Vol.2016-CVIM-201 No.6
2016/3/3

動画像認識のための 3 次元畳み込み RNN の提案

浅谷学嗣[†1]　田川聖一[†1]　新岡宏彦[†1]　三宅淳[†1]

概要：機械学習の一種である Deep Learning が画像認識や音声認識など様々な認識タスクにおいて，高い精度を発揮している．画像認識においては，Convolutional Neural Network が最も高い認識精度を示しており，動画認識への応用が期待されている．音声認識や自然言語処理など，可変長なデータに対しては Recurrent Neural Network が最適なモデルとして提案されている．本研究は，動画認識において end-to-end に学習を行う事が可能な 3D Convolutional Recurrent Neural Network を提案する．本手法を用いて 6 種類の人間の動作が格納された KTH のデータベースを分類したところ，86%の認識精度を実現した．

キーワード：Deep Learning, Machine Learning, Convolutional Neural Network, Recurrent Neural Network

Proposal of 3D-Conv-RNN for movie recognition

SATOSHI ASATANI[†1]　SEIICHI TAGAWA[†1]
HIROHIKO NIIOKA[†1]　JUN MIYAKE[†1]

Abstract: Deep Learning has achieved high recognition rates in various tasks such as image recognition and voice recognition. Convolutional Neural Network (CNN) has shown the best score of image recognition, and application of CNN has been expected in the movie recognition. Recurrent Neural Network (RNN) has been proposed as the best model that can recognize variable length data, such as voice recognition and Natural Language Processing. We propose 3D Convolutional Recurrent Neural Network to recognize movies by using end-to-end learning. When classifying KTH dataset, movie recognition precision of our model was 86%.

Keywords: Deep Learning, Machine Learning, Convolutional Neural Network, Recurrent Neural Network

1. はじめに

防犯を目的とした監視カメラによる人間の行動認識や，自動運転を目指した車載カメラによる周囲の環境認識など，動画像を用いた認識技術が様々な場面で応用され始めている．動画像を認識するための手法として機械学習が一般的に用いられるが，集められるデータはビッグデータとなる．ビッグデータの学習は，全体を自動化する事が望ましい．また，集められた動画像は長さが一定ではなく，可変長な動画に対して学習する手法が必要とされる．従来の動画認識技術では，時間毎の画像に対して人間が設定した特徴に応じたスコアを算出し，その時間的な変化をもとに教師あり学習を用いて認識する手法が提案されている[1]．このような手法では，用いた特徴量に分類性能が依存してしまい，未知の特徴に対応できない．また，考慮すべき時間的変化量も人間が設定したタイムスケールの範囲に限定され，認識できる時間幅が制限される．動画像認識を行うためには，時間毎に算出された特徴量を用いるのではなく，時間変化を考慮した特徴量を用いる事が重要であると考える．また，学習の自動化に適した特徴抽出部と識別部が一体となった end-to-end な学習も求められる．

画像認識や音声認識の分野で高い認識精度を実現する Deep Learning は，中間層を多層にしたニューラルネットワークである．Deep Learning はデータ構造にしたがって最適な構造が提案されている．画像認識においては人間の視覚野の機能を模擬したモデルである Convolutional Neural Network (CNN)が用いられる．CNN は画像を認識する際に用いられる特徴表現を，学習により獲得する事が知られている．音声認識においては，入力や中間層の時間的な影響をモデル化したニューラルネットワークである Recurrent Neural Network (RNN)が用いられている．RNN は前時刻の影響を考慮しながら時系列情報を認識する事が可能であり，可変長な入力に対応できるため，音声認識だけでなく自然言語処理などにも応用されている[2].

本研究では，画像認識に適用される CNN と時系列情報に最適な RNN を組み合わせ，end-to-end に教師あり学習する事で動画像認識を可能とする 3 次元畳み込み RNN を提案する．

2. 関連研究

動画像認識の研究は，動作認識や意味解析などの研究が多く，SIFT や SURF などの人間が設定した特徴量を用いた手法や，Deep Learning を応用した手法が提案されている．

†1 大阪大学
Osaka University

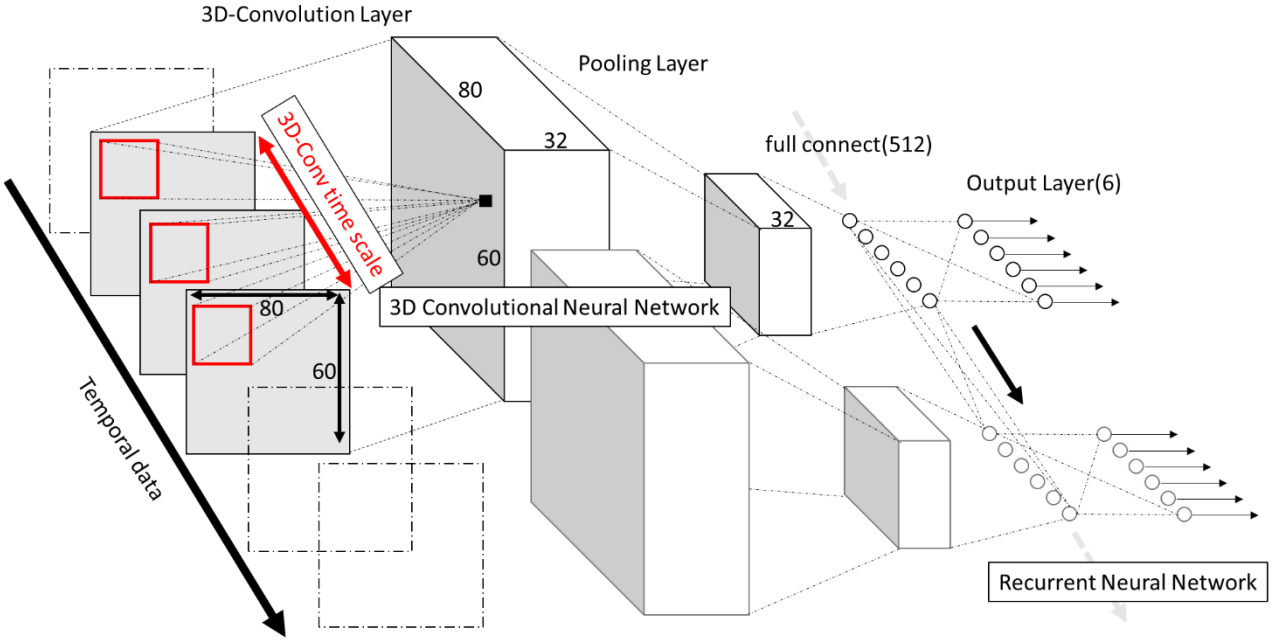

図 1　3D-Convolutional Recurrent Neural Network のネットワーク構造

前者の手法では，時空間特徴を抽出するために，画像認識に用いられている特徴を 3 次元に拡張する研究が行われている．Laptev らはハリス点検出を 3 次元的に拡張した検出手法[3]を提案し，野口らは SURF 特徴量の時間変化を時空間特徴とする手法[1]を提案している．これらの時空間特徴量を SVM などの分類器を用いて分類することで，動画認識を可能とした．

Deep Learning を動画像認識に応用した手法としては，3D-Convolutional Neural Network を用いた手法が提案されている[4]. 3D-Convolutional Neural Network は画像認識に応用される CNN の Convolution 層を 3 次元的に拡張したモデルである．Baccouche らは 3D-Convolutional Neural Network を用いて動画から時空間特徴を抽出し，その特徴記述ベクトルを LSTM などの機械学習手法により分類した[5].

これらの手法は，動画像から特徴を抽出する機構と抽出された特徴量を用いて分類する機構が分離しており，end-to-end に学習を行う事ができない．また，時間変化を考慮できる時間幅は，人間の設定に依存するため制約がある．

3. 提案手法

本研究で提案するニューラルネットワークは，CNN で用いられる Convolution 層のノードが時間軸方向にも結合した 3D-Convolution 層，画像の位置変動に対応するための Pooling 層，時系列の情報に対して認識を行う RNN が結合した構造である(図 1). ネットワーク前部は，動画像における短期的な画像変化の特徴を捉える特徴抽出部として機能し，後部の RNN は時系列変化に対応した分類器として機能する．これらを組み合わせる事で，短期的にも長期的にも画像変化を考慮した動画像認識が可能であると考える．

3.1 3D-Convolution 層

CNN 内の Convolution 層には入力データに対して局所的に結合したノード(フィルタ(図 2))が複数存在し，学習の過程で画像内の局所的な特徴を捉える事ができる(図 2(a)). 3D-Convolution 層は，任意の時刻 T だけの画像に結合するだけではなく，一定の時間幅内の画像すべてに対して局所的に結合したノードをもつ(図 2 (b)). 3D-Convolution 層内のノードは，入力画像データの時間的変化に対する特徴を，特徴表現として獲得することが可能である．

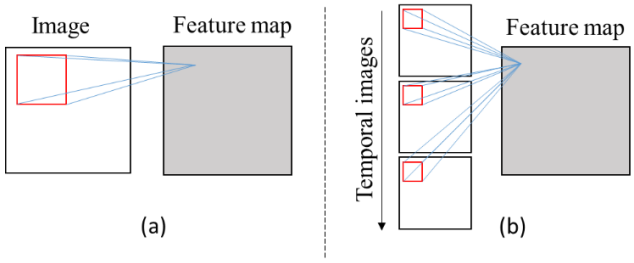

図 2　Convolution フィルタ(a)と 3D-Convolution フィルタ(b)

画像の縦横サイズが W×H，チャネル数が K の時，W×H 画素で K 枚の画像の形をとっている．以後 W×H×K と表記する．グレースケール画像の場合は K＝1 であり，RGB 画像の場合は K＝3 である．動画像はこの画像が連続的につながった画像群である．3D-Convolution 層内の任意の時間における画像に畳み込むフィルタサイズを S×S とした時，フィルタは画像内の局所領域 S×S×K に結合している．さらに，3D-Convolution 層内のフィルタは時間幅 T フレーム内の画像すべてにおいても同様の領域に結合している．これは画像内の時間変化を捉えるためであり，3D-Convolution 層内のフィルタは(S×S)×K×T のサイズをも

って畳み込み処理を行う事になる.

3.2 Pooling 層

CNN 内の Pooling 層は, 位置感度を低下させる事で, 画像特徴の微小な位置ずれに対する応答の不変性を実現するための層である. Pooling 層も Convolution 層と同様に, 局所的に結合したノードを持っており, 入力に対して一つの値を出力する. 本研究では入力の最大値を出力する max pooling を用いる. 特徴の位置ずれの補正し, より強い特徴を次の層に順伝播する.

Pooling 層の入力は 3D-Convolution 層の出力と結合しており, 画像中の位置感度を低下させ, 画像変化の微妙な位置ずれに対応する. Pooling 層のフィルタは, 1 画素ずつではなく数画素ずつずらして適応する場合があり, その適応間隔をストライド呼ぶ. ストライドを 2 以上にする事で次元の削減が行え, 計算時間を短縮することが可能である.

3.3 Recurrent Neural Network

3D-Convolution 層で抽出された特徴量には時間変化が考慮されているものの, フィルタの時間方向のサイズに依存してしまい, フィルタサイズよりも長い動作等を認識するためには, それらの特徴量を組み合わせて認識を行わなければならない. RNN は中間層のノードの出力が次の時刻の入力となる構造(図 3(a))を持っており, 時系列に入力される情報を, 前の情報を考慮しながら分類や認識を行う事ができる.

3.4 ネットワークの学習

ネットワークの学習手法は, RNN の学習手法の一つである Back Propagation Through Time(BPTT)を用いた. BPTT はニューラルネットワークの学習手法である Back Propagation を時間方向に拡張したもので, 最終時刻の出力層から順に誤差を入力層に伝播していく手法である. RNN を時間方向に展開することで, 順伝播型ネットワークに置き換え, Back Propagation を行っている(図 3(b)). 本手法では RNN の入力層と Pooling 層の出力層が結合しており, CNN の全結合層が時間方向に展開した構造を有している.

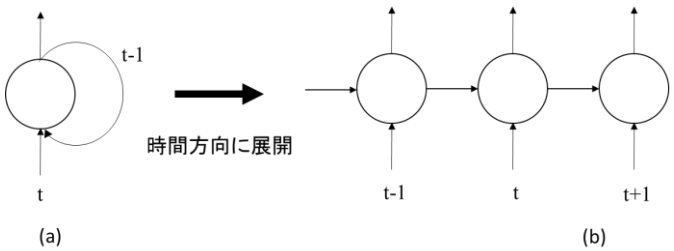

図 3 Recurrent Neural Network の構造

4. 評価実験

3 次元畳み込み RNN の認識性能を評価するために, 人間の動作認識を行った.

4.1 データセット

動作認識の研究分野において広く利用されている KTH データセットを用いた. KTH データセットは 6 種類の動作 (boxing, handclapping, handwaving, jogging, running, walking) の動画像がそれぞれ 100 データ程度格納されている(図 4). KTH 内の動画は, カメラが固定されており, 動作を行っている人が一人だけしか写っていない. 動画内の画像は, 160×120@25[fps]で撮影されたグレースケール画像である. 本実験では, 計算時間の短縮のため画像サイズを 80×60 に圧縮した. 交差検証を行うため, 訓練データは 80 データ, テストデータは 20 データとしたデータセットを各動作につき 5 セット用意した.

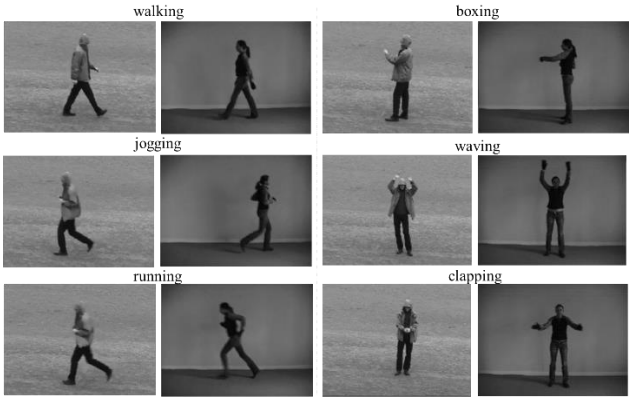

図 4 KTH 内の 6 種類の動作

4.2 実験環境

本実験環境は, 演算処理の高速化のため GPU (NVIDIA GTX 980 Ti)を用いた. CPU は Intel 社製 Core i7-4790K, メモリは 8GB である. 学習に用いたソフトウェアは, 我々が独自に開発・実装した Deep Learning ライブラリ Sigma を用いた. 実装には C++と NVIDIA が提供する GPU 向けの C 言語統合開発環境である CUDA を用いた.

4.3 構築したネットワーク

入力層は入力画像のサイズに依存し, 本実験では 80×60 のノード数となる. 第 2 層の 3D-Convolution 層のフレーム幅を持ったフィルタサイズは 5×5×1×6, フィルタ数は 32 とし, 活性化関数は ReLU [6]を用いた. 第 3 層の Pooling 層は, フィルタサイズを 5×5, ストライドを 3 とした. 第 4 層である RNN の中間層のノード数は 512 とし, 活性化関数はシグモイド関数を用いた. 出力層は, 分類すべきクラス数と同数の 6 ノードとし, 出力関数はソフトマックス関数を用いた.

入力のフレーム数は動画のフレーム数に依存し, 一つの訓練動画サンプルすべてのフレームを入力した後, BPTT を適応しノードの重みを更新した.

4.4 認識精度の結果と考察

交差検証を行った結果を表 1 に示す. 全身を動かす walking/jogging/running(前進動作)と, 身体の一部を動かす

boxing/waving/clapping を誤認識する事は無かった．誤認識している割合に注目すると，walking/jogging/running など類似している動作はお互いを誤認識する割合が高くなった．また，前進動作においては速度が近い動作と誤認識する確率が高い事がわかる．これは速度変化を特徴とした分類基準が，学習によって形成されていると考えられる．

本手法と他の認識手法の Accuracy を比較したものを表 2 に示す. 3D-Convolutional Neural Network のみを用いた手法(Shuiwang et al.)や，SURF 特徴量を用いた手法(Noguchi et al.)には及ばなかった．本提案手法は他の手法に比べて設定すべきパラメータが多く，ネットワーク構造の最適化が難しい．設定すべきパラメータは各層の層数やノード数などの構造に関係するパラメータだけでなく，入力動画のフレーム数や学習回数など構造に関係ないパラメータも重要となる. CNN では Convolution 層を積層化することで，より大域的な特徴を抽出できる事が知られており，2012 年の ILSVRC で優勝した CNN は 5 つの畳み込み層を持っていた．本手法でも，複数の 3D-Convolution 層を用いる事で，大域的な画像変化特徴を抽出でき，更なる分類性能向上が期待される．訓練データに関しては，人がカメラのフレーム外に出てしまい，画像変化がないフレームが多数存在していたため，学習の妨げになっていた可能性が考えられる．人が写っていないフレームを除去することで，学習効率を向上させられると考える．

	walking	jogging	running	boxing	waving	clapping
walking	0.76	0.16	0.08	0	0	0
jogging	0.09	0.81	0.1	0	0	0
running	0.06	0.12	0.82	0	0	0
boxing	0	0	0	0.9	0.03	0.07
waving	0	0	0	0.07	0.92	0.01
clapping	0	0	0	0.01	0.02	0.95

表 1　各動作に対する Accuracy

Method	walking	jogging	runnning	boxing	waving	clapping	Average
Our method	76	81	82	90	92	95	86
Shuiwang et al.[4]	97	84	79	90	97	94	90.2
Noguchi et al.[1]	99	94	85	96	98	93	94
Dollar et al.[7]	90	57	85	93	85	77	81.2
Niebles et al.[8]	82	53	88	98	93	86	83.3

表 2　他のモデルとの Accuracy の比較

5. 結論と今後の課題

3D Convolutional Neural Network と Recurrent Neural Network を結合したニューラルネットワークを用いる事で，動画認識可能なモデルを提案した．他の認識手法技術は特徴抽出部と分類部が分離しているが，本手法は動画像を end-to-end に教師あり学習可能な手法である．最新の分類手法よりは認識精度が低いものの，入力動画のフレーム数の制約が無く，学習の自動化との親和性が高い，応用範囲の広い技術であると考える．

また，3D-Convolution と RNN を組み合わせる事で，短期的な動作特徴だけでなく，長期的な動作特徴も学習できると考えている．様々な動きが連続した動作の認識など，より複雑な動作認識が求められる場面で応用し，更なる性能向上を目指したパラメータの最適化手法を検討する．

参考文献

[1] 野口顕嗣, et al. 動作認識のための時空間特徴量と特徴統合手法の提案. 画像の認識・理解シンポジウム (MIRU 2010). July 2010
[2] Ilya Sutskecer, et al. . Sequence to Sequence Learning with Neural Networks. Neural Information Processing Systems (NIPS). 2014.
[3] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *Proc. of IEEE Inter-national Conference on Computer Vision*, 2003
[4] Shuiwang Ji, et al. . 3D Convolutional Neural Networks for Human Action Recognition. Pattern Analysis and Machine Intelligence, vol. 35, No. 1, January 2013.
[5] Moez Baccouche, et al. . Sequential Deep Learning for Human Action Recognition. Human Behavior Understanding 2011, Lecture Notes in Computer Science 7065, pp. 29-39, 2011.
[6] V Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proc. International Conference on Machine Learning (ICML)*, 2010.
[7] P. Dollar, et al. . Behavior Recognition via Sparse Spatio-Temporal Features. In *Proc. IEEE Int'l Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65-72. 2005.
[8] J. C. Niebles, et al. . Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. Int'l J. Computer Vision, vol. 79, No. 3, pp. 299-318, 2008.