

2-Stream Recurrent Convolutional Neural Networkを用いた会話中の表情認識

Conversational Facial Expression Recognition using 2-Stream Recurrent Convolutional Neural Network

吹上雅樹[†]

若林哲史[†]

[†] 大学工学部 研究科

Masaki FUKIAGE[†]

Tetsushi WAKABAYASHI[†]

[†]Mie University

アブストラクト 従来の表情認識の研究は、Ekman らが定義した感情を表す 6 表情を認識対象としているが、コミュニケーションで見られる表情を無視している。そのため本研究では、Kaulard らが定義したコミュニケーション時に見られる表情を認識対象とする。コミュニケーション中の表情は顔の動きだけでなく頭部の動きを考慮する必要があるため、両方の動きを学習できる 2-stream RCNN アーキテクチャを提案し、MPI large facial expression database を対象として Leave one person out 交差検証による評価実験を行ったところ、認識精度が 40.4% となった。

表 1: 認識対象となる表情ラベル

Emotion			
Agree	Aha light bulb moment	Annoyed	Bored
Compassion	Confused	I don't know	Disagree
Disbelief	Imagine negative	Imagine positive	Insecurity
Not convinced	Thinking	Tired	Bambi-eyes

1 はじめに

我々は日頃、絶え間なく人とコミュニケーションをとっている。人と人がコミュニケーションをとる際には言語情報と非言語情報がやり取りされるが、言語情報よりも非言語情報が重視される場合が多い。非言語情報の代表である表情は、相手の感情を推定する重要な手がかりである。また、近年のコミュニケーションロボットの普及により、人とロボットがコミュニケーションをとる機会が増えている。ロボットが人の様々な表情を認識できることで、より高度なコミュニケーションが可能となり、ヒューマン・コンピューター・インタラクションの高度化につながる。そのため、より多くの表情を認識できる自動表情認識システムの実現と精度向上は重要な課題である。

認識に関するこれまでの研究 [4,5,8,9] の多くは、基本的な感情を表す 6 種類の表情 (怒り、悲しみ、喜び、驚き、軽蔑、恐怖)[6] を認識対象とした手法を提案している。本来、表情はコミュニケーション時にやり取りされる情報だが、従来手法はコミュニケーションでよく見られる表情を無視している。また、コミュニケーションでは表情の動きだけでなく、頭部の動きの情報が必要となる。そのため本研究では、コミュニケーション、特に会話中にみられる表情を対象とした動画像を入力とする表情の動きと

頭部の動きの両方を考慮した表情認識手法を提案する。

2 表情動画データベース

本研究では、実験データとして会話中の表情変化を再現した The large MPI Facial Expression Database[3,7] を用いる。このデータセットには 51 種類の表情動画があり、それぞれの表情動画は Emotional, Subordinate emotional, Conversational, Subordinate conversational の 4 種類に分類されている。今回の実験では、Conversational, Subordinate conversational に分類されている 29 種類の表情動画を使用する。また、表情や頭部の動きでは判断できない表情動画があるため、今回の実験では表情ラベルを 29 種類から 16 種類に再編した。今回の実験では表 1 で示す 16 種類の表情を認識対象とする。

3 提案手法

動画から会話中の表情の動きと頭部の動きを学習するために、表情のフレーム間差分画像を入力とする Facial ストリームと顔特徴点画像を入力とする landmark ストリームを組み合わせた 2 ストリーム Recurrent Convolutional

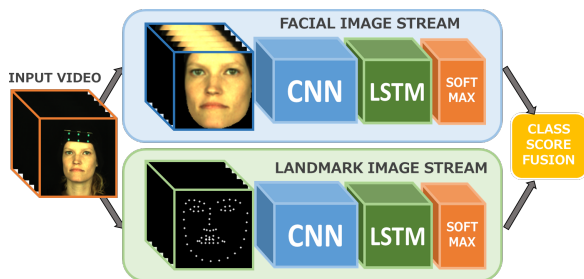


図 1: 2-stream アーキテクチャの構造

Neural Network(RCNN) アーキテクチャ[2] を提案する。このアーキテクチャの構造を図 1 に示す。入力は表情動画であり、出力は表情ラベルである。それぞれのストリームでは各入力画像に対して、Convolutional Neural Network (CNN) と Long-Short Term Memory Recurrent Neural Network (LSTM-RNN) を用いて softmax 値を算出する。それぞれの softmax 値の平均値から推定ラベルを決定する。

3.1 入力画像

Facial-stream には、フレーム間差分画像が入力され、Landmark-stream には顔特徴点画像が入力される。差分画像は 0 フレーム目の顔画像を正面化し、顔の輪郭を切り出したアライメント画像と入力フレームのアライメント画像の差の絶対値から生成した。顔画像ライブラリである OPENFACE[1] を用いてアライメントを行い、Python Image Library (PIL) を用いて差分画像を生成した。特徴点画像はアライメント画像から Dlib ライブラリを用いて生成した。特徴点の数は 68 個で、1 画素の白点を特徴点上に描画した。

3.2 CNN

画像から分類に必要な表情の動き特徴と頭部の動き特徴を抽出するために CNN を用いた。本研究で用いる CNN の構成を図 2 に示す。ここで Input は入力画像であり、チャンネル数×縦サイズ×横サイズで表す。また、Conv は畳み込み層、Pool はプーリング層、Fc は全結合を示す。活性化関数には Relu 関数を用いる。出力された 1024 次元の特徴ベクトルを LSTM-RNN に入力する。

3.3 LSTM-RNN

時系列データの特徴を識別するために LSTM-RNN を用いた。CNN から抽出された特徴ベクトルを入力し、得られた softmax 値を平均したものから推定ラベルを決定する。LSTM-RNN は中間層が 2 層、出力層が 1 層で構成されており、それぞれ中間層のユニット数は 400、出力層のユニット数は 16 である。損失関数として softmax cross

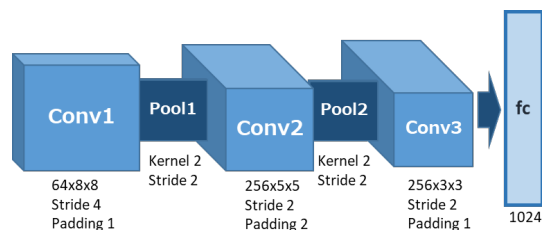


図 2: CNN の構造

entropy を用いた。

4 評価実験と実験結果

4.1 実験条件

本実験では、10 人分の表情動画を、学習用 8 人分、検証用 1 人分、評価用 1 人分に分割したデータセットを用いて、Leave One Person Out 交差検証を行う。また、学習用データを奇数フレームを含む動画と偶数フレームを含む動画に分けた。最適化関数は SGD momentum を用い、学習率は 0.003% とした。学習時に検証用データを入力した際の損失率が最も低かったモデルを評価に用いた。

4.2 実験結果

式 (1) を用いて提案したモデルの評価を行った。評価実験の結果、認識精度が 40.4% となった。

$$\text{認識精度} = \frac{\text{推定成功動画数}}{\text{全動画数}} \times 100 \quad (1)$$

5 評価実験と実験結果

本研究では、会話中の表情を認識対象とした 2 ストリームの RCNN を用いた手法を提案した。Leave one person out 交差検証での実験では、40.4% の認識精度となった。使用したデータセットは動画単位で表情ラベルが付与されているため、今回の実験では動画単位の表情推定を行っている。しかし、動画単位での表情推定は応用範囲が狭いため、リアルタイムの認識ができるフレーム単位の表情推定が望ましい。動画内の各フレームに表情なしのラベルと表情あり (動画単位の表情ラベル) を付与し、LSTM-RNN で学習することによって、フレーム単位で表情なしか、その会話中のどの表情かを推定することが可能になる。表情なしのフレームは自動判定によりラベル付けすること

を考えている。そのために表情の有無をいかに自動判定するかが今後の課題である。

参考文献

- [1] M. M. R. Tadas Baltrusaitis, “Constrained Local Neural Fields for robust facial landmark detection in the wild”
- [2] A. Z. Karen Simonyan, “Two-Stream Convolutional Networks for Action Recognition in Videos”
- [3] Cottbus-Senftenberg, “The large MPI Facial Expression Database”
- [4] M. H. M. Behzad Hasani, “Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks”
- [5] X. D. Y. Y.Fan, “Video-based emotion recognition using cnn-rnn and c3d hybrid networks”
- [6] P. E. a. W. V. Friesen, “Constans across cultures in the face and emotion,” Jounal of personality and social psychology”
- [7] C. D. B. H. W. C. Kaulard K1, “The MPI Facial Expression Database — A Validated Database of Emotional and Conversational Facial Expressions”
- [8] J.Donahue, L.Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Sachko, and T. Dattrell, “Long-term reccurnt convolutional networks for visual recognition and description”
- [9] Y.H.Byeon, K.C.Kwak, “Faxial Expression Recognition Using 3D Convolutional Neural Network”