VGG-Faceによるフレーム選択に基づく第一印象の予測

田中　祐仁† 　山田　一権†† 　白井　匡人†††

† 島根大学 自然科学研究科　〒 690-8504 島根県松江市西川津町 1060
†† エクスウェア株式会社　〒 140-0002 東京都品川区東品川 4 丁目 10
††† 島根大学 学術研究院理工学系　〒 690-8504 島根県松江市西川津町 1060
E-mail: †n19m106@matsu.shimane-u.ac.jp, ††k-yamada@xware.co.jp, †††shirai@cis.shimane-u.ac.jp

あらまし　本研究では, 映像・音声・テキストの 3 つのモダリティを用いた見かけの性格指標の予測を行う. 第一印象が評価される要因は, 表情や話し方, 発言内容など多岐に渡る. 近年マルチモーダル手法を用いた第一印象の予測に関する研究が盛んに行われている. 特にマルチモーダル深層学習は複数のモダリティを同時に考慮して学習を行うため, 複数の要因を加味して予測することが可能となる. 先行研究では, 映像特徴に関して単一フレームの抽出やランダムなフレーム選択を用いて予測しているものが多く, 決定的なフレームの選択方法は提案されていない. そこで本研究では, VGG-Face Network を特徴抽出器として用いた, 最適なフレーム選択方法を提案する. 実験により提案手法の有効性を示す.

キーワード　マルチモーダル深層学習, 第一印象の予測

1 前　書　き

他者にどのような第一印象を与えるかという問題は, 日常生活は勿論, 就職面接等においても非常に重要である. 第一印象を自動的に予測することにより, 自分が他者に一般的にどのような印象を与えるかを知ることができ, 自分自身をより良く見せる手助けとなる. しかしながら, 第一印象が評価される要因は, 表情や話し方, 発言内容など多岐に渡るため, 自動的に予測することは容易ではない.

本研究では, 代表的な性格特性である, Big Five の予測を行う. Big Five は外向性 (Extraversion), 協調性 (Agreeableness), 誠実性 (Conscientiousness), 神経症的傾向 (Neuroticism), 開放性 (Openness) の 5 つから成る性格特性である. 性格特性の予測には人手でラベル付けされた見かけの性格特性の予測と, アンケートにより得られる実際の性格特性を予測する 2 種類の研究があるが, 本研究では, 人手でラベル付された見かけの性格特性を対象とする.

近年, マルチモーダル手法を用いた第一印象の予測に関する研究が盛んに行われている. 一般的にマルチモーダル深層学習では, 映像・音声・テキストの 3 つのモダリティが対象となる. 先行研究では, 映像モダリティに関して, ランダムなフレーム選択や, 単一のフレーム選択により予測を行うものが多い. しかし, 使用するフレームがランダムに選択される場合, 特徴的なフレームが選択されるとは限らない為, 精度が安定しない. また, 映像には複数の表情が現れる場合があるが, 単一のフレームのみを用いる場合, 表情の変化を加味して予測を行うことができない. 本研究では, 動画から複数のフレームを決定的に選択することにより, 予測精度の向上を目指す.

第 2 章では関連研究について述べ, 第 3 章ではマルチモーダル深層学習について述べる. 第 4 章では提案手法について述べる. 第 5 章では実験により有効性を示す. 第 6 章で結論とする.

2 パーソナリティの予測

2.1 関 連 研 究

Zhang らは, 動画から 1 秒間に 6 つ, 動画全体から約 100 枚のフレームを抽出し, 同じ正解ラベルを付与することにより, 複数のフレームを考慮して実験を行っている [1]. Subramaniam らは, 動画を 6 つのセグメントに分割し, 各セグメントからフレームをランダムに 1 枚抽出, CNN による特徴抽出後, 時系列順に LSTM への入力とする方法が提案されている [2]. Kampman らは, 動画からランダムに単一のフレームを選択し, 学習済み VGG-Face CNN モデル [3] を転移学習することにより, 映像特徴量を学習している [4]. これらの手法では, 特徴的なフレームの選択ができていない可能性がある.

3 マルチモーダル深層学習

一般に, 人が他者の第一印象を評価する際は, 声や発言内容, 表情等, 複数の要因が総合的に考慮される. 同様に, マルチモーダル深層学習は, 音声・テキスト・映像等の複数のモダリティを同時に考慮して学習を行うことができる為, 複数の要因を加味して予測することが可能である. 本研究では, マルチモーダル深層学習として Kampman ら [4] の手法に基づき, 映像特徴量に関して改良を加える.

3.1 音声特徴量

本研究では, 音声特徴として波形データを用い, これを CNN への入力とする. 音声を 8kHz にダウンサンプリングし, 録音音量によるバイアスを避けるために, 訓練データについて 10^U(U は区間 [-1.5〜1.5] の一様分布からランダムに抽出) を乗算することにより, 振幅をランダム化する. また, 音声のエネルギー成

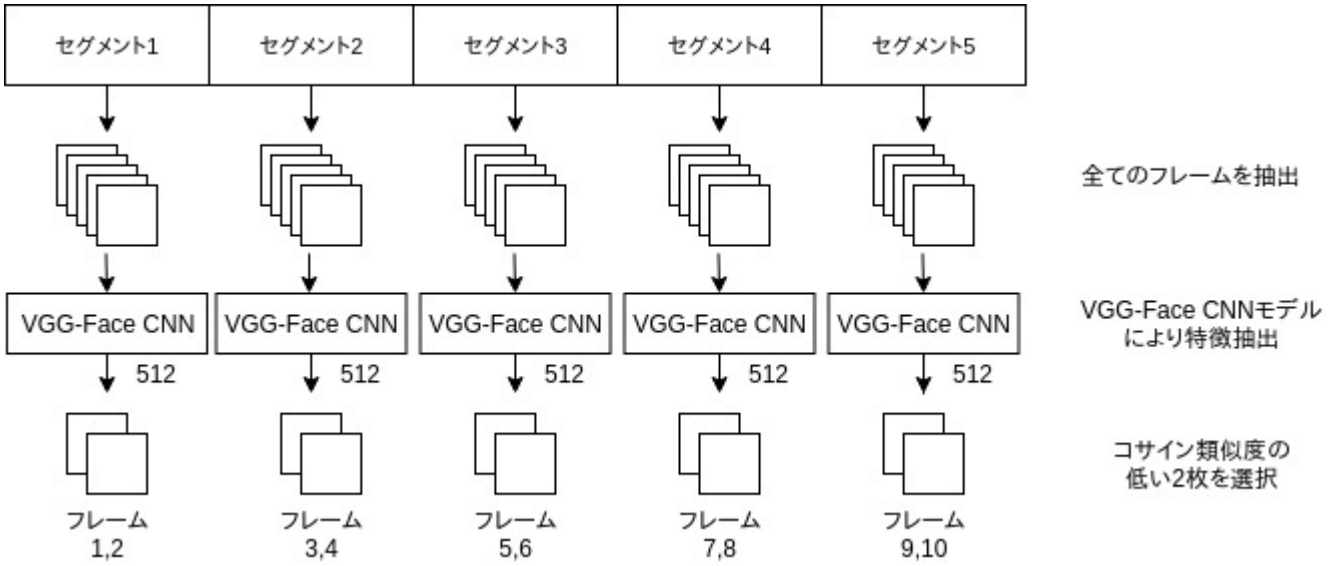

図 1　フレームの選択

分を取得するために, 振幅を 2 乗したものも CNN への入力として用いる. Bertero ら [5] を参考に, 4 つの畳み込み層, 1 つのグローバルアベレージプーリング層, 全結合層を経て 64 次元のベクトルに変換する.

3.2　テキスト特徴量

用いるデータセットは動画のみであり, テキスト情報が付与されていない為, Google Cloud Speech API を用いて文字起こしを行ったものを用いる. テキストを word2vec [6] の Google ニュースデータ (3000 億語) の学習済みモデルを用いて単語ごとに 300 次元のベクトル表現に変換する. Kim ら [7] を参考に, それぞれサイズ 3,4,5 のカーネルを用いて畳み込み, 最大プーリングを行った後接続し, テキストを 64 次元のベクトルに変換する. 過学習を防ぐため, 最後の層にドロップアウト (p=0.5) を適用する.

3.3　映像特徴量

映像からランダムに単一のフレームを抽出し, 学習済み VGG-Face CNN モデルを転移学習することにより, 画像から特徴を抽出する. 入力画像を縦幅 224, 横幅 224 のカラー画像 (224,224,3) に変換する. VGG-Face CNN モデルの全結合層を取り外して新たに全結合層を追加し, 上から 14 層目までの重みを凍結して学習を行う. 入力を 512 次元のベクトルに変換する.

3.4　結　合　層

結合層では, 音声特徴量 (64 次元), テキスト特徴量 (64 次元), 映像特徴量 (512 次元) を結合した 640 次元のベクトルを, 全結合層を通して Big five 性格特性の 5 次元に対応づける. 出力層の活性化関数にはシグモイド関数を用いる.

4　提 案 手 法

先行研究 [4] では, 動画からランダムに単一フレームを選択していたが, 本研究の提案手法では複数のフレームを用いる. フレームの抽出方法を図 1 に示す. 動画から複数のフレームを抽出するために, 1 つの動画を等間隔に 5 つのセグメントに分割する. それぞれのセグメントから全てのフレームを抽出し, 学習済み VGG-Face CNN モデルを特徴抽出器として用いることにより, 512 次元のベクトルに変換する. 最も特徴的なフレームを選択するために, それぞれのセグメントから抽出された特徴量について, ベクトル間のコサイン類似度が最も小さい 2 つのフレームを学習に使用する. したがって, 1 つの動画から 10 枚のフレームが選択される. それぞれのフレームについて, 学習済み VGG-Face CNN モデルを転移学習することにより, Big five 性格特性の 5 次元に対応づける. フレームを時系列順に並べ, CNN の中間層の出力 (512 次元) を LSTM への入力として用い, 300 次元のベクトルに変換し, 結合層への入力とする. モデルの概要を図 2 に示す.

5　実　　験

実験では, Chalearn First Impressions Challenge データセット [8] を用いてマルチモーダル深層学習を基に第一印象の予測を行う. 提案手法の有効性を確認するために, ランダムにフレーム選択を行う場合と予測精度を比較する.

5.1　実 験 準 備

Chalearn First Impressions Challenge データセットは約 15 秒の YouTube vlog クリップであり, Amazon Mechanical Turk を利用し, 人手で Big Five 性格特性のラベルが付与されている. データセットは全て, 英語でカメラに向かって話している動画で構成されており, 様々な年齢, 性別, 国籍の人物が含まれる. Training Data(6000 件) を訓練データ, Validation Data(2000 件) をテストデータとする. 実験に使用するパラメータを表 1 に示す.

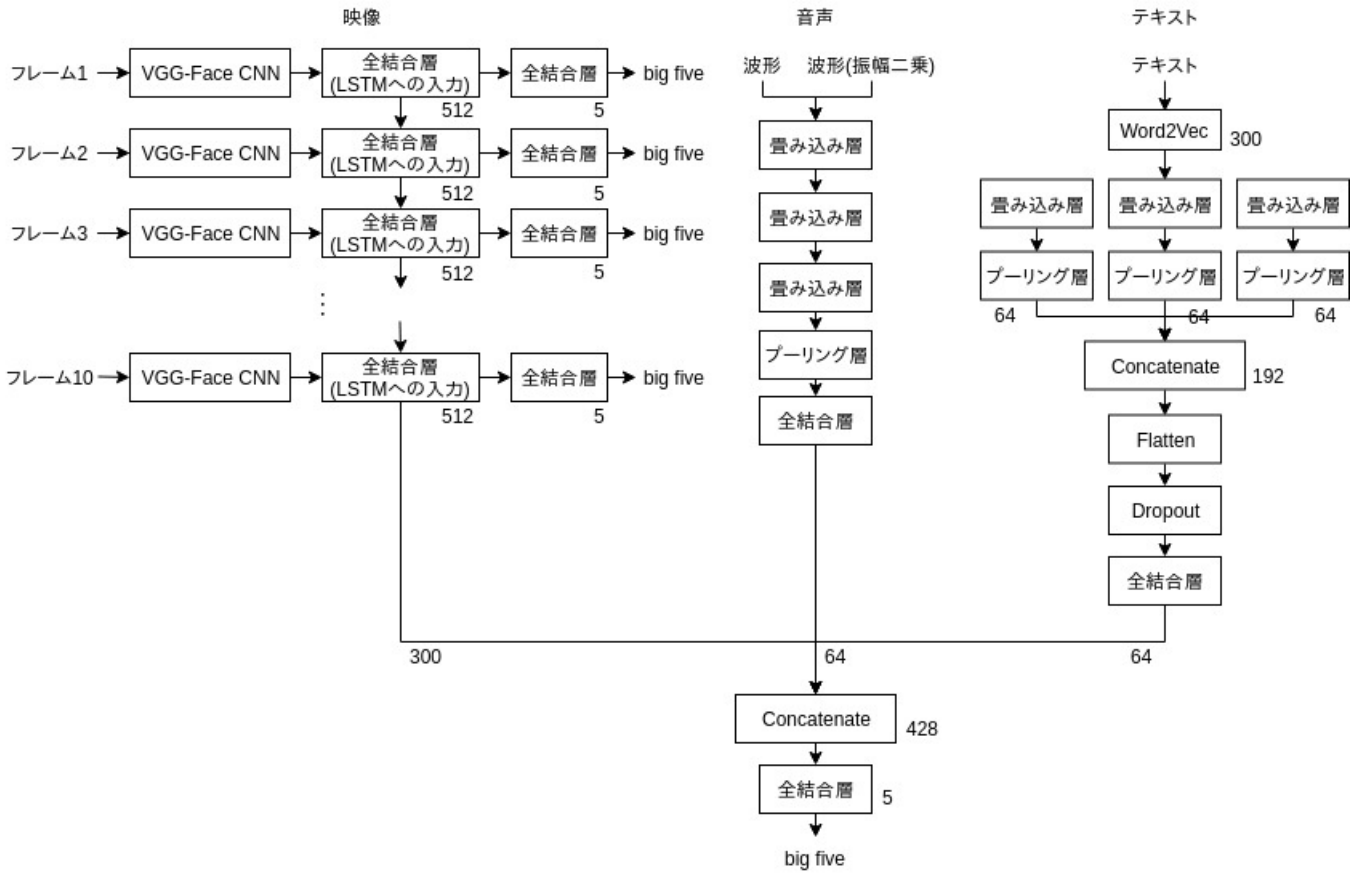

図 2　モデルの概要

表 1　使用するパラメータ

		特徴抽出部分	モデル全体
活性化関数	中間層	ReLU	ReLU
	出力層	sigmoid	sigmoid
損失関数		平均二乗誤差 (MSE)	平均二乗誤差 (MSE)
最適化関数		Adam [9]	Adam [9]
バッチサイズ		32	32
エポック数		100	100

5.2 評価方法

評価尺度は平均絶対誤差 (MAE) を用いる．Big Five の 5 つの性格特性において, 以下の式より予測値 p_i と正解ラベル y_i 間での MAE の平均を求める．

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|p_i - y_i| \tag{1}$$

5.3 実験結果

実験結果を表 2 に示す．表は, 音声のみ, テキストのみ, 映像のみを用いる場合の MAE と, モデル全体を統合的に学習する場合 (Fusion) の MAE を示している．モデル全体を統合的に学習した場合の MAE は, 単一のフレームを使用した既存手法で 0.0962, 複数のフレームを選択して用いる提案手法で 0.0887 となる．また, 個別のモダリティに着目すると, 音声・テキスト・映像の各モダリティのみを用いた MAE はそれぞれ 0.1138, 0.1166, 0.0888 となっており, 映像モダリティのみを使用した場合が最も高く, 音声, テキストモダリティのみを用いた場合の精度はこれと比較して低くなっている．各性格特性については, 外向性, 協調性, 誠実性, 神経症的傾向, 開放性のそれぞれについて, 既存手法では 0.0959, 0.0932, 0.0931, 0.0975, 0.0958, 提案手法では 0.0887, 0.0902, 0.0845, 0.0870, 0.0927, 0.0890 となる．

表 2 実験結果

MAE	Big Five 性格特性					
モデル	平均	E	A	C	N	O
音声	.1138	.1157	.1021	.1223	.1177	.1111
テキスト	.1166	.1203	.1034	.1193	.1215	.1183
映像 (既存手法) [4]	.0962	.0972	.0910	.0973	.0995	.0960
映像 (提案手法)	.0888	.0900	.0847	.0873	.0927	.0891
Fusion(既存手法) [4]	.0951	.0959	.0932	.0931	.0975	.0958
Fusion(提案手法)	.0887	.0902	.0845	.0870	.0927	.0890

E は外向性, A は協調性, C は誠実性, N は神経症的傾向, O は開放性をそれぞれ表す.

5.4 考　　察

表 2 より, ランダムなフレームを使用した既存手法と比較して, 提案手法により MAE が 0.064 改善していることがわかる. また, 映像モダリティのみを用いる場合の精度と比較して, 音声モダリティのみ, テキストモダリティのみを用いる場合の精度が低く, マルチモーダル深層学習による精度の向上はほとんど見られない. したがって, 音声モダリティについては MFCC 等の特徴量, テキストモダリティについては内容の解析が可能な手法を用いるなど, 特徴抽出, 解析方法については改善の余地がある. 各性格特性については, 提案手法, 既存手法同様に神経症的傾向の予測が最も困難であり, 協調性と誠実性の予測は比較的容易であることが分かる.

6 結　　論

本研究では, マルチモーダル深層学習による見かけの性格予測において, フレームの選択方法の改善により, 予測精度を向上させる手法を提案した. 既存手法である, 単一のフレームのみを用いた場合の平均 MAE が 0.0951 であるのに対し, 提案手法による平均 MAE は 0.0887 であった. また, 音声, テキスト, 映像の各モダリティによる個別の評価実験の精度より, 映像モダリティの精度が最も高く, 音声, テキストモダリティの精度向上が今後の課題である.

文　　献

[1] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, "Deep bimodal regression for apparent personality analysis," European Conference on Computer VisionSpringer, pp.311–324 2016.

[2] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," European Conference on Computer VisionSpringer, pp.337–348 2016.

[3] O.M. Parkhi, A. Vedaldi, A. Zisserman, et al., "Deep face recognition.," bmvc, vol.1, p.6, 2015.

[4] O. Kampman, E.J. Barezi, D. Bertero, and P. Fung, "Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp.606–611, 2018.

[5] D. Bertero, F.B. Siddique, C.-S. Wu, Y. Wan, R.H.Y. Chan, and P. Fung, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp.1042–1047, 2016.

[6] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," Advances in neural information processing systems, pp.3111–3119, 2013.

[7] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.

[8] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H.J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," European Conference on Computer VisionSpringer, pp.400–418 2016.

[9] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.