

# 時系列骨格データにおける時間的処理を用いた 人間行動認識

藤 亮太†

黄 潤和‡

法政大学大学院 情報科学研究科†

法政大学 情報科学部‡

## 1. まえがき

近年、人間の行動認識を機械にさせる研究が盛んに行われている。映像データは高さ、幅、フレーム（時間）の3次元のデータであることを考慮し、3DConvolution を使った研究がなされている。しかし、高さ、幅と同時に時間の処理を行うことは動画の構造を軽視しているのではないかという問題点[1]もある。さらに、映像には背景があるが、背景が行動認識に依存している問題[2]もある。したがって本研究では、フレームごとの座標差分をとることで時間的処理をあらかじめ行い、空間軸と時間軸を分けて処理する。背景に依存しない人間の行動により着目している骨格データを用いて、人間の行動分類を行う。そのデータを Long-Short Term Memory (LSTM) に学習させる。どの程度差分をとるべきかと、その威力を調査する。

## 2. 使用するデータセット

使用するデータセットは NTU データセット[3]に含まれる骨格データである。表1に詳細を示す。本研究では、1人のみ取得しているデータを使用するため分類クラスも60から40となる。この骨格データは25関節分あるが、そのうち両手の親指と先端、両足の先端を除いた19関節分を用いる。また、関節にはx,y,zの3軸の座標があるが、x,yのみ使用する。これは将来的に撮影した動画から骨格データが取れたとしても、深度を測定することができない場合にも対応するためである。したがって、19関節、2座標の38次元のベクトルを1フレーム分の情報とする。

表1 NTU データセット情報

サンプル数	56800	→	36286
行動クラス数	60	→	40
FPS	30		
センサー	Kinect v2		

Human Action Recognition Using Temporal Processing for Sequential Skeleton Data

†Ryota To, ‡Runhe Huang

†Graduate School of C.I.S, Hosei University

‡Faculty of C.I.S, Hosei University

## 3. フレーム座標差分の取り方と整形

NTU データセットはノイズを含んでいるデータである。指数平滑フィルタを用いてノイズを軽減させた。1フレーム前の出力からの更新パラメータは0.85とした。フレーム差分をdiffとして、あるフレームfの関節番号jの座標を $c_j^f$ としたとき、 $c_j^{f+diff} - c_j^f$ を学習データとする。データによってフレーム数が違うデータであることから、全データの最大フレーム数に届かないデータに関してはZeroPaddingをする。nフレーム分のデータがあったとき、差分をとるとn-diffフレーム分のデータになる。本研究では、diffを1, 4, 7, 10, 13, 16と設定する。また、データを0~1に正規化するが、差分データは方向も指標となるので-1~1に正規化する。

## 4. LSTM の構造と諸パラメータ

本研究では2層LSTMを用いる。各層の隠れ層は256とした。図1に2層LSTMを含んだ本研究の全体の構造を示す。TensorFlowをバックエンドとするKerasで実装した。

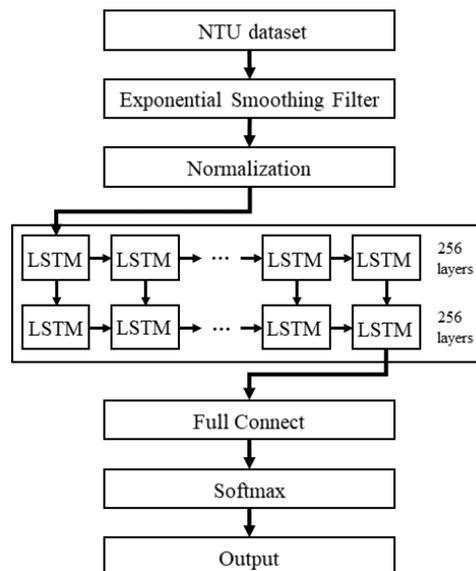


図1 2層LSTMを含んだ全体図

学習データは29018件、テストデータは7268件の約8:2の割合で分割する。フレーム差分を

とらないデータの結果を基準として、差分をとったデータの結果を考察していく。Epochs100, 損失関数は Cross Category Entropy, 最適化アルゴリズムは Adam を用い, 学習率は $10^{-4}$ とした。3 回 loss が低下しなかった場合は早期停止をする。

### 5. 結果と考察

各差分のテストデータの精度と損失関数の値を図2に、適合率, 再現率, F1 値を表2に示す。

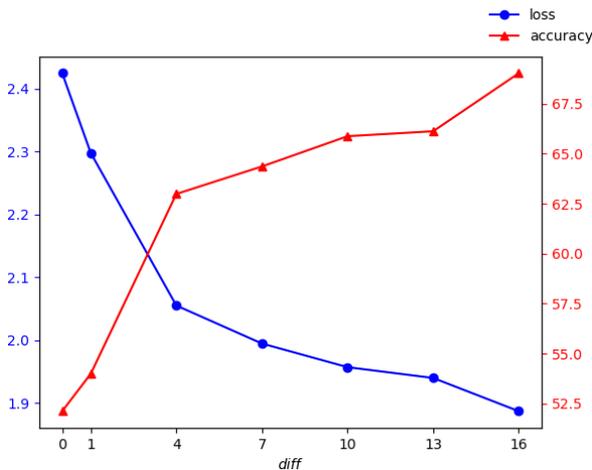


図2 各差分の分類精度

表2 各差分の適合率, 再現率, F1 値

diff	適合率	再現率	F1 値
0	0.52	0.52	0.50
1	0.52	0.54	0.51
4	0.64	0.63	<b>0.62</b>
7	0.64	0.64	<b>0.63</b>
10	0.66	0.66	0.66
13	0.66	0.66	0.65
16	0.69	0.69	<b>0.69</b>

結果から、差分をとった方が精度上昇の傾向が見られ約69%まで記録し、損失関数の値も減少した。特に、 $diff = 4$ からは大きく変動している。これらのことから、フレーム差分間あたりの座標変位が速度となり、人間の骨格点の動きをより捉えられているといえる。 $diff = 4$ は時間にとると約0.13秒である。この間隔で動きを捉えていると考え、 $diff = 1$ の微小時間よりも変化を捉えられているといえる。各 $diff$ の精度と F1 値は、 $diff$ が上昇するにつれ、共に上昇した。しかし、この関係が常に成立するわけではない。わずかではあるが、 $diff = 10$ あたりから過学習が生じやすくなった。 $diff = 10 \sim 16$ ということは、約 0.33~0.53 秒である。これは時系列的特徴が欠損し始めていると考えられる。3 章のデータ数の減少により、精度と F1 値の上昇は見込める

が、過学習が発生する要因となっていくと考えられる。したがって、 $diff = 4 \sim 7$ として差分をとっていくべきであると考え。次に顕著に発生した誤分類について考察していく。 $diff = 0$ であるとき、“drink water”の分類精度が約 0.54 %であった。“brushing teeth”などに約 27%が誤分類されていたので、これは指の先端のデータを学習させていなかったためだと考えられる。しかし、 $diff = 1$ となると、精度が約 50.2%となる。差分をとることで、細かな動きにも対応できたことが伺える。しかし、 $diff = 4 \sim 7$ であっても捉えられない動きがあった。例えば“reading”, “writing”が“typing on a keyboard”に分類されてしまった。読み書きをしているときに関節点の変位はなかなか現れない。タイピングをしているときの姿勢に酷似していることも原因である。これらの行動は $diff = 4$ が他の $diff$ よりも F1 値が約0.15高くなった程度に過ぎなかった。動きがあまりないデータに対しては、時間差分という指標はあまり有用ではないことが考えられる。

### 6. むすび

本研究では、 $x, y, t$ で構成されているデータに対して、フレーム差分をとる処理を事前に行い、LSTM による学習を行った。結果として差分をとったデータの精度が上昇する傾向が伺えた。最近では時間を考慮した複雑な LSTM や3次元畳み込みニューラルネットワークを構築することが流行っているが、時間軸と空間軸を分け、時間軸に対し事前処理をすることで精度が上がるということが判明した。課題として、時間経過により変化しない行動の識別ができていなかった。それらは人間が所持している物体に依存している行動が多かった。物体認識を考慮した学習にすることで、精度が上がると思われる。

### 参考文献

- [1] De-An Huang. et al. What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Dataset. In Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018, pp.7366-7375.
- [2] Yun He. et al. Human Action Recognition without Human. In Proceedings of European Conference on Computer Vision Workshops (ECCV), 2016, pp.11-17.
- [3] Amir Shahrudiy. et al. NTU RGB+D: A Large Scale Dataset for 3D Human Action Analysis. In Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016, pp.1010-1019.