5V-01

ポーズデータとNNを用いた動作識別手法の調査

高崎 智香子† 竹房 あつ子‡ 中田 秀基§ 小口 正人†

†お茶の水女子大学 ‡国立情報学研究所 §産業技術総合研究所

1. はじめに

センサ機器やクラウドコンピューティングの普及により，一般家庭で取得，蓄積した動画像が子供やお年寄りの見守りサービスや防犯対策，セキュリティに活用されるようになってきた．しかし，家庭のセンサで取得した動画像をリアルタイムに機械学習を用いて解析するにはデータサイズと解析計算量が大きいため，サーバやストレージを用いてデータの分析や蓄積を行う必要がある．

我々は，姿勢推定ライブラリ OpenPose[1][2][3][4] を使ってセンサ側で動画像から特徴量を抽出し，クラウドでその特徴量データのみを用いて機械学習による動作識別を行うことで，処理遅延やプライバシーの問題に対処する分散処理手法を提案している [5]．また，STAIR Actions[6] データセットのうち，3 カテゴリを用いた識別を行い，80%以上の精度で識別可能であること，センサからクラウドへのデータ転送量を大幅に削減できることを確認した．しかし，3 カテゴリの識別では汎用性が低く，過学習の改善も課題となっていた．本研究では，より多様な動作識別を活用した実アプリケーションへ応用を目指し，同データセットの全 100 カテゴリを用いた識別を行う．

2. 背景

本研究では，図 1 のようなシステムを想定している．各一般家庭に設置されたセンサで動画像を取得し，前処理を行うことで特徴量を抽出する．その特徴量のみをクラウドに集約し，機械学習処理を行うことで動画に含まれる動作を識別する．クラウド側では特徴量データのみを使用して十分に動作を解析できるのか，どの機械学習手法を用いると高い精度が得られるのかを調査する．

提案システムでは，センサ側における前処理に OpenPose を，クラウド側における機械学習処理に Keras[7] を用いる．OpenPose は，深層学習を用いて人物のポーズをリアルタイムに抽出する手法である．加速度センサなどの特殊センサを使わずに，カメラによる画像や動画のみで解析できる．Keras はニューラルネットワーク (NN) を実装するためのライブラリで，様々な NN に対応可能である．Keras により簡単にモデルを記述することができる．

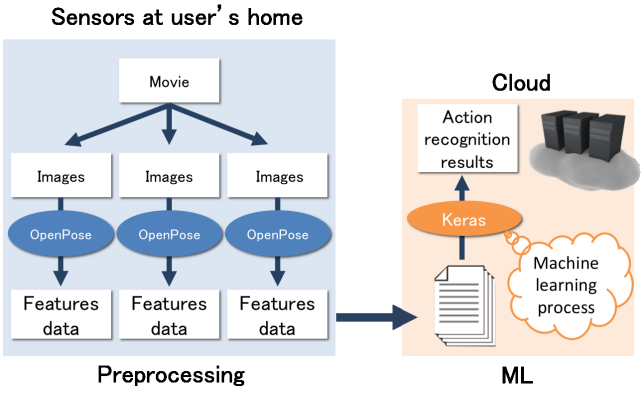

図 1: 提案する分散動画像解析システム

3. 実験

OpenPose を用いて画像から抽出した関節点の座標データを使用し，Keras を用いて構築した複数の機械学習モデルを用いて動作識別精度を比較する．

3.1 本研究で使用する機械学習手法

機械学習手法として，1. 全結合 NN モデル，2. LSTM モデルの 2 手法で動作識別を行い，予測上位 5 カテゴリに正解カテゴリが含まれる精度を比較した．NN は人の神経細胞を模したニューラルネットワークに畳み込み処理を導入したモデルである．LSTM は RNN の長期記憶ができないという欠点を解消し，データの長期依存を学習可能にした手法である．

実験では，時間ステップ数を 10，20，30 と設定したモデルを使用し，過学習を抑制するため，無効化率 2 割の dropout と batch normalization(BN) を導入した．NN では，dropout のみ，BN のみ，dropout と BN の両方を導入した場合とどちらも導入しない場合の 4 パターンで識別した．LSTM の dropout には，入力の時点でノードを無効化する dropout と，再帰の時点でノードを無効化する recurrent dropout の 2 種類があり，dropout のみ，recurrent dropout のみ，dropout と recurrent dropout の両方を導入した場合とどちらも導入しない場合の 4 パターンで実験を行った．

3.2 使用データセット

データセットには，日常の動作 100 カテゴリの動画を約 1000 ずつ集めた STAIR Actions を利用した．各動画から等間隔に複数枚の静止画を抽出した後，OpenPose を用いて人間 の 25 の関節点の画像上の x，y 座標を取得して特徴量 50 のデータを作成した．各モデルで使用したデータの詳細は表 1 の通りである．1. NN モデルでは，各画像の特徴量を時系列順に並べて使用し，2. LSTM モデルでは，各画像の特徴量を 1 step ごとの入力として使用する．

A Study on Action Recognition Method with Estimated Pose by using NN

Chikako Takasaki†
Atsuko Takefusa‡
Nakada Hidemoto§
Masato Oguchi†
†Ochanomizu University
‡National Institute of Informatics
§National Institute of Advanced Industrial Science and Technology (AIST)

表 1: データ数

モデル	画像数	間隔	データ数
(1a) NN	10	0.1 sec	87923
(1b) NN	10	0.3 sec	96807
(2a) LSTM w/10 steps	10	0.1 sec	87923
(2b) LSTM w/10 steps	10	0.2 sec	96807
(2c) LSTM w/20 steps	20	0.1 sec	128039
(2d) LSTM w/30 steps	30	0.1 sec	85553

3.3 識別精度の比較

図 2，図 3 に NN および LSTM による識別精度を示す．図 2 から，NN では BN の導入では識別精度が悪化しているが，dropout の導入により精度が改善されていることがわかる．また，図 3 の結果においても dropout の導入による精度の改善が見られた．時間ステップ数を 20 に設定した際の精度が良くなる傾向にあるが，(2d) の時間ステップ数 30 のモデルに dropout のみを導入することで最も良い精度を得ることができた．NN と LSTM の識別精度を比較すると，LSTM の方が良い結果が得られ，LSTM が時系列の学習に適していることがわかる．

3.4 学習処理時間の比較

今回の実験で使用した計算機の性能を表 2 に示す．OpenPose を用いたキーポイント抽出にかかる時間の 5 回平均は画像 10 枚あたり 2.14 秒で，1000 データあたりの機械学習の推論にかかる平均時間は NN モデルでは 0.082 秒，LSTM モデルでは 0.451 秒であった．OpenPose の処理時間が機械学習の推論にかかる時間より長く，センサ側での処理が重くなってしまっている．リアルタイム処理を行うための適切な前処理時間，動画から静止画を取得する間隔などを今後調査する必要がある．

4. 関連研究

Hara ら [8] は，動画から行動を識別するため，3D Residual Network(ResNet)[9] による性能改善を示した．しかし，動作識別処理は計算量が膨大であるため，一般家庭において深層学習を使用した解析を行うことは難しい．

本研究では，動画像に含まれる人間のキーポイントの座標値のみを使用して，行動を十分に認識できるかについて調査する．リアルタイムに動画像を解析するため，データ量を削減し，エッジとクラウドで処理を分散した後でも，十分な認識精度を確保することが目的である．

5. まとめと今後の予定

STAIR Actions データセットから取得した画像を OpenPose を用いて関節点の座標値に変換し，複数の機械学習モデルで動作の識別精度を比較した．Dropout の導入によって精度を改善できることがわかり，時間ステップ 30 の LSTM による動作識別精度が最も良くなることがわかった．

今後の課題として，家庭におけるエッジデバイスを用いてより実環境に近い環境での実験と評価を行い，リアルタイム識別を目指した処理時間を調査する．

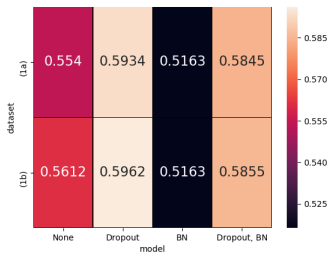

図 2: NN による識別精度

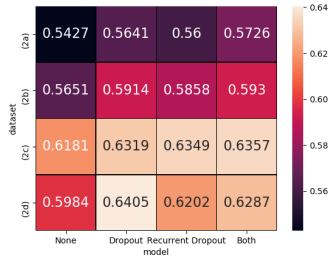

図 3: LSTM による識別精度

表 2: 実験で使用した計算機の性能

OS	Ubuntu 16.04LTS
CPU	Intel(R) Xeon(R) CPU W5590 @3.33GHz
GPU	NVIDIA GeForce GTX 980
Memory	49Gbyte

謝辞

この成果の一部は，JSPS 科研費 JP19H04089，国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO)，JST CREST JP-MJCR1503 の委託業務及び，2019 年度国立情報学研究所公募型共同研究 (19S0501) の助成を受けたものです．

参考文献

[1] Z. Cao, et al. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[2] T. Simon, et al. Hand keypoint detection in single images using multiview bootstrapping. In *Proc. IEEE conference on Computer Vision and Pattern Recognition*, pp. 1145–1153, 2017.

[3] Z. Cao, et al. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[4] SE Wei, et al. Convolutional pose machines. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, 2016.

[5] C. Takasaki, et al. A study of action recognition using pose data toward distributed processing over edge and cloud. In *Proc. the 11th IEEE International Conference on Cloud Computing Technology and Science (CloudCom2019)*, pp. 111–118, 2019.

[6] Y. Yoshikawa, et al. Stair actions: A video dataset of everyday home actions. *arXiv preprint arXiv:1804.04326*, 2018.

[7] Keras: The Python Deep Learning library . `https://keras.io/`.

[8] K. Hara, et al. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proc. the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018.

[9] K. He, et al. Deep residual learning for image recognition. In *Proc. the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.