J-045

時系列を考慮した生体情報と映像の物理特徴量及び音源分離による音響特徴量を用いた感情推定に関する検討

A Consideration on Emotion Estimation Using Bio-signals and Physical Features of Video with Sound-source Separated Audio Features Based on Time-series Analysis

井上 竜一† 菅沼 睦‡ 亀山 渉‡‡
Ryuichi INOUE Mutsumi SUGANUMA Wataru KAMEYAMA

1. はじめに

動画共有サービスの普及によって，ユーザは様々な種類の動画コンテンツを視聴できるようになった反面，大量の動画の中からユーザの趣味や嗜好に沿ったコンテンツを探すことが困難になった．この問題を解決するために，筆者らは，コンテンツ推薦システムの補助として用いることを目的に，映像視聴者の感情を機械学習により推定する研究を行っている．

筆者らの研究[1]においては，映像視聴中の被験者に提示する映像本数を増やし，生体情報に加えて映像から得られる輝度，色彩，エッジ等の物理特徴量を機械学習の説明変数に加えることで，感情推定精度が向上することを確認した．また，筆者らの研究[2]では，映像音声から音源分離により抽出された人の声やドラム音等の音響特徴量を説明変数として追加することで，感情推定精度が向上することを確認した．加えて，筆者らの研究[3]においては，音源分離をしないデータに時系列解析を行うことで，感情推定精度が向上することを確認した．そこで，本稿では，更なる感情推定精度の向上を目指すために，音源分離に加えて時系列を考慮した解析を行った結果について報告する．

2. 実験概要

本稿で使用する生体情報データは文献[1]の実験にて取得された被験者3名（男性3名，平均年齢23歳，標準偏差1）のものである．実験では，被験者が 1 本 180 秒の映像を視聴中に生体情報を取得し，その視聴直後に，同一の映像を見ながら生起した感情についてジョイスティックで回答を行う．この手順を映像 160 本にわたって繰り返し実施する．取得する生体情報は，基礎律動，瞳孔径，視線情報，眼球の 3 次元位置変化，顔特徴点，心拍間隔（RRI）である．

3. 解析データセット

本稿で使用するデータセットは，文献[3]のものと同様であるが，音響特徴量のデータには音源分離を適用して抽出した音源ごとの音響特徴量を用いた．音源ごとの音響特徴量の抽出方法は，文献[2]と同様に，Spleeter[4]によって，実験映像の音を「声，声以外，ドラム，ベース，ピアノ，その他」の 6 成分に音源分離を行った後に，audioFeatureExtractor[5] 及び MoSQITo[6] を用いて，melSpectrum，mfcc，Loudness 等の計 54 次元の音響特徴量を各分離成分で抽出するものである．以下では，この新たな解析データセットを「音源分離あり」と呼称し，文献[3]と同じデータセットを「音源分離なし」と呼称する．

† 早稲田大学大学院基幹理工学研究科，Waseda Univ.
‡ 多摩大学経営情報学部，Tama Univ.
‡‡ 早稲田大学理工学術院，Waseda Univ.

4. 解析手法

解析では，以下に説明する決定木ベースの機械学習モデル及び Neural Network ベースの機械学習モデルの計 5 個で，「愉快な，興奮した，怖い，悲しい，リラックスした」の 5 クラス分類を行った．また，検証手法は，映像 160 本の内，20 本をテストデータとする 8 分割交差検証である．

4.1 決定木ベース機械学習モデル

決定木ベースの機械学習として，scikit-learn の Random Forest（RF）及び sktime の Canonical Interval Forest[7]（CIF）を用いる．CIF は，Random Forest 分類器を時系列データに適合できるようにしたモデルである．CIF では，時系列データをランダムな位置と長さでいくつかの interval に分割した後，各 interval において，平均，標準偏差，勾配に加えて catch22[8]という時系列特徴量を算出し，これらの値から決定木を学習するという手順を繰り返すことで，決定木のアンサンブル学習を行う．表 1 に，RF 及び CIF の実行条件を示す．

表 1　RF と CIF の実行条件

RF		CIF	
パラメータ	値	パラメータ	値
n_estimators	30	n_estimators	30
max_features	100	n_intervals	100
		min_interval	20
		max_interval	40
		シーケンス長	40 (10[s])

4.2 Neural Network ベース機械学習モデル

Neural Network ベースの機械学習として，文献[3]と同様の Deep Neural Network（DNN）と Long Short-Term Memory（LSTM），1 Dimensional Convolutional Neural Network（1DCNN）を用いる．LSTM は，DNN の中間層を時系列の考慮ができるものに改良したモデルであり，1DCNN は，データに対して時間軸方向に畳み込み処理を行うことで，時系列の考慮を可能としたモデルである．それぞれの実行条件を以下の表 2，表 3 及び図 1 に示す．LSTM の実行条件は，表 3 に示すもの以外は表 2 の DNN と同様である．

表 2　DNN の実行条件（文献[3]より引用）

中間層	出力次元	最適化関数
256→128→128→64	5	Adam
ドロップアウト率	エポック	バッチサイズ
0.5	50	128
損失関数	学習率	活性化関数
Categorical_crossentropy	0.0001	Relu

表 3　LSTM の実行条件（文献[3]より引用加筆）

中間層	シーケンス長
256(LSTM)→128(LSTM)→128→64	40 (10[s])

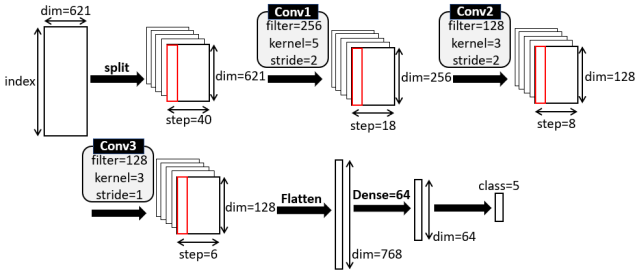

図 1　1DCNN の実行条件（文献[3]より引用）

5. 解析結果

各解析手法における感情の 5 クラス分類の推定精度を表 4 に示す．なお，推定精度は 8 分割交差検証の平均値を示している．

表 4　各モデルでの推定精度

被験者	データセット名	RF	CIF	DNN	LSTM	1DCNN
1	音源分離あり	0.708	0.747	0.729	0.758	**0.776**
	音源分離なし	0.680	0.738	0.718	0.730	**0.750**
2	音源分離あり	0.701	0.728	0.716	0.732	**0.749**
	音源分離なし	0.679	**0.725**	0.694	0.702	0.717
3	音源分離あり	0.663	0.685	0.684	0.693	**0.701**
	音源分離なし	0.631	0.658	0.656	0.667	**0.676**
平均	音源分離あり	0.691	0.720	0.710	0.728	**0.742**
	音源分離なし	0.663	0.707	0.689	0.700	**0.714**

表 4 より，どの被験者のどのモデルでも音源分離ありデータセットの推定精度が音源分離なしデータセットの推定精度よりも高いことから，提案する音源分離を用いた音響特徴量の有効性が確認された．また，RF モデルに比べて CIF モデルの推定精度が高いこと，並びに，DNN モデルに比べて LSTM，1DCNN モデルの推定精度が高いことから，音源分離ありデータセットにおいても時系列の考慮が有効であることが確認された．また，音源分離ありのデータセットにおいて，最も推定精度が高いモデルは 1DCNN であり，被験者間平均推定精度は 0.742 となった．

次に，音源分離ありの 1DCNN，DNN 及び音源分離なしの 1DCNN における感情ごとの F 値，最も精度の高かった音源分離ありの 1DCNN を基準とした F 値の差について表 5 に示す．ここで，F 値は 8 分割交差検証の平均値の被験者間平均を示している．

表 5　各感情の F 値

モデル名	データセット名	愉快	興奮	怖い	悲しい	リラックス
1DCNN	音源分離あり	0.893	0.741	0.812	0.559	0.647
DNN	音源分離あり	0.885	0.694	0.764	0.503	0.641
1DCNN	音源分離なし	0.887	0.645	0.793	0.502	0.644
音源分離ありの 1DCNN と DNN の差		0.008	**0.047**	**0.048**	**0.056**	0.006
1DCNN の音源分離ありとなしの差		0.006	**0.096**	0.019	**0.057**	0.003

表 5 より，音源分離ありの 1DCNN と DNN の差から，音源分離を用いたデータセットにおいて，時系列を考慮することによって，特に「興奮した，怖い，悲しい」の感情の F 値が向上することが確認された．また，1DCNN モデルにおいて，音源分離を用いた音響特徴量の利用によって，特に「興奮した，悲しい」の感情の F 値が向上することが確認された．

6. まとめと今後の課題

本稿では，音源分離を用いたデータセットにおける時系列を考慮した解析の有効性について検証した．どのモデルにおいても音源分離を用いることで推定精度が向上し，どの時系列モデルにおいても，時系列考慮の有効性が確認された．また，音源分離を用いた場合，最も高い推定精度となったモデルは 1DCNN であり，推定精度は被験者間平均で 0.742 であった．今後の課題として，被験者数を増やし汎化性能の評価をすることや，更なる特徴量の追加，モデルのパラメータチューニング等が挙げられる．

参考文献

[1] 小野浩輝, 井上竜一, 菅沼睦, 亀山渉, “生体情報と映像の物理特徴量を用いた DNN による映像視聴者の未学習映像における感情分類精度の向上に関する検討,” 電子情報通信学会, HCG シンポジウム 2022, C-2-6, Dec. 2022.
[2] 井上竜一, 小野浩輝, 菅沼睦, 亀山渉, “生体情報と映像の物理特徴量及び音源分離による音響特徴量を利用した映像視聴者の感情推定に関する検討,” 電子情報通信学会, 2023 年総合大会, H-2-12, Mar. 2023.
[3] 小野浩輝, 井上竜一, 菅沼睦, 亀山渉, “時系列分析による生体情報と映像の物理特徴量を用いた未学習映像での感情分類精度の向上に関する検討,” 電子情報通信学会, 2023 年総合大会, H-2-15, Mar. 2023.
[4] Deezer, “deezer/spleeter: Deezer source separation library including pretrained models,” https://github.com/deezer/spleeter (最終確認：2023 年 6 月 6 日).
[5] MathWorks, “Streamline audio feature extraction - MATLAB,” https://jp.mathworks.com/help/audio/ref/audiofeatureextractor.html (最終確認：2023 年 6 月 6 日).
[6] Green Forge Coop., MOSQITO [Computer Software], https://doi.org/10.5281/zenodo.5284054 (最終確認：2023 年 6 月 6 日).
[7] M. Middlehurst, J. Large, A. Bagnall, “The Canonical Interval Forest (CIF) Classifier for Time Series Classification,” IEEE International Conference on Big Data, pp.188-195, 2020.
[8] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, N. S. Jones, “catch22: CAnonical Time-series Characteristics,” Data Mining and Knowledge Discovery, vol.33, no.6, pp.1821-1852, 2019.