ディープラーニングによる表情認識を利用したコミュニケーションロボットの研究

○楢木　翔*1　　入江　寿弘*2
新宮　清志*3

キーワード：AI　深層学習　画像認識　ロボット制御　機械工学

1.　はじめに

近年、プロセッサーの性能の向上やネットワークインフラの充実に伴い、機械学習に必要なデータが簡易に集められるようになった．また、ロボットの分野では、産業用ロボットにとどまらず人との対話を図るためのコミュンニケーションロボットの開発も盛んに行われている．

ロボットが人と同様の方法でコミュニケーション取るためには、人同士がコミュニケーションをする方法と同様に、表情から感情の推定を行う必要がある．そこで本稿では、ディープラーニングを用い表情の認識を行い、それに付随して機械的機構の制御の研究を行なっている．

2.　ディープラーニング

画像認識において用いられるディープラーニングのアルゴリズムは、畳み込みニューラルネットワーク(Convolutional Neural Network)が一般的である．現在では、それを応用して様々なアルゴリズムが考案されている．以下に本研究で用いたディープラーニングのアルゴリズムと、学習モデルの作成過程の詳細を示す．

2. 1 Faster R-CNN

研究では、2015 年に考案された一般物体検出用アルゴリズム Faster R-CNN[1)]を応用して、表情の認識を行う．図-1 にその全体構造の概略図を示す．

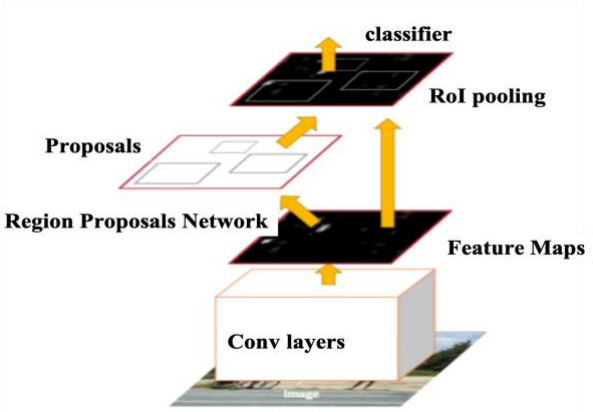

図-1 Faster R-CNN[1)]

Faster R-CNN は、上層に置かれた VGG16(図-2)と言うニューラルネットワーク(以下、NN)で畳み込みを行い画像の特徴マップを作成する．

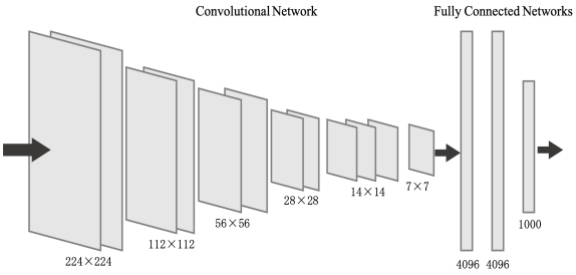

図-2 VGG16のネットワーク構造

作成した特徴マップは下位の層にある Region Proposal Network(以下、RPN)(図-3)に渡される．RPNは任意のサイズの入力から物体候補を出力するため、そのスコア(cls layer)と物体の領域(reg layer)の2つを同時に出力する．画像全体の特徴マップから事前に決められたk個の固定枠(Anchor)を用いて特徴を抽出し、RPNの入力とすることで、各画像位置において物体候補であるかの推定を行う．候補として推定された領域をプーリング層(RoI pooling)の入力としクラス識別のネットワークの入力とすることで最終的な物体検出を行う．RoI pooling層(図-1)は、SPPnet[2)]で考案された画像の入力サイズが一定でなくてもpoolingを行えるプーリング層である．

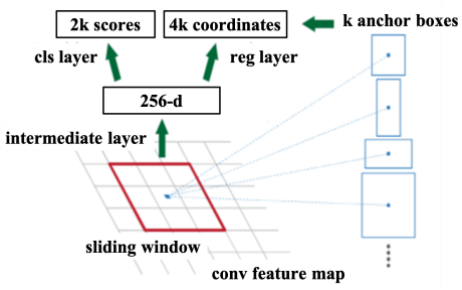

図-3 Region Proposal Network

2. 2 Alex Net[3)]

2. 1節で顔の候補領域の判定を行った後、AlexNet(図-4)と呼ばれるディープラーニングのアルゴリズムで表情の認識を行う.

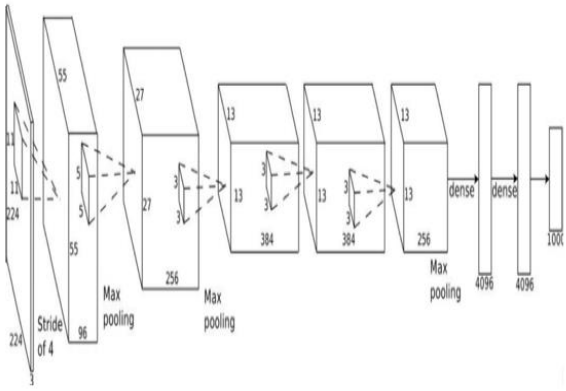

図-4 Alex Net

AlexNet は GPU の性能が高くない時に用いられていたアルゴリズムであることから、初期のネットワーク構造は GPU を２枚用いて学習を行えるように２層に分岐したような形であった．現在の GPU 性能であれば、１枚の GPU で学習を行えるため図-4 のような構造を持つ．本研究で用いた AlexNet の上層の畳み込み層は表情の認識を行う層であることから、次節で説明する表情データセットを使用して学習を行わせた.

2. 3 最適化手法

NNは算出した予測値と正解との誤差を算出し、その誤差の総和が最小になるように重みとバイアスを決定することで学習を行っていく．その誤差の総和を算出する手法を最適化と呼ぶ．本研究で使用した最適化を行う関数はMomentum SGD[4)](Stochastic Gradient Descent：確率的勾配降下法)と呼ばれるもので、以下にその原理を示す.

Momentum SGDとは、最適化手法の中でも初期に提唱された最も基本的なアルゴリズムであるSGDに慣性項(Momentum)を適用した手法である．重みの更新は、以下の式(1)のように行う．なお、αは慣性パラメータ、wは重み、Eは誤差関数、ηは学習係数を表す.

$$W_{t+1} \leftarrow W_t - \eta \frac{\partial E(W_t)}{\partial W_t} + \alpha \Delta W_t \qquad (1)$$

前回の更新量にもとの値αを加算することで、パラメータの更新をより慣性的なものにすることができる．なお、今回はηを0.005、αを0.9とした.

2. 4 表情データの学習

本研究では、画像データセットを2種類用いて学習を行う．個人でのデータセット収集には時間がかかる上に、学習に十分なデータ量を集めるのが困難なためである．そのため、表情認識においても二段階の認識を行うことになる.

まず、1つ目のデータセットは香港中文大學より提供されているWIDER FACE[5)](図-5)用いて学習を行う. その画像をテスト画像と学習用に分類して顔の認識行うための学習データを作成する.

図-5 香港中文大學 「WIDER FACE」

2つ目のデータセットは人の表情認識のために使用される．表情を「笑顔」「怒った顔」「泣き顔」「驚き」「その他」で定義し、それらの画像をgoogle画像検索などから収集する．「その他」には、一般物体を始めとする様々なデータを用意する．表-1にその画像枚数を示す.

表-1 各表情の画像枚数

データセット	画像枚数
笑顔	1996
怒った顔	1746
泣き顔	1080
驚き	828
その他	520
合計枚数	6170

3. 学習結果

表情のデータセットに関する学習結果を、以下に示す.

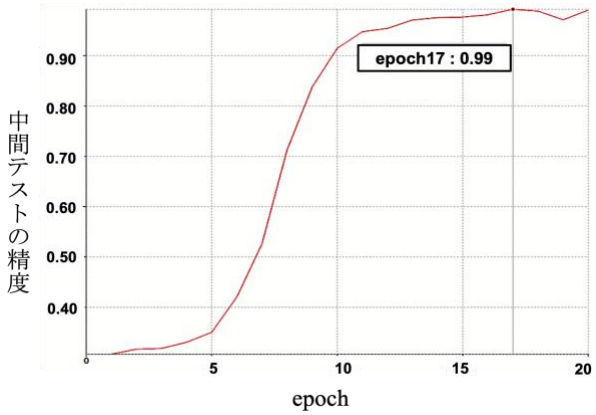

図-6 表情データセットの学習結果

図-6に示す学習データのパラメータは学習の中間テスト回数であるepoch数を20とし、最終的に最も精度の良い結果を実験に使用する．今回は17回目のデータで99.2[%]の精度を得られた為このデータを使用する．

4. 表情認識実験

4.1 実験概要

今回の実験は、本研究で用いた二つのネットワークによる顔の認識と表情の認識の二つの性能を評価するために行う．実験に使用した機器や開発環境については3.2節に、実験結果については3.3節に示す．

webカメラを用い画像の認識を行う際、人間同士の会話が行われるのに不自然でない距離を考え、800mm前後で実験を行った．

4.2 実験機器

本研究で用いた実験機器は、ディープラーニングを行う為の GPU に NVIDIA の GeForce シリーズ「GTX TitanX」[6]を使用した．この GPU の主な仕様は以下の表-2 に示す．

表-2 GPU スペック

CUDA コア数	3072
クロック数[MHz]	1075(最大)
メモリ[GB]	12

また、画像認識に用いるカメラに Buffalo の Web カメラを使用した．

ディープラーニングを実装する際に用いるフレームワークには Preferred Networks が開発している Chainer の画像認識ライブラリである「ChainerCV」を用いた．

５章の「サーボ制御」において、Tower Pro の Micro Servo 9g SG90 を使用した．また、サーボを制御するためにマイコンボード Arduino UNO を使用した．

4.3 実験結果

研究はデータを動画で取得している為、本節に示す図は実験結果の一部になる．

4.3.1 顔検出の精度

まず、顔の検出精度の結果を以下の図-7 に示す．

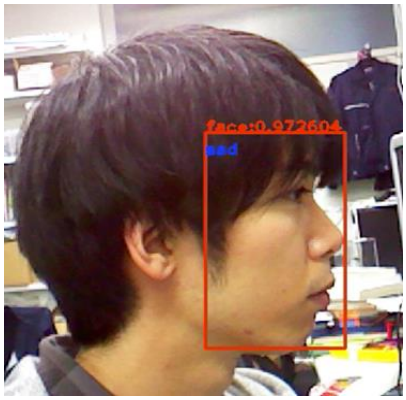

図-7 真横を向いている顔の認識精度

図-7 内の赤い矩形で囲まれた部分が顔として認識されており、その上部の「face：0.97…」と言う文字が顔である確率を表している．また、この際の表情の判定は「泣き顔」として認識されている．

4.3.2 表情の検出

次にリアルタイムでの表情の認識精度について検証を行った．その結果を以下の図-8 に示す．図-8 は紙に印刷された表情の画像を実験に使用した．対象物が静止している為、結果は動いている対象物の検出と比較して一定の値を示した．よって、各表情の検出結果を表-3 に示す．この画像の検出では「smiling」が最も大きな値を出力し、その値は 45%であった．また、「others(その他・無表情)」と笑顔の検出結果を合計すると 83%の結果を得ることとなり、良い結果を示したと思われる．

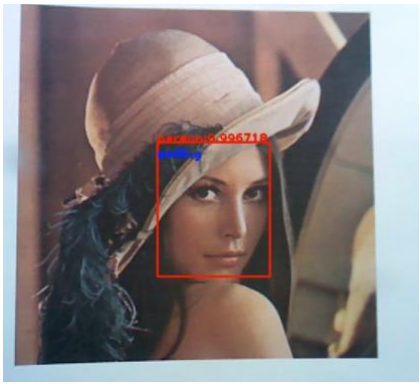

図-8 表情の認識結果

表-3 各表情の検出結果

表情	出力[%]
笑顔	45
泣き顔	1
怒った顔	2
驚き	16
その他・無表情	38

5. サーボ制御

前章より表情認識を用いた顔の識別を行った．本研究では、その顔認識を利用して人の顔を追従する機構を作成した．そのためのサーボ制御はディープラーニングによる顔認識及び表情判別と並行して処理を行う．そのため、プロセッサーへの負荷が大きくならないよう簡単な比例制御を用いる．

この機構ではカメラフレームのディスプレイ座標を考え、サーボの制御はx軸方向のみで行う．カメラフレームに顔の候補領域が検出された時、その矩形の左側をx1、右側をx2、中心座標をcxとする．カメラフレームのサイズは(480、640)に設定として以下の式で制御を行った．

図-9に機構の概略図を示す．また、図-10に実際に使用した機材の写真を示す．

$$cx = \frac{(x_2 - x_1)}{2} \tag{2}$$

$$\theta = 90 + \{30 \times \frac{(800 - cx)}{800}\} \tag{3}$$

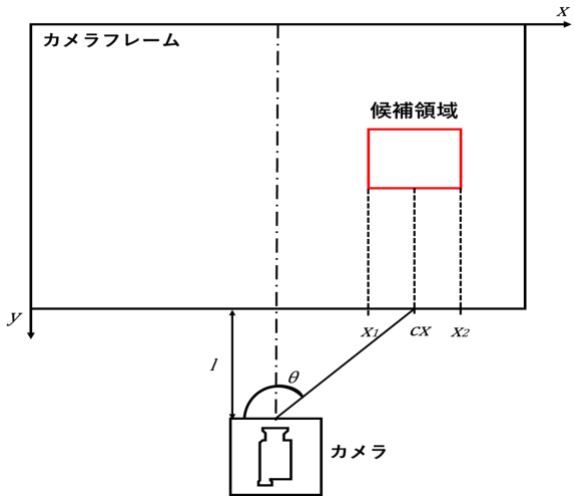

図-9 R/Cサーボモータ制御機構

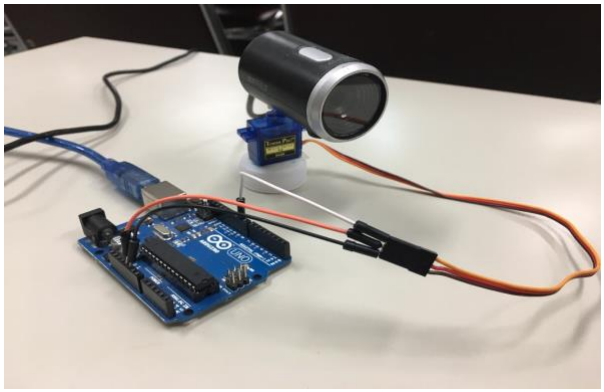

図-10 実際に使用した機材

4章で使用したウェブカメラによる表情認識システムを上記の機構に搭載し、顔を検出した際にそれを追従することで大きな範囲で顔認識を行えるようになった．

6. まとめ

ディープラーニングを用いて、独自のデータにより表情の判別を行った．

実験結果4. 3. 1における顔認証の精度の検証では学習に十分な量のデータセットを用いたため、高い出力を得ることができたと考えられる．

図-8の写真の女性は微笑みを浮かべている写真であり、今回の実験結果も笑顔に最も高い出力が得られた．

また、5章で作成したサーボ機構はフレームが垂直方向に動かせないことや簡単な比例制御ではフレームがぶれて誤認識を起こしてしまうことが問題点として挙げられる．今後は顔を追従する際に穏やかな挙動を示すように、新たな制御方法を考える必要がある．

また、ディープラーニングの新しいアルゴリズムを導入し、リアルタイムで表情判別を行う際の認識速度を向上させる予定である．

[参考文献]

1) Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, 2015.
2) Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. : Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, 2015.
3) Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, 2012
4) Ning Qian, On the Momentum Term in Gradient Descent Learning Algorithms, 1999
5) Yang, Shuo and Luo, Ping and Loy, Chen Change and Tang, Xiaoou. : WIDER FACE: A Face Detection Benchmark, 2016
6) https://www.nvidia.co.jp/object/geforce-gtx-titan-x-jp

＊1　日本大学大学院　理工学研究科　精密機械工学専攻　大学院生
＊2　日本大学　理工学部　精密機械工学科　教授　博士（工学）
＊3　日本大学　名誉教授　工学博士，総合資格学院　特別顧問