2M-01

畳み込みニューラルネットワークを用いた人物画像の認識と評価

細川皓平 † 川村秀憲 ‡
北海道大学情報科学研究科

1 はじめに

複数のネットワークカメラを利用した人物追跡より，人物の移動軌跡を求めることができる．この軌跡情報により，その領域における人物の行動パターンを解析するのに役立つ．また，カメラから取得した動画から分析するということは特殊な装置を用いることなく, 監視カメラなどを利用することで低コストで人物追跡をすることを可能とする．

画像から人物を抽出することだけを考えれば背景差分などの技術により比較的容易に実現可能であるが，ひとつのカメラの動画において複数の人物が映っている場合，それらを個別に追跡するには画像による人物判別や移動の予測などをする必要がある．さらに，複数のカメラを利用してより広範囲の追跡を考えると，カメラ同士で情報を共有して追跡していかなければならない．そこで，動画から人物画像を抽出しその人物画像を複数比較することにより追跡していくシステムを構築したい．

本研究ではこれらのシステムのうち, 抽出した人物画像に対しての判別をするシステムの構築を行う．また，この判別を実現する手法として，画像認識に広く応用されている畳み込みニューラルネットワーク (Convolutional Neural Network)[1] を使用する．

2 人物判別

本研究ではネットワークカメラから取得された画像のなかからすでに人物の画像が切り抜かれたと仮定し，複数の画像を比較して同一人物か否か判別するシステムを構築する．これを実現する初歩として，2 枚の人物画像を比較しそれらの人物が同一人物であるかを判別することを目標とする．

複数の人物の様々な状態，視点から撮影した画像を用意する．ここでは 30 人分の画像，累計 50000 枚を作成した．この 50000 枚の画像からランダムにペアを作り，それらの画像の人物が同一人物であるか否かを学習する．ただし，服装が違う場合は同一人物でも違う人物として扱う．

2.1 データセットの作成

データセットの作成にはビデオカメラを用い，撮影した動画を画像に分解した．判別する画像はすでに人物の部分を切り取ったものと仮定するため，データセットとする画像も全体が人物を収めたものとした．この際，立つ，歩く，座る，といった動きをすることによって判別に必要なパターンを網羅するように撮影した．動画の長さは一人あたり 1 分とし，これを 30 人分撮影した．なお，撮影には大学生の男子 9 名に協力してもらった．動画のフレームレートは 30fps であるため，一人あたり 2000 枚，全体で 60000 枚の画像となった．また，前処理としてすべての画像を 128x256 に縮小した．

2.2 ネットワークの構築

ここでは Convolutional Neural Network (CNN) を利用した学習を考える．CNN は大規模画像認識のコンテストである, IMAGENET Large Scale Visual Recognition Challenge(ILSVRC) において上位を独占しているアルゴリズムであり [2]，画像認識において最も有力な手段の一つとされている．

学習をするにあたって, DeepLearning 用フレームワークである chainer[3] を利用して実装した．chainer は python のライブラリであり，DeepLearning の実装を容易にする他，CUDA[4] を利用した GPU プログラミングを可能とすることで効率的な学習を実現する．ネットワークの具体的な構成は Fig.1 に示す．このネットワークは AlexNet[5] を参考に構築した．

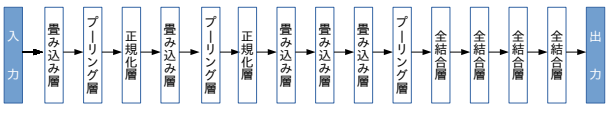

図 1 ネットワークの概要

入力には 128x256 の画像 2 枚を 256x256 の 1 枚の画像に結合したものを使用した．出力は 2 枚の画像が同一人物のものである確率，違う人物のものである確率の 2 つの値とした．

Human Recognition using Convolutional Neural Networks
†Kouhei Hosokawa, Hokkaido University
‡Hidenori Kawamura, Hokkaido University

2.3 実験

2.3.1 学習

データセットの 30 人分の画像のうち 20 人分の画像 37200 枚を訓練データ，10 人分の画像 17100 枚をテストデータとして使用した．学習は各画像ごとに 10 回 (10epoch) ずつ行った．この時，各 epoch ごとに画像を読み込む際にランダムな位置で 113x226 のサイズにトリミングをすることでデータ拡張を行う．ここでは画像を 8 枚読みこむごとにミニバッチを作成した．これにより 226x226 の画像 28 枚をひとつのミニバッチとした．学習過程における誤差率の遷移は Fig.2 に示す．

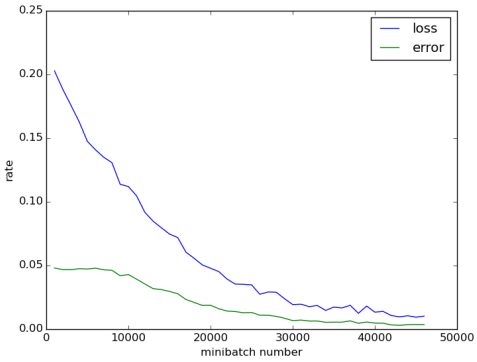

図 2　誤差率の遷移

横軸は学習したミニバッチの数 (更新回数) を表しており，縦軸の loss は出力層を通した値，error は出力層を通していない値に対する誤差率を表している．これにより学習が進んでいること，収束していることがわかる．

2.3.2 結果

2.3.1 で学習したネットワークを利用して，テストデータを用いた分類を行った．テストデータをこのネットワークに入力することで，2 枚の画像が同一人物のものである確率，違う人物のものである確率がわかる．この 2 つの確率のうち高い方を判別結果とし，正解データと比較する．2.3.1 にあるようにテストデータには 17100 枚の画像を使用したが，この際同じクラスの画像の組み合わせと違うクラスの画像の組み合わせが同一となるようにした。

表 1　テスト結果

		真の結果	
		True	False
予測結果	True	12119	1667
	False	14638	25090

表 1 の結果を元に評価した結果が表 2 である。

表 2

精度	(accuracy)	69.5%
適合率	(precision)	88.0%
再現率	(recall)	45.3%
F 値	(F-measure)	0.598

表 2 より，適合率と比較して再現率が非常に低いことがわかる。これは表 1 を見てもわかるように，真の結果が False の際の精度は非常に高いが，True の際は精度が低いことを表している．これは学習に使用された画像の内，違うクラス同士の画像の組み合わせが同じクラス同士の画像の組み合わせと比較して非常に多くなってしまっていることが問題であると考えられる．

3 まとめ

今回はネットワークカメラでの人物追跡の初歩として 2 枚の人物画像の判別を行い，精度としては約 70% という高い正答率を得ることができた．しかし，再現率の点から見ると 2 つの画像の組み合わせが同一である場合の精度が低い結果となってしまった．そのため，今後は学習の際の 2 つの画像の組み合わせが同一のものと違うもので数を揃える工夫が必要だろう．また，今回の実験で使用したデータセットは限れられた人物のものであることや同じ環境で撮影されたものである．そのため，より様々な人物や環境ではどのような結果となるかも調査する必要がある．

参考文献

[1] Y.LeCun, L.Bottou, Y.Bengio and P.Haffner. Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE, 86(11):2278-2324, 1998

[2] Results of ILSVRC2014, http://image-net.org/challenges/LSVRC/2014/results.

[3] Chainer, http://chainer.org/

[4] nVIDIA CUDA ZONE, https://developer.nvidia.com/cuda-zone

[5] A.Krizhevsky, I.Sutskever and G.E.Hinton. ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012