2H3-OS-35a-4in1

深層学習を用いた会話中の人物頭部ジェスチャ認識

Automatic Recognition of Human Head Gestures in Conversations using Deep Learning

大薮 将士 *1*2 Masashi Oyabu　Debora Zrinscak*1 Debora Zrinscak　大塚 和弘 *1 Kazuhiro Otsuka

*1NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories　*2長岡技術科学大学 Nagaoka Institute of Technology

This report addresses a deep learning approach to the automatic recognition of spontaneous head gestures occurred in multiparty face-to-face conversations. To that purpose, a deep convolutional neural network (CNN) is designed and trained to discern the presence/absence of head nodding at each time step, from the time series of 3 degree-of-freedom head pose angles, measured with attached sensors. Experiments confirmed that the proposed CNN significantly outperformed a previous method that uses wavelet features and support vector machines (SVM).

1. はじめに

対面会話はコミュニケーションの基本的な形態であり，その中で生じる会話参加者の非言語行動の自動認識は，社会的信号処理の重要な課題である．そのような非言語行動の中でも，頷きや首振りと言った頭部運動や頭部ジェスチャは，発話のリズム取り，傾聴の合図，同意・非同意など態度の表出，話者交替の制御など，会話を進行する上で多様かつ重要な機能を担っている [Maynard 87]．過去，頭部ジェスチャの自動認識のため，頭部姿勢の時系列信号の特徴抽出と教師あり機械学習を組み合わせた方法が提案されているが（例えば [Otsuka 07] など），会話中に自発的に生じる頭部ジェスチャはその強度や周期性，個人性など多様であり，高精度な自動認識は依然として困難である．一方，近年，画像認識の分野では，畳み込みニューラルネット (convolutional neural netowork, 以下 CNN）が，一般物体認識のタスクに極めて有効であることが知られている．また，最近では，時系列信号からの人物動作認識のタスクとして，CNN を意図的な屋内日常動作（ドアの開閉など）の認識に適用した事例も報告されている [Yang 15]．本稿では，CNN を用いた時系列認識手法を会話中の頭部ジェスチャ認識に適用し，その有効性を検証したので，その結果を報告する．

2. 提案モデル

本稿では，会話中の人物頭部ジェスチャとして，特に「頷き」に着目して，各時刻においてその有無を識別するために畳み込みニューラルネット (CNN) を用いる．CNN は，順方向ニューラルネットの一種で，畳み込み層，プーリング層，正規化層の繰り返しの後，全結合層を経て，最終的にクラス識別の結果が出力される．本稿の提案法では，入力として，3 自由度の頭部姿勢角の角速度の時系列が与えられる．頭部姿勢角の各成分は azimuth, elevation, roll と呼ばれ，azimuth は首振り方向，elevation は頷き方向，roll は傾げ方向の回転にそれぞれ対応する．実際には認識対象となる時刻を中心とした前後一定の窓幅分の時系列が CNN に入力される．出力としては，頷きの有無が 2 クラス分類の結果として出力される．図 1 に具体的な CNN の構造を図示する．層数は Yang らの CNN[Yang 15] を踏襲しているが，各層のノード数は，今回の入力信号の次元数や性質に応じて調整された．最初の畳み込み層 conv1 では，各々の頭部姿勢角の角速度成分について独立に畳み込みフィルタが適用され，その後，活性化関数（正規化線形関数 ReLU）を経て出力が得られる．次のプーリング層 pool1 では，最大値プーリングによる局所最大値の検出とサブサンプリングによる時間方向のデータの縮減が行われ，その後，局所コントラスト正規化が成される (norm1)．その後，図 1 の様に畳み込み層，プーリング層，正規化層が繰り返された後，全結合層 fc4 では，前段 (norm3) の出力が全結合された出力ノードへと至り，前段まで別々に扱われてきた azimuth, elevation, roll の情報が統合される．最後にその結果がジェスチャの有無に相当する 2 ノードの出力層 fc5 へ全結合され，ソフトマックス活性化関数の適用を受け，最終的な識別結果が出力される．

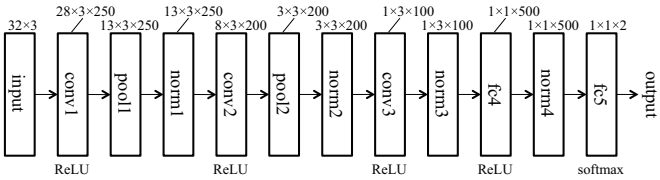

図 1: 畳み込みニューラルネットの構造．conv*: 畳み込み層，pool*: プーリング層，norm*：正規化層，fc*: 全結合層を示す．各層上部の $X \times Y \times Z$ はノード数を表す．X は入力，Y はチャンネル数（azimuth, elevation, roll に相当），Z は出力の次元数にそれぞれ対応．

3. 実験

提案モデルの有効性を検証するため，2004 年に著者らが収録した女性 4 人による対面会話データを用いた．2 グループ (G1, G2 と表記) による各 2 セッションの会話を対象とした (G1-C1, G1-C2, G2-C1, G2-C2 と表記)．データ長はそれぞれ 9.6 分，5.5 分，5.6 分，5.9 分であった．会話のタスクはグループ内での合意形成であった．人物の頭部姿勢は，頭部にバンドで固定された磁気式センサ Polhemus Fastrak により計測された（30Hz）．会話に参加していない女性 1 名によって，各時間フレーム毎に頭部ジェスチャの種別のラベル（頷き，首振

連絡先: 大塚和弘，NTT コミュニケーション科学基礎研究所，〒 243-0198 神奈川県厚木市森の里若宮 3-1，TEL: 046-240-3639，Fax: 046-240-3145，E-mail: otsuka.kazuhiro@lab.ntt.co.jp

表 1: グループ特化モデルの認識精度（同一グループの別会話間での交差検定）

手法	適合率	再現率	F 値
SVM	0.782	0.518	0.623
CNN	0.765	0.728	0.746

表 2: 汎用モデルの認識精度（グループ間での交差検定）

手法	適合率	再現率	F 値
SVM	0.702	0.408	0.516
CNN	0.785	0.702	0.741

り，傾げ）が付与され，今回はその内，頷きかそれ以外かの二値のデータを CNN の教師データとして用いた．

提案法の実装は，Yang による公開 Matlab コード [Yang 16] をベースとした．CNN の各層のノード数は図 1 の通りである．学習に用いたパラメータは，窓幅 32，畳み込み層 (conv1, 2, 3) のフィルタ長 5, 6, 3，バッチサイズ 12，エポック数 4，重み減衰率 5×10^{-4}，慣性係数 0.9 とした．学習率は，0.01, 0.001 とした（エポック数の増加につれ減少）．

比較対象とする従来法として，頭部姿勢時系列の離散ウェーブレット分解によって特徴抽出を行い，その後，SVM(Support Vector Machine) によってジェスチャの有無を識別する手法 [Otsuka 07] を用いた．ウェーブレット基底として，Daubechies10 を用い，スケール 2 から 4 まで多重解像度分解したときの各高周波数成分の統計量（平均値，最小値，最大値, 標準偏差）を特徴量とした．窓幅は 17 とした．SVM には RBF カーネルを使用し，パラメータは探索の結果，RBF カーネルパラメータ $\gamma = 0.03$，コストパラメータ $c = 128$ とした．

以下，二種類の評価結果を示す．まず，表 1 には，各グループに特化したモデルの評価結果を示す．これは同一グループ 4 人の 1 会話分を教師データとして，同グループの別会話をテストデータとした交差検定を 2 グループ分実施したときの平均値である．表 1 から，従来法と比較して提案法は，F 値にて 12.3 ポイントの大幅な性能向上が読み取れる．次に表 2 には，汎用モデルの評価結果を示す．これは一つのグループの 2 会話分を教師データとし，別のグループの 2 会話をテストデータとし，グループ間で教師データ，テストデータを入れ替えて評価を行った時の平均値である．表 2 からは，従来法と比較して提案法は，F 値にて 22.5 ポイントの大幅な性能向上が見て取れる．表 1 のグループ特化モデルと比較して，SVM の F 値は大きく低下している一方，CNN はほぼ同等の性能を保っており，汎化性能においても CNN の優位性が顕著といえる．

図 2 には，表 2 の汎用モデルによる認識結果を時系列として可視化した例を示す（G2-C2 の人物 3）．SVM の結果に検出漏れが多い反面，CNN の結果は，教師データ（正解データ）と近く，より高精度な認識がなされていることがわかる．

図 3 には，汎用モデルの最初の畳み込み層 (conv1) で学習された畳み込みフィルタ（elevation 成分）の重みの一部を図示する．人物の頷き時の頭部姿勢の時系列に含まれる様々な特徴（低周波成分や高周波成分）が抽出できるようなフィルタが自動的に構成されていることが伺える．

4. むすび

本稿では，会話中の人物頭部ジェスチャを自動的に認識する手法として，畳み込みニューラルネットを用いた方法が有効

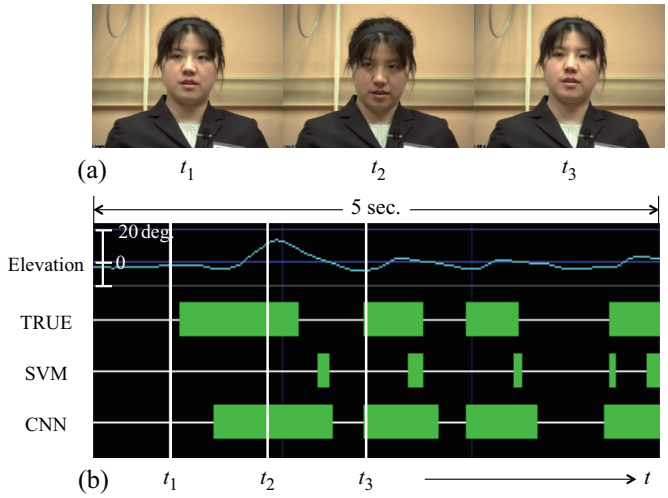

図 2: ジェスチャ時系列と認識結果の例．(a) 人物顔画像のスナップショット，(b) 時系列表現，Elevation: 頭部姿勢角，TRUE: 正解（教師）データ，SVM: SVM モデルの認識結果，CNN: CNN モデルの認識結果

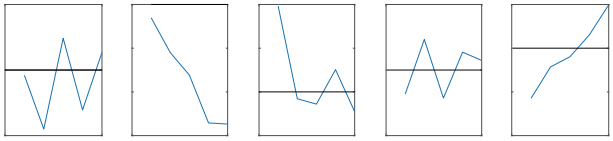

図 3: 学習された畳み込み層 (conv1) の重みの例．

であることを示した．今後の課題としては，認識カテゴリーを首振りや傾げなど複数の頭部ジェスチャに拡大することや，入力手段をセンサー以外の画像入力にも対応すること，さらには同意・否定といったより高次の心的状態やジェスチャ機能の分類なども視野に入る．また，時系列データの識別などにおいて有効とされるリカーレントニューラルネット（RNN）や LSTM(Long Short-Term Memory) など他の深層学習モデルとの比較も検討課題として上げられる．以上を含め，今後，様々な社会的信号処理に対して深層学習の活用が期待される．

参考文献

[Maynard 87] Maynard, S. K.: Interactional Functions of a Nonverbal Sign: Head Movement in Japanese Dyadic Casual Conversation, *J. Pragmatics*, Vol. 11, pp. 589–606 (1987)

[Otsuka 07] Otsuka, K., Sawada, H., and Yamato, J.: Automatic Inference of Cross-modal Nonverbal Interactions in Multiparty Conversations, in *Proc. ACM ICMI'07*, pp. 255–262 (2007)

[Yang 15] Yang, J. B., Nguyen, M. N., San, P. P., Li, X. L., and Krishnaswamy, S.: Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition, in *Proc. IJCAI'15*, pp. 3995–4001 (2015)

[Yang 16] Yang, J. B.: https://github.com/sibosutd/cnn-timeseries (2016)