Vol.2017-CVIM-206 No.17
2017/3/10

3D畳み込みニューラルネットワーク及び LSTM を用いた動画像認識

勝田 裕貴　　斎藤 博昭
慶應義塾大学大学院 理工学研究科
{katsuta, hxs}@nak.ics.keio.ac.jp

概 要

動画像認識ではその特徴量は画像の縦横軸に加えて，時間軸を考慮する必要があるため，時間変化を考慮した特徴量を用いることが重要である．一方，深層学習は画像認識や音声認識で高い認識精度を誇っており，学習過程において特徴の自動設計がなされる．深層学習技術を動画像認識に適応する際，画像の縦横軸及び時間軸方向に畳み込み計算を行う 3D Convolutional Neural Network(3D-CNN) による手法がある．また，音声認識等では時系列データを扱う Long-Short Term Memory(LSTM) によって時間変化を考慮した認識を可能としている．本論文では，3D-CNN により特徴抽出を行った後，LSTM に接続することでより長時間の動画像の認識を可能とするシステムを提案する．実験では，行動認識の分野で用いられる 25 人による 6 行動の動画像からなる KTH データセットを用いて行い，提案手法は 93.4%の精度となった．

1 はじめに

動画像を認識するための手法として機械学習が用いられるが，従来の動画像認識技術では，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．例えば動画像から局所特徴量を抽出して行動認識を行う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が依存されてしまい，扱う問題ごとに適切な特徴量の設計が求められる．

一方，動画像認識や音声認識の分野で高い認識精度を実現する深層学習は，中間層を多層にしたニューラルネットワークである．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) が高い精度を誇っている [2]．CNN は二次元カーネル用いて元の画像を畳み込み計算を行う畳込み層と，特徴を汲み取るプーリング層を多層に繰り返したあと，全結合層により分類を行う構造をしている．これにより画像を認識する際に用いられる特徴表現を，学習によって獲得することが知られている．また，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ことが可能であり，可変長な入力に対応できる．また，RNN の拡張として登場した Long-Short Term Memory(LSTM) は従来の RNN では学習できなかった長期依存の学習を可能となる [3]．

深層学習を動画像認識に応用する場合，CNN の畳み込み計算を時間軸方向にも行う 3D Convolutional Neural Network(3D-CNN) が提案されており，時間変化を捉えた特徴を抽出することができる [4]．本研究では 3D-CNN と LSTM を組み合わせたモデルを提案し，3D-CNN による特徴抽出ののち LSTM による長期的な時間変化を捉えた動画像認識を実現する．

2 関連研究

動画像認識の研究は，動作認識や意味解析の研究が多く人間が設定した特徴量を用いる手法や深層学習を用いた手法などがある．小林らは自己相関関数を三次元に拡張した Cubic Higher-order Local Auto-Correlation(CHLAC) を提案した [5]．また，野口らは SURF 特徴量の時間変化を時空間特徴とする手法を提案した [6]．

深層学習を用いた手法として Shuiwang らは 3D-CNN を提案し，CNN を三次元拡張したモデルを用いて動画像認識を行った [4]．また，浅谷らは 3D-CNN に RNN を組み合わせた 3 次元畳み込み RNN を提案し，3D-CNN による短期的な時間変化と RNN による長期的な時間変化を捉えたモデルとなっている．

深層学習を用いた手法は特徴の自動設計が行われる点で画期的であり，さまざまな応用が考えられる．

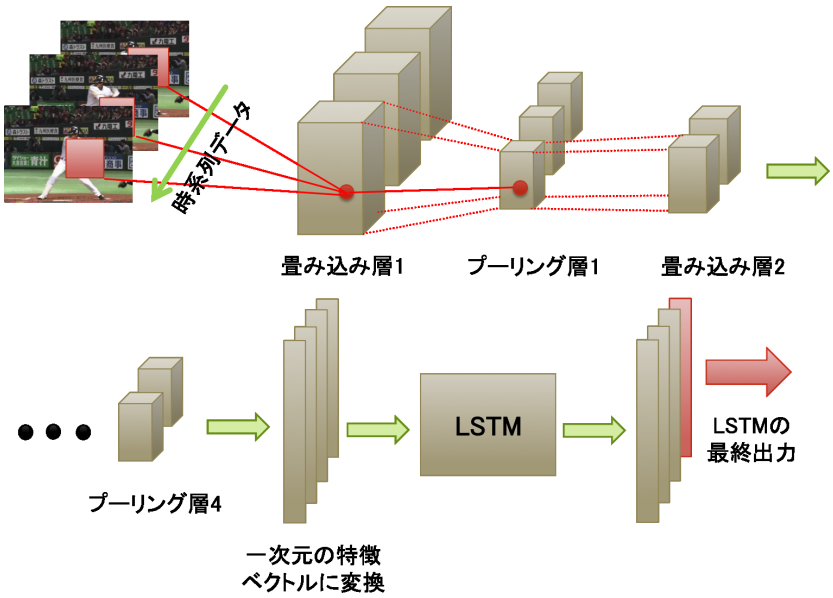

図 1: 提案手法の基本構成

3 提案手法

動画像認識の手法は従来の人手で特徴を設計する手法に対し，深層学習の手法は特徴を自動で計算できる点で勝り，深層学習を用いた高性能な手法が求められる．本研究では 3D-CNN によって特徴抽出した後 LSTM による分類を行うネットワークを提案し長期的な時間変化を捉えた動画像認識を実現する．提案手法は構成は図 1 のようになっており、以下の手順からなる．

1. 学習データの前処理
2. 3D-CNN のカーネルの事前学習
3. 学習されたカーネルを用いて特徴マップを抽出
4. 特徴マップを入力とした LSTM による分類

3.1 学習データの前処理

CNN 等のクラス分類の精度を向上させるために十分な学習データ数が必要である．学習データにコントラスト調整やガウス分布に基づくノイズを付加し画像の水増しを行う．

3.2 3D-CNN

3D-CNN は縦・横・時間軸方向に畳み込み計算を行う畳み込み層，プーリング層，全結合層の構成で成り立つ．これにより時間軸を持つ動画像の短期的な画像変化を捉える特徴を抽出することができる．ネットワークは畳み込み層とプーリング層を多層に重ねることで高い認識性能を実現することができる．本提案手法では事前に 3D-CNN で学習を行った後、学習された 3D-CNN の全結合層前の層の出力を 3D-CNN の特徴抽出をし、その特徴を LSTM によって分類を行う構成をしている．

3.2.1 畳み込み層

3D-CNN の畳み込み層ではカーネルの次元を一つ増やして時間方向の畳み込み計算を行えるようにしたものである．入力は縦横のサイズが $W \times H$，チャネル数が K の画像が T 枚のとき，$W \times H \times T \times K$ となり，フィルタのサイズは $S_x \times S_y \times S_t \times N$ となる．畳み込み後の画素値 $u_{i,j,k}$ は，

$$u_{i,j,k} = \sum_{p=1}^{S}\sum_{q=1}^{S}\sum_{r=1}^{S} x_{(i+p),(j+q),(k+l)} f_{p,q,r} + b \quad (1)$$

と計算される．

3.2.2 プーリング層

プーリング層は，位置感度を低下させることで画像特徴の微小な位置ずれに対する不変性を実現するための層である．プーリングの方法には各小領域に属する画素の値の平均をとる「平均化」プーリング，各小領

域内から画素値が最大のものを用いる最大化プーリングなどがある．通常 2 以上で設定されたストライド数 s によるプーリングで出力される画像サイズは縦横それぞれ $1/s$ 倍され $W' \times H' \times N$ となる．

3.2.3 全結合層

畳み込み層とプーリング層を多層に渡り繰り返した後，出力のすべての画素値 x_m と n 個のノードからなる次の層 y_n を全て結合する．これらの重み付き和にバイアスを加算した

$$y_j = \sum_{i=1}^{m} w_{ij} x_i + b_j \tag{2}$$

によって計算される。

3.3 LSTM

LSTM は時系列データを扱うモデルであり，従来の RNN では学習が困難であった長期期間の記憶を可能としている．カーネルを学習させた 3D-CNN から抽出された特徴を入力とし，LSTM による長期時間変化を捉えた動画像認識を実現する．

3.3.1 LSTM の構造

LSTM は中間層が再帰構造を持つニューラルネットである RNN を拡張したものであり，長期期間の共起関係を学習する際に，勾配の極端な縮小・拡大を避けるために，再帰的な入力を受け付けるノードに関しては高等関数の活性化関数を用いる．また，再帰的な入力に関する重み行列を単位行列に固定する．

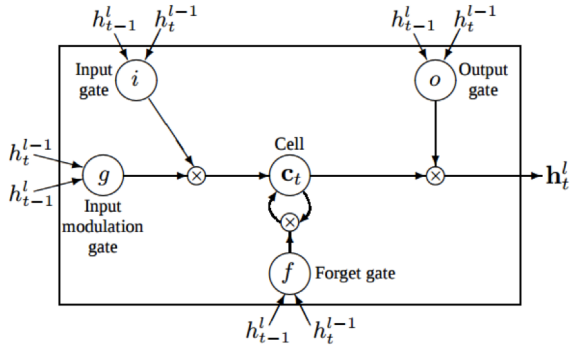

図 2: LSTM の構成 [7]

LSTM は図 2 のような構成をしており，入力判断ゲート，忘却判断ゲート，出力判断ゲートの判断ノードを導入して表現力を向上させている．判断ゲートは前時刻からの加算処理を行うか行わないか，また隣接層から入力の加算を行うか行わないか，出力を行うか行わないかを現在/直前のメモリノードの値や隣接層からの入力ベクトルをもとに判断するシグモイドノードである．LSTM は次のような式で表すことができる [7].

$$LSTM : h_t^{l-1}, h_{t-1}^l, c_{t-1}^l \rightarrow h_t^l, c_t^l$$

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} sigm \\ sigm \\ sigm \\ sigm \end{pmatrix} T_{2n,4n} \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix} \tag{3}$$

$$c_t^l = f \odot c_{t-1}^l + i \odot g \tag{4}$$

$$h_t^l = o \odot \tanh(c_t^l) \tag{5}$$

ここで，$\odot$ は行列の各成分の積であり，h_t^l は時刻 t における l 番目の隠れ層の出力，c_t は過去の状態を記憶するメモリセルを表し，$T_{2n,4n}$ は h_t^{l-1}，h_{t-1}^l から入力及び三つの判断ゲートへの写像である．各判断ゲートのパラメータは誤差逆伝播法によって学習される．

4 評価実験

4.1 データセット

提案手法の評価は動作認識の分野で広く利用されている KTH データセットを用いた [8]．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おり，画像サイズは 160 × 120 の 1 チャネルである．本実験では画像サイズを 80 × 60 × 1 に圧縮した後，動画像から取り出した 80 フレームを一つの学習データとした．

4.2 ネットワーク構成

入力層は 80 × 60 × ×1 × 80 のノード数を持つ．第 1 層，第 3 層，第 5 層，第 7 層が畳み込み層となっており，それぞれのカーネルサイズは第 1 層目では 3 × 3 × ×3 × 1 のフィルタを 20 枚とし，3 層目以降は 3 × 3 × ×3 × 20 のフィルタを 20 枚ずつ用意し，それぞれの層に活性化関数 ReLU を用いて出力を行った．第 2 層，第 4 層，第 6 層，第 8 層はプーリング層である．第 8 層の出力を隠れ層 25 の LSTM の入力とし，その出力をソフトマックス関数を用いて分類すべき 6 ノードの値を得た．

4.3 結果と考察

表 1: 提案手法と従来手法の精度の比較

手法	精度%
提案手法	93.4
3D-CNN[4]	90.1
3D-CNN+RNN [9]	86
SURF 特徴量 [6]	94.0

表 2: 各動作に対する精度

	walking%	jogging%	running%	boxing%	waving%	clapping%
walking	99.1	0.9	0	0	0	0
jogging	4.2	86.0	9.7	0	0	0
running	2.1	8.8	90.0	0	0	0
boxing	0	0	0	93.4	1.1	6.6
waving	0	0	0	1.2	98.6	0.1
clapping	0	0	0	1.2	5.2	93.6

実験結果を表 1 に示す．結果から従来手法の深層学習を用いたものの中でもっとも良い結果が得られた．また，表 2 に各動作に対する精度を示す．表から動く速度や体の動きなどを特徴として捉えることができていると考えられる．また，似ている動作同士で誤認識してしまう傾向が見られた．これは人間の目で見ても判断がつきにくいものが誤認識に繋がっている．

本手法は特徴を自動設計されるという点で従来の人手で特徴を設計する手法より優れており、様々な応用が考えられる．一方で，ネットワーク構築に関しては様々なパラメータが存在しており，モデル構築者はパラメータ設定に注力を注ぐ必要がある．特に畳み込み層とプーリング層の構成は一般にどのような組み合わせが良い性能が出るか知られておらず，手探りで探す必要がある．また，実験を行う中でカーネルの畳み込みサイズを 3 より大きな数値にすると途端に学習が進まなくなってしまったが，これは時間軸方向にも畳み込み計算を行うために畳み込み計算の総数が多くなりすぎることが原因と考えられる．LSTM 部においては，隠れ層の数が小さすぎるとうまく学習を行うことができず，逆に大きすぎると過学習を起こしてしまう原因となる．本実験では 25〜30 程度で高い認識性能を出すことができたが，さらなる調整が必要であると考える．

深層学習において精度をあげる要因として学習データの質と量が重要である．そのために学習データの水増しも精度をあげるための重要な要因と考えられ，動画像においてはどのように水増しすべきか熟慮すべきである．

5 結論と今後の課題

本研究では，3D-CNN による特徴抽出した後，LSTM による動画像認識を実現させた．深層学習を用いた学習によりネットワーク構築に柔軟性があり様々なタスクに応用できる一方，パラメータ設定に注力を注ぐ必要がある．

今後はパラメータの最適化手法などの提案によりより柔軟性の高いネットワーク構築を行い，さらなる性能向上を目指していくことを検討する．また，本提案手法の長期記憶性はどの程度まで有効なのかさらなるじ実験を行い確認していく必要がある．さらに，LSTM を複数繋げた構成や分岐を持つ構成などのネットワークの複雑性が性能にどのような効果が表れるか検討の対象となると考えられる．

参考文献

[1] 勝手美紗, 内海ゆづ子, 黄瀬浩一. 物体と動き特徴を用いた行動認識. 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 111(430), pp. 125–126, 2012.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Proc. NIPS*, 2012.

[3] Felix A. Gers, Jurgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation 12.10*, pp. 2451–2471, 2000.

[4] Ji Shuiwang, Wei Xu, Ming Yang, , and Kai Yu. 3d convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, vol. 35*, 2013.

[5] Takumi Kobayashi and Nobuyuki Otsu. A three-way autocorrelation based approach to human identification by gait. *In Proc. of IEEE Workshop on Visual Surveillance*, p. 185192, 2006.

[6] 野口顕嗣. 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 (MIRU2010), 2010.

[7] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *Under review as a conference paper at ICLR 2015*, 2015.

[8] Ivan Laptev. Recognition of human actions. http://www.nada.kth.se/cvap/actions/, 2005.

[9] 浅谷学嗣, 田川聖一, 新岡宏彦, 三宅淳. 動画像認識のための 3 次元畳み込み rnn の提案. 情報処理学会研究報告コンピュータビジョンとイメージメディア (CVIM)Vol.2016-CVIM-201 No.6, pp. 1–4, 2016.