

深層ニューラルネットワークを利用した顔認識の検討

Study of Face Recognition using Deep Neural Network

河合 吉彦† 遠藤 伶† 藤森 真綱† 望月 貴裕†

Yoshihiko Kawai† Rei Endo† Naotsuna Fujimori† Takahiro Mochizuki†

1 はじめに

テレビ番組の効率的な制作支援を目指し、映像内容を自動認識する技術について研究をすすめている。その中でも、映像に映る人物を識別する顔認識技術は非常に応用範囲の広い技術であり、映像検索や映像加工などへの活用が期待されている。顔認識の既存研究としては、近年、深層ニューラルネットワーク (DNN) を利用した手法が多く提案されている [1]。本稿では、DNN をベースとした顔認識手法において、ネットワークの入力部分に複数の画像変換処理を追加し、認識精度を改善する手法を提案する。具体的には、入力画像に対して、エッジ検出画像、グレースケール画像、左右反転画像を連結する構造を DNN の前段に設ける。また学習の際は、学習データからノイズラベルを除去する事前処理を実施し、精度向上を図る。実験では、LFW データセット [2] を利用して提案手法の有効性を検証する。

2 提案手法

図 1 に提案手法の概要を示す。提案手法は、入力部とネットワーク部からなる。入力部は顔画像であり、出力は特徴ベクトルである。出力された特徴ベクトルの距離が近いときは、入力画像は同一人物であると判定し、遠いときは別の人物と判定する。

2.1 入力部

DNN の利点は特徴抽出の機能を学習によって構築できる点であるが、経験的に有用と思われる画像特徴を意図的にネットワークに追加することで、認識精度を向上させたり、学習時間を短縮したりすることが可能となる。また、畳み込み演算の組み合わせのみでは算出が難しい特徴も認識に利用できる。図 2 に提案手法の入力部の構造を示す。まず、微分フィルタ dx , dy によるエッジ検出画像と、深さ方向の畳み込みフィルタ ($depthwise$) によるグレースケール画像を、元の入力画像と連結 ($concat$) する。次に、連結画像とその左右反転画像 ($h-flip$) を連結してネットワーク部の入力に渡す。

2.2 ネットワーク部

ネットワーク部は、入力部で生成された連結データを DNN によって特徴ベクトルに変換する。物体認識で使用される AlexNet や ResNet, VGGNet などのネットワーク構造がそのまま利用できる。従来研究 [1] では、Inception 構造を利用する手法が提案されている

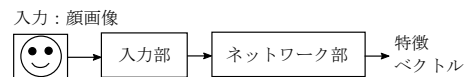


図 1 提案手法の概要

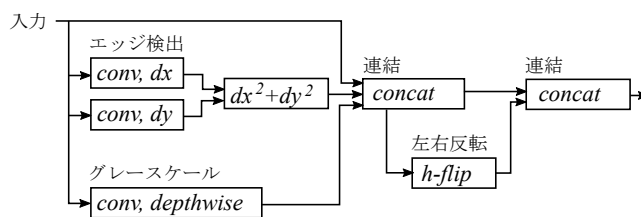


図 2 入力部の構造

が、提案手法ではより深いネットワークが構築できる Inception-ResNet V2 構造 [3] を採用する。また、従来技術 [1] では特徴ベクトルを 128 次元としていたが、提案手法では予備実験の結果を考慮して 512 次元とする。

3 ネットワークの学習

3.1 学習データセット

学習データとして、約 10 万人、合計 1,000 万枚の画像からなる MsCeleb データセット [4] を利用する。しかし、MsCeleb には誤ったラベル (人物名) が付与された画像が多く含まれており、そのまま利用すると精度低下の原因となる。画像数が膨大であるため人手による修正も難しい。そこで提案手法では、既存の軽量の DNN [5] を利用して、明らかに誤りがあると予測されるラベルの自動修正を試みる。手順を以下に示す。

1. 元のラベルを用いて DNN を学習する
2. データセットに適用し、ラベルを割り当て直す
3. 修正後のラベルを用いて DNN を再学習する
4. データセットに適用し、ラベルを再度割り当て直す

ラベルを割り当て直す際は、誤りの可能性が十分に高い (認識スコアが閾値より低い) データのみを対象とする。

3.2 学習の方法

学習の際は、学習データに含まれるラベルと同数の出力を持つ全結合層をネットワークの最終層に接続し、正解ラベルとの交差エントロピー損失を最小化する。交差エントロピー損失を利用するにあたり、パラメータ数の削減と学習処理の効率化ため、データ数が 100 画像以下のラベルを学習データから除外した。除外後の学習データのサイズは、ラベル数が約 13,000 人、画像の合計が

† NHK 放送技術研究所

表 1 学習データのラベル修正による精度比較

学習データの修正	認識率
なし	0.9897 ± 0.0039
あり	0.9945 ± 0.0032

表 2 入力部の構成による精度比較

#	エッジ	グレースケール	左右反転	認識率
1	-	-	-	0.9938 ± 0.0043
2	○	○	-	0.9942 ± 0.0042
3	○	○	○	0.9945 ± 0.0032

表 3 他手法との精度比較

手法	認識率
DeepID [7]	0.9745 ± 0.0026
人間 [2]	0.9753
FaceNet [1]	0.9920 ± 0.0035
提案手法	0.9945 ± 0.0032

約 160 万枚となった。なお、文献 [1] では、Triplet 損失の有効性が報告されているが、学習の安定性に課題があるため [6]、本手法ではシンプルで安定している交差エントロピー損失を採用した。

学習アルゴリズムには Adam を利用する。入力画像のサイズは 160×160 画素、学習ステップ数は 32 万、バッチサイズは 90 とする。初期の学習率は 0.1 に設定し、16 万、24 万、28 万ステップにおいて $1/10$ 倍する。GPU に NVIDIA GeForce GTX 1080Ti を利用して学習したところ、学習時間は約 100 時間であった。

4 評価実験

評価実験には LFW データセット [2] を利用した。LFW は、画像のペアが同一人物か否かを判定するベンチマークである。評価尺度は、正しく判定できた割合を表す認識率であり、10-fold 交差検定で評価する。

4.1 学習データのラベル修正の効果検証

始めに、学習データのラベル修正の効果を検証した。表 1 に実験結果を示す。「なし」は元の学習データをそのまま利用した場合であり、「あり」はラベルを修正した上で学習した場合を表す。修正ありの方が 0.5% 程度精度が向上していることが確認できた。

4.2 入力部の効果検証

次に、入力部の追加による効果を検証した。表 2 に結果を示す。エッジ検出画像、グレースケール画像、左右反転画像の利用による認識精度の変化を表している。いずれも利用しない構造 1 の認識率は 99.38% となり、エッジ検出画像とグレースケール画像のみを使用する構造 2 は 99.42% となった。一方、すべてを利用する提案手法の構造 3 は認識率が 99.45% となり、最も精度が高くなることが確認できた。



(a) 別人と誤認識 (b) 同一と誤認識

図 3 誤認識の例

4.3 従来手法との比較

表 3 に従来手法との比較結果を示す。DeepID[7]、FaceNet[1] は、DNN に基づく手法である。また、「人間」は人が目視で判別した際の認識精度を表す。比較の結果、提案手法の認識精度は 99.45% となり、すべての手法の中でもっとも高い精度となった。学習データのラベル修正、入力部の追加によって認識精度が向上したものと考えられる。ここで、FaceNet は本手法と同じデータで学習、テストした結果である。顔位置のアライメントを実施していないため、論文 [1] で報告されている認識精度 (99.63%) よりも低い結果となっている。

図 3 に、提案手法の誤認識の例を示す。(a) は同一人物を別人と、(b) は別人を同一人物と誤認識した例である。顔器官が隠ぺいされた場合への対応や、類似した顔の判別方法などを検討する必要がある。

5 おわりに

本稿では、深層ニューラルネットワークを利用した顔認識手法を提案した。提案手法では、ネットワークの入力部分に、エッジ検出画像、グレースケール画像、左右反転画像を連結する構造を追加することで、認識精度の向上を図った。LFW データセットを用いた実験では、認識精度が 99.45% となり、従来よりも高い精度が達成できた。今後は、ネットワーク構造を改良するなどして、さらなる精度向上を目指したい。また、テレビ番組への適用に向けて、データセットの整備をすすめたい。

参考文献

- [1] F. Schroff, *et al.*, "FaceNet: A Unified Embedding for Face Recognition and Clustering," Proc. IEEE CVPR (2015).
- [2] G.B. Huang, *et al.*, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," University of Massachusetts, Technical Report 07-49 (2007).
- [3] C. Szegedy, *et al.*, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," abs/1602.07261 (2016).
- [4] Y. Guo, *et al.*, "MS-Celeb-1M: Challenge of Recognizing One Million Celebrities in the Real World," (2016). www.microsoft.com/en-us/research/project/ms-celeb-1m-challenge-recognizing-one-million-celebrities-real-world/
- [5] X. Wu, *et al.*, "A Light CNN for Deep Face Representation with Noisy Labels," IEEE Trans. on Information Forensics and Security, Vol.13, No.11, pp.2884–2896 (2018).
- [6] J. Deng, *et al.*, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," abs/1801.07698 (2019).
- [7] Y. Sun, *et al.*, "Deep Learning Face Representation from Predicting 10,000 Classes," Proc. IEEE CVPR (2014).