーク (Convolutional Neural Network, 以下 CNN) と LSTM を組み合わせた機械学習器を使用することで, 非手指信号を考慮した手話認識を行うことができると考えた．特に特徴量抽出に関しては, Kinect を用いて手話を撮影し, 指や手の位置情報だけでなく動画全体の深度画像と可視画像を入力として CNN に学習させることで, 非手指動作も含めた特徴量を自動的に抽出できると考えた．Kinect と CNN を用いた手話認識は Pigou[12] らによって行われている. Pigou らは, 上半身全体と手領域のみの時系列動画像をそれぞれ 3 次元配列として CNN に入力し, 20 種類のイタリア手話を精度 91.7%で判別できたと報告している．また, 時系列の認識に関して, 筆者らは、恣意的にフレームを分割する代わりに, LSTM に動画像を 1 フレームずつ入力することによって手話動作の時間的な構造も考慮した手話認識が可能となると考えた．CNN と LSTM の組み合わせは Oriol[13] らによって画像からのキャプション生成タスクにおける有用性が示されている．そこで筆者らは CNN-LSTM を採用した手話認識システムを構築した上で実際の手話動画を用いて認識実験を行った.

3 データセット

3.1 データセット作成

2015 年 7 月時点では, 手話動画の公開データセットがなかったため, データセットを自ら作成した．言語や地域によって多様な手話が世界中で使用されているが, 日本で主に使われている手話には, 日本手話, 日本語対応手話, 中間手話 (日本手話と日本語対応手話の混在) が存在する．なかでも今回は, 日本手話を対象としてデータセットを作成した．手話者 7 人 (内ろう者 2 人、健聴者 5 人) を, 35 単語 20 例文を原則として 4 回ずつ撮影した．単語セットは手話技能検定 [14]5 級程度の簡単な語彙のなかでも, 個人や地域によらず手指の動作が統一されているものを選択した．文章セットも, 手話技能検定 5 級程度の簡単な語彙を組み合わせた文を選択した．撮影には深度情報を撮影できる Microsoft 社の Kinect Version2 を用いた．手話者は白い壁を背景として椅子に座り、膝から上が映るように撮影した.

3.2 訓練データ構成

収集した映像について, RGB 映像はグレースケール化し, 132x132 に縮小し, 深度画像も 132x132 に縮小した．フレームレートは 30fps とした．Kinect センサーの不調によってフレーム間の時間の遅延が発生した箇所については予め閾値を設定してカットした．最終的に 815 動画の各フレームに対して 37 種類の教師ラベル (35 単語ラベルと開始/終了記号) のいずれか一つを付与し, これらを訓練データとして使用した．また, 文章データについての実験は今回は行わなかったが, 教師ラベルを付与したものについては訓練データとして使用し, 計 97 の教師ラベルを付与した.

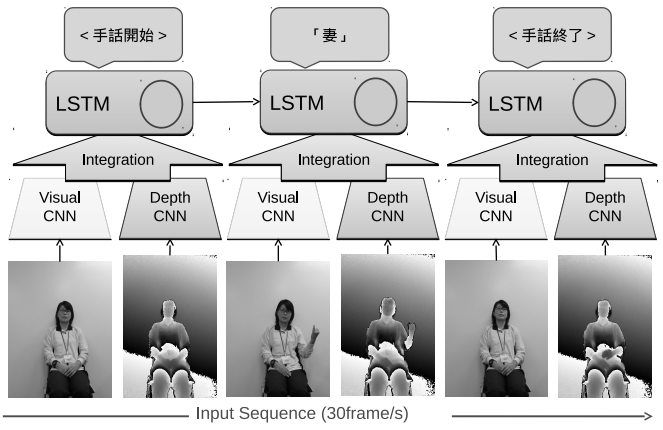

図 1: 提案モデル概要

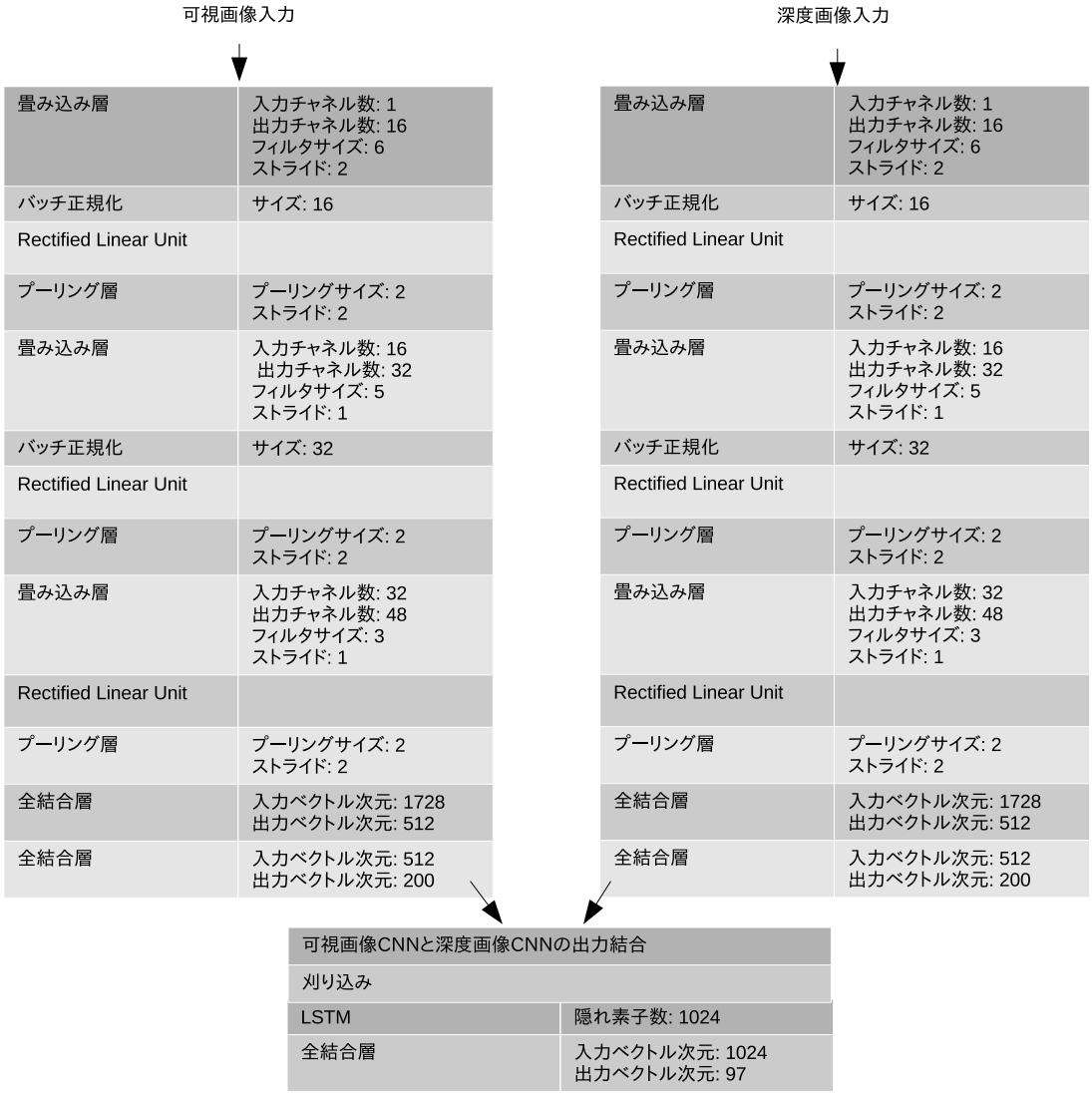

図 2: 提案モデルの構造

4 提案モデル概要

図 1 に提案モデルの概観を示した. (1) 各フレーム毎に可視画像, 深度画像をそれぞれの CNN を通すことによって高次特徴を抽出し, (2) CNN の Full Connection 層を通過した後の可視特徴ベクトル, 深度特徴ベクトルを連結して統合した後, (3) LSTM によるフレーム毎の手話予測を行った. 訓練データ構成で述べたように, 各フレーム毎に教師ラベルを付けたため, CNN-LSTM 一括で誤差逆伝播法による教師あり学習によって学習した. また, 実装には python の機械学習フレームワークである chainer[15] を使用した. 提案モデルの構造は図 2 のとおりである.

5 実験

実験は 2 種類行った.

[手話者依存実験] 訓練データ 734 動画, テストデータ 196 動画, ラベル数：97 ラベル手話者依存実験では訓練データとして, 1 単語につき, 7 人の手話者の 4 回の試行のうち 1,2,3 回目の試行を使用し, 4 回目の試行をテストデータとして使用した.

[手話者非依存実験] 訓練データ 766 動画, テストデータ 164 動画, ラベル数：97 ラベル手話者非依存実験では訓練データとして, 6 人の手話の全試行を使用し, 7 人目の手話者の全試行をテストデータとして使用した.

表 1: 各モデルの単独単語認識率 (単位: %)

	$CNN-LSTM$	CNN のみ	可視画像のみ	深度画像のみ
手話者依存実験	86.4	82.2	63.9	75.1
手話者非依存実験	26.3	-	-	-

5.1 実装モデル

提案する CNN-LSTM モデルに加えて, 比較として可視画像と深度画像を入力とする CNN のみのモデル, 可視画像のみを入力とする CNN-LSTM のモデル, 深度画像のみを入力とする CNN-LSTM のモデルを作成した.

5.2 実験結果

各単語につき, 毎フレームごとに, 単語ラベルを出力し, 5 フレーム以上同一の予測が得られるラベルのうち, 最も遅く検知されたものを単語認識結果として出力し, 正解率 (Accuracy) を算出した. これを単独単語認識率と名付けた. 実験結果は表 1 のようになった. 手話者依存実験では CNN-LSTM モデルは単独単語認識率 86.4%であった. CNN のみを用いたモデルおよび可視画像/深さ画像のみを入力した CNN-LSTM モデルの単独単語認識率はそれぞれ 82.2%, 63.9%, 75.1%であった. 手話者非依存実験では CNN-LSTM モデルのみ実験を行い, 単独単語認識率は 26.3%であった.

6 考察

CNN-LSTM モデルは, CNN のみのモデルに比べて単独単語認識率が高かったことから, LSTM の導入は今回の手話認識システムの精度向上に重要であることが示された. また, 可視画像と深度画像の両方を入力した学習器は, 可視画像のみを入力した学習器, 深度画像のみを入力した学習器に対して認識率が高かった. よって, 今回の手話認識システムにおいて, 可視画像と深度画像を両方用いたことの有用性が示された. 訓練データに入っていない手話者の手話を認識した手話者非依存実験の単独単語認識率は, 手話者依存実験の単独単語認識率よりも低い. これは, 手話者非依存実験では 1 単語当たり, 訓練データが約 20 動画で, 手話の個人差を吸収するにはデータ数が不十分であったからと考えられる. より多数の手話者からの訓練データを学習器に入力することで認識率を改善することができると考えられる.

7 今後の展望

最終的な出力方法については, 本研究では 1 フレームにつき 1 つの教師ラベルを付与し, 出力も 1 フレームごとに行ったが, 複数フレームからなる一連の動作に対して 1 つのラベルを出力する方法も可能である. 後者の方法は, 一連の時系列変化から特徴量を抽出する LSTM の特性をより有効に利用できる可能性がある, その他の認識率を向上させる戦略としては, より明示的に非手指信号を特徴量として抽出するために, 重要な非手指信号の一つである表情を抽出して CNN への入力に加えることや, 大局的な動作と手や顔の微細な動作の双方を考慮に入れるために, 上半身全体の画像, 手領域のみの画像, 顔領域のみの画像をそれぞれ入力することなどが考えられる.

謝辞

本研究のデータセット作成に協力してくださった東京大学バリアフリー支援室の職員の方々, 東京大学施設部環境課の職員の方々に感謝いたします. また, 本研究をご支援いただいた株式会社ドワンゴの大垣慶介氏, 宮脇康介氏, 全脳アーキテクチャイニシアチブの方々, そして本稿の執筆をご指導いただきました電気通信大学大学院情報システム学研究科の栗原聡教授に感謝いたします.

参考文献

[1] http://enabletalk.com/

[2] L. Yoshino et al.: Recognition of Japanese sign language from image sequence using color combination, Proc. 3rd Int. Conf. Image Processing, pp. 16–19 (1996)

[3] 吉冨康成, 永山しづえ, 杉山雅祥: 温度画像処理による手軌跡の抽出と手話認識, 信学技報, SP2002-112, Vol. 102, No. 418, pp. 21–26 (2002)

[4] 岡澤裕二, 堀内靖雄, 市川熹: オプティカルフローによる手話の大局的動作の認識について, 信学技報, PRMU2002-77, Vol. 102, No. 317, pp. 39–44 (2002)

[5] 李昌宏, 中薗薫, 長嶋祐二, 張鴻徳: 動きベクトルを用いた手話単語分類, 信学技報, WIT2003-62, Vol. 103, No. 746, pp. 65–70 (2004)

[6] 川東香菜, 白井良明, 島田伸敬, 三浦純: 手話の HMM 作成のための状態分割, 信学技報, WIT2005-21, Vol. 105, No. 67, pp. 55–60 (2005)

[7] 柳哲, 柳生雄午, 徳田恵一, 北村正: 手の動作と形状を用いた HMM 手話認識, 電子情報通信学会総合大会講演論文集, D-12-119, pp. 285 (2004)

[8] 市川熹, 長嶋祐二, 寺内美奈: 手話における ”顔 ” のはたらき, 情報処理学会コンピュータビジョンとイメージメディア研究会資料, 2005-CVIM-148, pp. 66–72 (2004)

[9] 土肥修, 堀内靖雄, 市川熹, 長嶋祐二, 寺内美奈: 手話対話における頷きの影響に関する実験的検討, 電子情報通信学会福祉情報工学研究会資料, WIT2002-21, pp. 45–50 (2002)

[10] 白井良明 他: 手の動きと形を用いた動作分割による手話認識, 第 13 回画像の認識・理解シンポジウム, (2010)

[11] 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, ヒューマンインターフェース学会論文, Vol. 15, No. 1, pp. 85–94 (2013)

[12] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen: Sign language recognition using convolutional neural networks, In ECCVW (2014)

[13] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan: Show and tell: A neural image caption generator, arXiv:1411.4555 (2014)

[14] http://www.shuwaken.org/

[15] http://chainer.org/