7U-06

Attention機構を用いた深層学習による表情認識

狩野悌久 †　　長尾智晴 ‡

† 横浜国立大学 大学院環境情報学府　　‡ 横浜国立大学 大学院環境情報研究院

1　はじめに

近年，人間の感情を定量的に評価する感情認識はサービス応用などの面から注目されている．その中でも，顔画像を用いた表情認識の研究は盛んに行われ，深層学習を用いた手法が良好な結果を示している [1][2]．表情認識において，データセットの多くは静止画像に対して離散的なアノテーションされたものが主であり，動画像に対して連続的なアノテーションが施されたものは少ない．そのため静止画像を対象とした認識が主流である．しかし，静止画像に対する認識器では，時間的な変化を捉えることができないため，動画像フレームに適用した場合に，微妙な表情変化であっても認識結果が大きく異なってしまうという課題がある．そこで本研究では，Attention 機構を導入し，時間方向の特徴に重み付けを行うことで，静止画像で学習したモデルを動画像に対して効果的に適用するための手法を提案する．

2　提案手法

提案モデルの概要を図 1 に示す．提案モデルは「静止画像を対象とした認識モデルの学習」と「動画像を用いた時系列 Attention 機構の学習」の 2 段階の学習が行われる．

2.1　静止画像を対象とした認識モデルの学習

このフェーズでは，大量に準備可能な静止画像データを用いて，Convolutional Neural Network ベースの表情認識モデルの学習を行う．このフェーズの目的は，大量のデータを用いて，顔画像から汎化性の高い表情特徴を抽出する下位レイヤ（図 1(a)）と抽出された表情特徴から分類を行う上位レイヤ (図 1(b)) の学習を行うことである．学習は，softmax cross entoropy を誤差関数とした一般的なクラス分類タスクとして行われ，学習終了時に全てのユニットの重みは固定される．

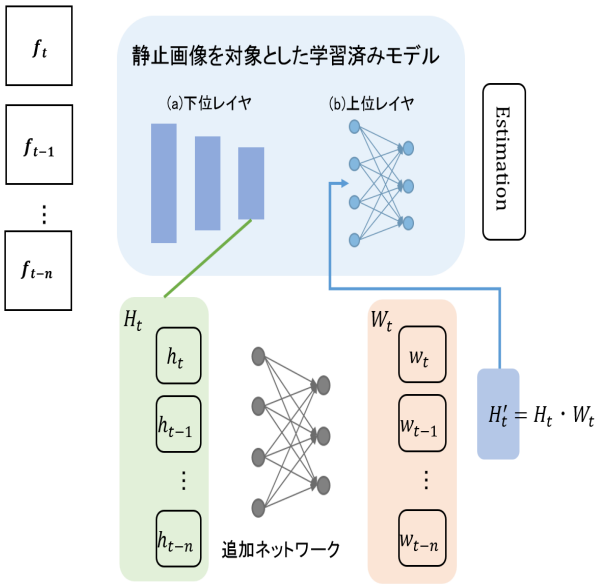

図 1: モデル概要

2.2　動画像を用いた時系列 Attention 機構の学習

このフェーズでは，2.1 で学習を行った静止画像を対象とした表情認識モデルを動画像データに効果的に適用するために，時間方向に対して重み付けする追加ネットワークの学習を行う．このネットワークは，動画像フレーム f_{t-n}〜f_t をそれぞれ学習済みモデルの下位レイヤに入力した際の表情特徴を結合・ベクトル化した $\mathbf{H_t} = (h_{t-n}, ..., h_{t-1}, h_t)$ を入力として，それぞれの特徴に対する重み係数 $\mathbf{W_t} = (w_{t-n}, ..., w_{t-1}, w_t)$ を出力する．そして，フレームごとに重み付けされた表情特徴 $\mathbf{H'_t} = \mathbf{H_t} \cdot \mathbf{W_t}$ として学習済みモデルの上位レイヤに入力され，分類が行われる．ネットワークの学習は同じ動画像内で 1 フレームだけずれた特徴量 $\mathbf{H'_t}, \mathbf{H'_{t+1}}$ をそれぞれ上位レイヤに入力した際の出力の平均二乗誤差を誤差関数とすることで行う．これにより，追加ネットワークは隣接したフレーム間では似通った分類結果となるように，それぞれのフレームに対して重み付けを行うことを目的として学習される．

Deep Learning with Attention Mechanism for Facial Expression Recognition

†Yoshihisa Kanou　‡Tomoharu Nagao

†Graduate School of Environment and Information Sciences, Yokohama National University

‡Faculty of Environment and Information Sciences, Yokohama National University

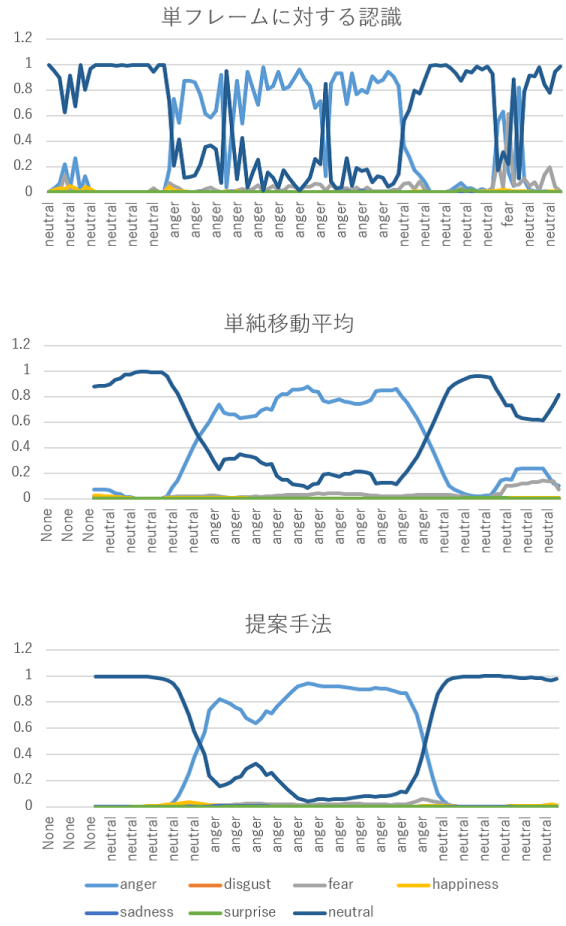

図 2: 表情認識結果

3 表情認識実験

3.1 実験設定

提案手法を動画像に適用し，予測結果が時系列的にどのように変化するか検証を行った．静止画像を用いて学習を行う表情認識モデルには，Residual Attention Network[3] で提案された Atteniton 機構を導入し，空間方向の Attention を考慮できるモデルとした．また時間方向に重み付けを行う追加ネットワークは 3 層の全結合ニューラルネットワークを用いた．静止画像・動画像データセットはそれぞれ fer2013[4]・The MUG Facial Expression Database[5] を用い，時系列考慮のフレーム幅 $n = 10$ とした．適用対象として The MUG Facial Expression Database 内の動画で，neutral から anger，そして neutral へもどる動画データを用いた．

3.2 実験結果と考察

図 2 に実験結果を示す．結果は上から単フレームに対する認識，単フレームの結果に対して単純移動平均をとったもの，提案手法を表している．

単フレームに対する認識では認識結果が隣接フレームで大きく振動していることが分かる．また単純移動平均をとった場合であっても，大きな認識結果のブレにより，曲線が乱れている．提案手法では，フレームごとの特徴量に対して適切な重み付けが行われているため，局所的な認識結果のブレに影響されにくく，より滑らかな表情変化として捉えることが可能になっている．

4 まとめと今後の課題

本稿では，静止画像を対象としたモデルに対して，時間方向に重み付けを行うネットワークを追加することで，効果的に動画像に適用する手法の提案を行った．

今後は照明変化などの外乱により，単フレームでの認識結果がより大きく振動する動画データに対して，提案手法の有効性の検証を行う．また正負の感情を考慮し，感情間の関係性をフレームの重み付けに組み込んで行きたい．

参考文献

[1] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, A. Courville, P. Vincent, et al. ”Emonets: Multimodal deep learning approaches for emotion recognition in video.” Journal on Multimodal User Interfaces 10.2 ,2016.

[2] H. W. Ng, V. D. Nguyen, V. Vonikakis , and S. Winkler. ”Deep learning for emotion recognition on small datasets using transfer learning.” Proceedings of the 2015 ACM on international conference on multimodal interaction. ACM, 2015.

[3] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. ”Residual Attention Network for Image Classification.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[4] P. L. Carrier, A. Courville, I. J. Goodfellow, M. Mirza, and Y. Bengio. ”FER-2013 face database.” Universit de Montral, 2013.

[5] N. Aifanti, C. Papachristou, and A. Delopoulos. ”The MUG facial expression database.” Image analysis for multimedia interactive services (WIAMIS), 2010 11th international workshop on. IEEE, 2010.