計測自動制御学会東北支部第316回研究集会(2018.6.22)
資料番号316-7

Deep Learningを用いた顔認証システムに関する研究

A Study of Facial Recognition System Using Deep Learning

○佐賀健志, 釜谷博行, 工藤憲昌
Takeshi Saga, Hiroyuki Kamaya, Norimasa Kudoh
八戸工業高等専門学校
National Institute of Technology, Hachinohe College

キーワード: 機械学習(Machine Learning), ディープラーニング(Deep Learning),
顔認証(Facial Recognition), アフィン変換(Affine Transformation)

連絡先: 〒039-1192 八戸市田面木字上野平16-1 八戸工業高等専門学校　産業システム工学専攻
Tel.: 0178-27-7283, E-mail: kamaya-e@hachinohe-ct.ac.jp

1. はじめに

近年、日本の高齢者人口比率は上昇し続けている。それに伴い高齢者に関係する社会問題も深刻になってきた。特に深夜徘徊や高齢ドライバーによる交通事故数の増加は無視することができない社会問題となっている。

そこで、本研究では高齢者の夜間外出を未然に防ぐための顔認証システムの開発を目的とする。顔認証システムの実現方法として、近年画像分類研究において高い認識精度を挙げているディープラーニング(Deep Learning)[1]を用いた。

2. 開発環境

プログラミング言語はPython3.5、ライブラリはTensorFlow (機械学習)とOpenCV、Dlib (画像処理)を使用した。また、計算の高速化のためにGPU(GTX Geforce 1080Ti)を使用した。

3. Deep Learning

ニューラルネットワーク(以下 NN)とは生物の神経回路網の数学モデルであり、NNを多層にしたものがDeep Learningである。各層はユニットと呼ばれる図1の円に相当するものから構成される。

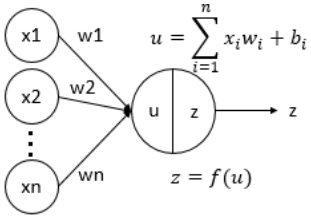

図 1　ユニットの入出力

図1のxは入力、wは重み(ユニット間の結合の強さ)、bはバイアス、uは入力と重みの線形和、fは活性化関数、zは出力を表す。出力値はwとbによって変化するため、これらの変数を適切な出力が得られるように調節することがNNにおける学習である。

表 1 ネットワーク構成 1

	kernel_size	stride	out_channel	route
input			3	
conv	3*3	1	32	
conv	3*3	1	32	
pool	2*2	2		
conv	3*3	1	48	
conv	3*3	1	48	
pool	2*2	2		
conv	3*3	1	72	
pool	2*2	2		
conv	3*3	1	108	
pool	2*2	2		
conv5	1*1	1	48	1
conv5	5*5	5	64	1
conv3	1*1	1	64	2
conv3	3*3	3	96	2
conv3	3*3	3	96	2
ave_pool	2*2	2		3
conv1	1*1	1	32	3
fc			1024	
fc			256	
output			2	

input：入力層，conv：畳み込み層，
max_pool：最大プーリング層，fc：全結合層，
avg_pool：平均プーリング層，output：出力層

4. 畳み込みニューラルネットワーク

本研究では、画像認識に有効であるとされる畳み込みニューラルネットワーク（以下 CNN）を用いた。CNN では、畳み込み（convolution）やプーリング（pooling）といった操作が行われる。この操作により、画像内の物体の平行移動に対する不変性を得る。畳み込みは、入力画像にフィルタを適用し、画像からフィルタと類似した特徴的な部分を抽出する働きを持つ。本研究は平均プーリングと最大プーリングを使用する。この操作により重要ではない情報を削減することで計算コストを抑えることができる。

CNN の主要なパラメータを以下に示す。

・学習係数（0.0002）
　重みの更新量の大きさを決める
・フィルタサイズ
　畳み込みに使用するフィルタのサイズ
・バッチサイズ（128）
　重みを更新するまでに与えるサンプル数

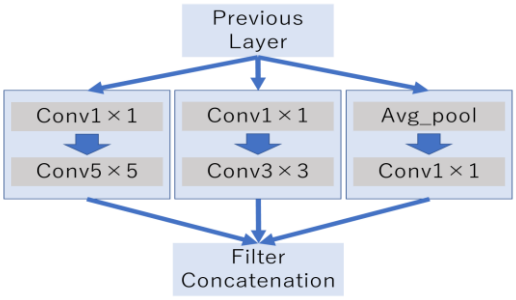

図 2　Inception 構造

・ユニット選出率（0.5）
　ドロップアウト時のユニット有効化率
・Gradient Clipping（0.005）
　急激に変化する勾配を切り捨てる閾値
・損失関数
　交差エントロピーを用いて正解値との誤差を評価
・最適化関数
　Adam を勾配計算、最適化に使用

5. 2 クラス分類

5.1　ネットワーク構成

本実験では表 1 に示す 17 層のネットワークを用いる [(2)(3)(4)]。表 1 の灰色部分は図 2 に示すような Inception 構造にしている。これにより計算が並列に行われ、計算時間を短縮できる。表 1 における Kernel size はフィルタのサイズ、stride は各演算あたりのフィルタ移動量、out_channel は出力される画像の枚数、route は Inception の枝路番号である。

5.2　データセット

本研究で用いるデータセットは、椅子に座り、撮影プログラムのインストラクションに従って、顔を動かすことで収集した。一人あたり約 1000 枚のカラー画像を用意し、約 800 枚を学習画像、約 200 枚を検証画像として用いた。また、撮影条件の異なるテスト画像を約 100 枚用意した。

用意した画像から顔領域のみを抽出するために、前処理には Dlib の顔検出器を利用した。この顔検出器は元画像から顔を検出し、

(a)　(b)　(c)

図 3　入力画像例(左から元画像、アフィン変換後画像、アフィン変換後の特徴点隠蔽画像)

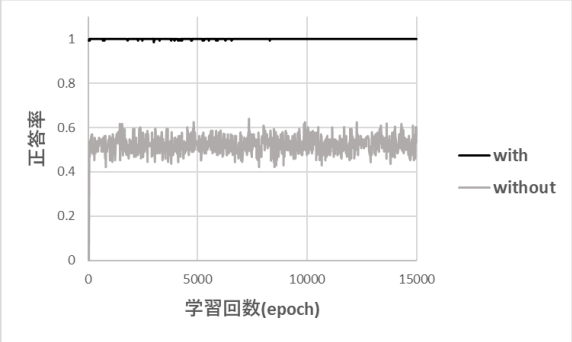

図 4　検証画像に対する認識精度

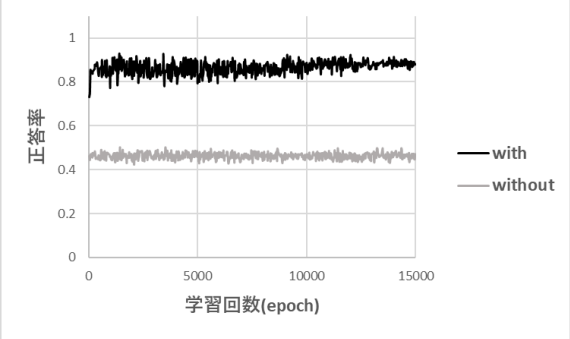

図 5　テスト画像に対する認識精度

その顔の座標平面上縦軸の最大値と最小値、座標平面上横軸の最大値と最小値を返り値として出力する。これを基準として OpenCV の矩形トリミング(cv2.rectangle)を利用して画像を矩形に切り取る。その後、112×112 になるように、バイキュービック法を用いて拡大縮小を行い、jpg 形式で保存する。

入力画像は RGB カラーのものを用い、96×96 のサイズになるようにランダムにクリッピングした後、明度差による影響を減らすために正規化を施した。

CNN の弱点である画像の傾きによる影響をなくすために顔の特徴点を利用し、常に同じ位置に特徴点が重なるようにアフィン変換を用いて傾きを補正した（図 3）[5]。

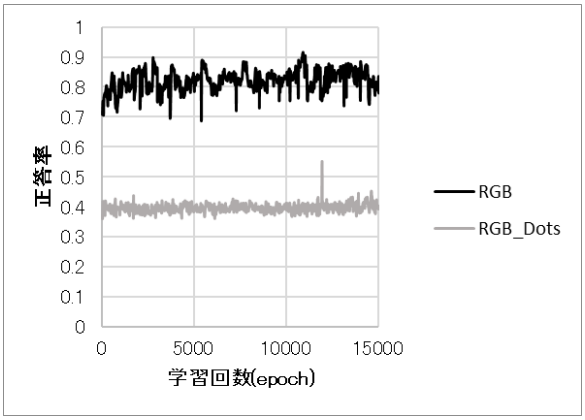

図 6 顔の特徴量隠蔽による影響

これ以降の実験では、2 クラス分類で性能評価する

5.3 実験 1：アフィン変換の効果

前処理としてアフィン変換を行った場合の効果を検証するために、アフィン変換の有無の認識精度を比較する実験を行った。

検証画像に対する結果を図 4 に示す。アフィン変換を行わなかった場合には約 55%の正答率であったが、行った場合は約 99%の正答率となった。

また、学習画像とは撮影環境の異なるテスト画像での結果を図 5 に示す。アフィン変換を行わなかった場合には、正答率は低いままであった。アフィン変換を行った場合の正答率は約 85%となり、検証画像に比べて正答率の低下がみられた。

以上のことから、本研究においてアフィン変換の有用性が確認できた。

5.4 実験 2：顔の特徴量隠蔽による影響

顔のどの部分が認識精度に影響を与えるのかを調べるために、両目と鼻領域が欠落した画像の認識精度について評価する。

アフィン変換を行った画像(図 3(b))と、その後に両目と鼻領域に白点でマーキングした画像(図 3(c))の 2 種類の画像を用い、実験では、テスト画像に対する正答率を調査する。

実験結果を図 6 に示す。白色マーキングを

表 2　ネットワーク構成 2

type	kernel_size	stride	out_channel
input			3
conv	3*3	1	64
conv	3*3	1	64
max_pool	2*2	2	
conv	3*3	1	128
conv	3*3	1	128
max_pool	2*2	2	
conv*3	3*3	1	256
max_pool	2*2	2	
conv*4	3*3	1	512
inception*3			
fc			1024
fc			256
output			5

表 3　ネットワーク構成 2 の Inception

	kernel_size	stride	out_channel	route
conv1	1*1	1	64	1
conv3	1*1	1	48	2
conv3	3*3	1	64	2
conv3	1*1	1	64	3
conv3	3*3	1	96	3
conv3	3*3	1	96	3
conv3	1*1	1	64	4
conv3	3*3	1	96	4
conv3	3*3	1	96	4
conv3	3*3	1	96	4
conv3	1*1	1	64	5
conv3	3*3	1	96	5
conv3	3*3	1	96	5
conv3	3*3	1	96	5
conv3	3*3	1	96	5
ave_pool	2*2	2		6
conv	1*1	1	32	6

しなかった場合の正答率は約 85%であったが、白色マーキングをした場合の正答率は 40%となり、約 45%低下した。この結果より、目と鼻の部分が重要な情報源であることがわかった。

6. 多クラス分類

実環境での応用を視野に入れ、クラス数を 5 に増やして実験を行う。

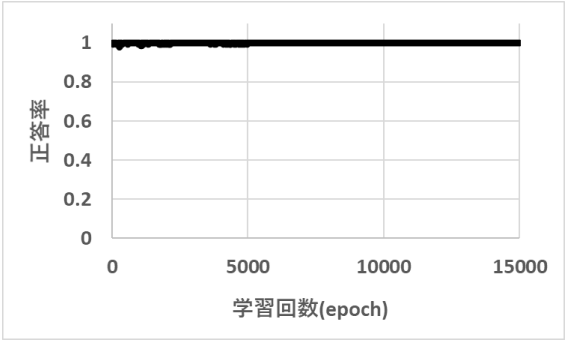

図 7　5 クラス検証用画像学習曲線

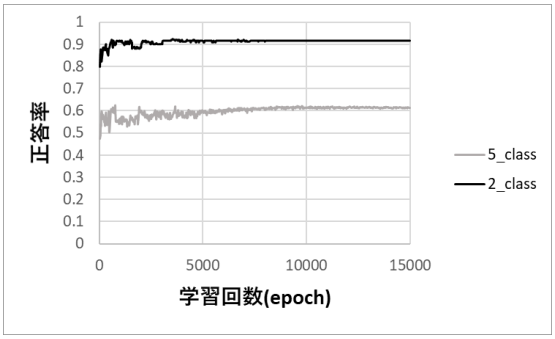

図 8　5 クラステスト用画像学習曲線

6.1 ネットワーク構成

多クラス分類においては精度向上の観点から前述の実験で用いたネットワーク構成を 23 層に改良したものを用いる(表 2, 表 3)。この改良にあたって畳み込みに用いるフィルタサイズを一種類に統一した。通常異なる特徴量を抽出するために異なるフィルタサイズの畳み込みを組み合わせるが、構造が複雑になってしまうという問題点があった。それを改良するためにフィルタサイズを 1x1、3×3 に固定し、畳み込み回数を重ねることでより大きなフィルタサイズの畳み込みの代用をすることができる。構造が簡単になるということだけではなく、この変更により同時に計算量も減らすことができることが証明されている[6]。使用するパラメータについては前述の実験と同様の値を用いる。

6.2 実験結果

図 7 に検証画像に対する学習曲線、図 8 に

表４　５クラス分類の分類結果

		推論値				
		A	B	C	D	E
真値	A	73.4	1.6	1.6	0.8	22.7
	B	4.0	73.4	11.3	0.8	10.5
	C	2.8	26.2	55.1	5.6	10.3
	D	2.7	19.5	6.2	68.1	3.5
	E	30.9	1.8	25.5	10.9	30.9

テスト画像に対する学習曲線を示す。図８には参考データとして同様の実験条件での２クラス分類の学習曲線も示す。学習画像と撮影環境が同一の検証画像では、ほぼ100%の正答率を示しているが、撮影環境の異なるテスト画像では最終的な正答率が 60%程度と大きく低下する結果となった。これは、2 クラス分類では正答率が約 90%であることから、約30%低下したことになる。推論結果を詳細に分析するため、A～E の５人のテスト画像に対するクラスごとの推論結果を表４に示す。表4 の対角部分が正答率を、その他の部分が誤答率を表す。

この表からわかるように特にＥの正答率が低い。これは本実験において、A と E は眼鏡をかけていたことが原因であると考える。このため、本来 E に分類されるべきところ、A に分類されてしまったと推測される。

また、D は日本人ではなく白人系アメリカ人であったため、高い正答率になると考えていた。しかし、実験結果を見ると正答率に大きな差はなかった。これは、鼻の高さや目の堀の深さは特徴量の違いとして学習されなかったためと考えられる。

一方、B の正答率が 73.4%と比較的高い。学習画像を比較してみると片目をつぶっている画像があったり、口の一部が指で隠れていたりなど特徴量の組み合わせのバリエーションが多彩だったことが一因だと考察する。

ところで、今回は画像データ収集を自動化したプログラムで行っていた。実験終了後に学習画像を確認したところ、明らかに顔画像の切り取り位置が異常な画像や、斜めに歪んでいたりする画像が一部に見受けられた。このため、このような画像をうまく取り除くことができれば、認識精度をさらに向上できるものと考えられる。

7. おわりに

今回は CNN を用いた２次元カラー画像による顔画像分類を行った。

まず、実験１の結果より、撮影した顔画像を学習にそのまま利用しても効果的に学習は進まなかったが、アフィン変換を行うことで劇的に学習精度が向上することが確認できた。しかし、アフィン変換はその特性上、どうしても画像が歪んでしまう場合がある。そのため、さらなる精度向上を狙うためには、正常な画像のみを選んで学習させる必要がある。一方、撮影環境の異なる画像で実験したところ、正答率が低下した。より精度を向上させる方法として考えられるのは学習画像のバリエーションを増加させることである。現在は椅子に座り、撮影プログラムのインストラクションに従って、顔を動かすことで多様な角度の学習画像を取得している。しかし、この方法ではカメラに対する顔の角度は変更できても、光源が変更できないためバリエーションが制限されてしまう。学習画像の収集方法については今後さらに検討する必要がある。

つぎに、実験２の結果より、目や鼻を白色のマーキングで隠蔽したところ、正答率が約45%低下した。このことから、目や鼻がディープラーニングを用いた顔認証において重要な情報源であることが確認できた。

実験３では、５クラス分類を行った。結果は、正答率が２クラス分類に比べて約 30%低下した。今後は、ネットワーク構成に関してもさらなる検討が必要である。

将来的に本技術は、高齢者一人暮らし家庭の入退室管理・オートロックシステムへの応用や、個人を識別することによって実現されるひとりひとりに寄り添ったパーソナルロボットへの応用が期待される。実社会への応用を目指すにあたり、認識精度のさらなる向上が望まれる。

文　献

(1) 岡谷貴之、深層学習、講談社、pp.1-110 (2015)

(2) C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions”, In Proc. the IEEE Conference on Computer Vision and Pattern Recognition, pages1-9, 2015

(3) C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, “Rethinking the Inception Architecture for Computer Vision”, arXiv preprint arXiv:1512.00567,2015

(4) F. Schroff, D. Kalenichenko, J. Philbin, “FaceNet:A Unified Embedding for Face Recognition and Clustering”, In Proc. CVPR, 2015.

(5) B. Amos, B. Ludwiczuk, M. Satyanarayanan, “OpenFace : A general-purpose face recognition library with mobile applications”, CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

(6) Karen Simonyan, Andrew Zisserman, “VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION”, arXiv preprint arXiv:1409.1556v6, 2015