

動画像内における複数人の動作識別への取り組み

Yiwen Fang[†]小林 一郎^{††}[†]お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース^{††}お茶の水女子大学大学院 基幹研究院 自然科学系

1 はじめに

近年、2次元上に表現される人の動作識別に関する研究が多くされている。また、動作識別においては3次元空間で識別が可能であれば、人の動作をより包括的に捉えることができ、精度を向上させることができる。具体的な取り組みとして、人が映る2次元の動画像から3次元の姿勢を推定し、姿勢情報を関節座標に変換する。この際、複数人の関節座標を共通空間に保持することにより、包括的に複数人の相互作用の識別が行えるようにする。関節座標データに対応する動作ラベルを含むデータセットを作成し、関節とラベルの対応関係を捉えたグラフ畳み込みニューラルネットワーク(GCN)を通じて関節情報と関節で結合されている骨格情報から動作識別を行う。

2 関連研究

2.1 動画像内の複数人の3次元姿勢推定

Dongら[1]は、複数のカメラに映る同一人物を特定するためには、1) 外観の類似度の計算。人の重要な関節を複数のカメラ画角で識別できるネットワーク(Person Re-Identification ネットワーク)を使い、複数の動画像に映る被写体の類似度を計算する。2) 幾何学上の一致性の検証。同じ人物を異なる画角から見たときに、外観が似ている場合や人物が不明瞭な場合に同一人物と判断する。カメラからの被写体の距離のデータを取得し、異なる画角から撮った画像間の相対運動を利用して、2次元姿勢の位置を予測する。最後に、同一人物の複数の画角における2次元姿勢をマッチングさせた後、人体骨格の物理構造を統合できる3D pictorial structure(3DPS)モデルを用いて最終的な3次元姿勢を構築する手法を提案している。

2.2 3次元姿勢推定に基づく動作識別

Shiら[2]の研究を、2つに分けて説明する。1つ目は、入力として与えられた人の3次元関節ごとに異なるネットワークポロジを自動的に学習できるデュアルストリーム適応型グラフ畳み込みネットワーク(2s-AGCN)を提案している。GCNのユニットは、Spatial GCN(SGCN)とTemporal GCN(TGCN)に分けられる。SGCNでは、各フレームで人の骨格における接続情報を扱う。TGCNでは、連続する2つのフレームで同じ関節点を追従する。2つ目は、関節と関節で結合されている骨格情報を入力として、関節の系列を扱うJoint Streamと骨格の系列を扱うBone Streamから構成されるデュアルストリームネットワーク構造を提案している。Joint stream(Js)は、人の関節情報に基づいて、動作識別のため重要な関節(例：手首)を識別し、Bone stream(Bs)は、人の骨格情報に基づいて、重要な関節の連結(例：腕)を識別する。

3 提案手法

本研究では、上記2つの先行研究[1][2]を元に、関節座標データを対象とした動作識別モデルの構築を行った。構築したモデルの概要を図1に示す。

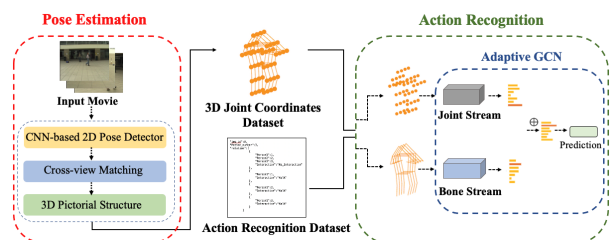


図1: 提案手法の概要図

CampusとShelfは、多画角から撮影した2種類の3次元人体姿勢データセットである、それぞれ補正された3台と5台のカメラを使って、屋外で3人が交流するシー

An Approach to the recognition of multiple people's movements in movies

[†]Yiwen Fang(g2140677@is.ocha.ac.jp)

^{††}Ichiro Kobayashi(koba@is.ocha.ac.jp)

ンと室内で4人が棚を作るシーンを撮影している。また、人の関節数を17個に設定し、Dongら[1]の研究に基づき、同一人物の2次元姿勢からそれに対応する3次元姿勢を推定し、関節座標を抽出した後、正規化し、共通の座標空間で表現した。そして、人の関節座標データに対応する動作ラベルをペアにしたデータセットを作成した。具体的に、単一の人物に対して、一人の動作を逐次的に識別する（例：Person 1: 'Walk'）。複数人に対して、何人が何をしているかを識別する（例：Person 2,3,4: 'Talk'）。さらに、これら複数人の骨格情報に対する3次元情報に対して、Shiらの手法[2]で多角的に撮影された複数の動画画像に映る人物の動作識別を行うモデルの構築を行った。

個々の人の骨格構造を異なることから、2s-AGCNを利用して、それぞれ3次元関節構造を表現可能なネットワークポロジータに変換する。その上で、GCNユニット中のSGCNとTGCNに対して、動作識別のために重要な関節の注意機構(Attention)を導入し動作識別の推定精度を向上させた。SGCNに対して、より重要な関節と骨格情報を注意機構によりそれにつけられているラベルの対応関係を強化させ学習させた。例えば、「Fight」の識別では、2人の手首の関節と足首の関節を捉えるだけでなく、腕と脚を捉えて学習させた。TGCNに対して、連続する複数のフレームで同じ重要な関節と骨格情報を学習させた。例えば、連続する複数のフレームのインタラクション動作ラベルが全部同じ場合、それに対応するフレームの関節と骨格情報を用いて動作識別を学習させた。最後に、関節と骨格情報を入力して、JsとBsを含むデュアルストリームネットワーク構造を利用して、単一人物および複数人の動作を識別する。

4 実験

4.1 実験設定

モデルの学習に関するパラメータは、先行研究[1][2]の設定に基づいた。CampusとShelfから学習は作成した人の3次元関節座標データと動作ラベルをつけるデータセット（「Campus」、「Shelf」、「Campus+Shelf」と呼ぶ）によって、単一人物および複数人の動作を識別した。

4.2 実験結果

単一人物および複数人の動作識別結果をそれぞれ表1, 2に示す。

複数人物の動作識別実験のモデルについては、比較の対象となる複数人物の3次元関節座標からの動作識別

表1: 単一人物の動作識別精度

Model*	Test Dataset**		
	Campus	Shelf	Campus+Shelf
ST-GCN	73.33%	70.11%	86.11%
Js-AGCN	74.52%	72.38%	73.28%
Bs-AGCN	81.22%	80.73%	79.72%
2s-AGCN	79.37%	77.86%	84.19%
Js-AGCN-new (Ours)	78.82%	77.59%	75.69%
Bs-AGCN-new (Ours)	84.71%	76.51%	75.32%
2s-AGCN-new (Ours)	82.35%	76.72%	88.43%

* ModelのTrain Data: 80% of Campus+Shelf Dataset

** Test Data: 20% of Campus+Shelf Dataset

実験がないので、比率の異なるデータを入力して実験を行う。

表2: 複数人物の動作識別精度

Model*	Test Dataset**		
	Campus	Shelf	Campus+Shelf
Campus	74.18%	85.82%	86.67%
Shelf	58.74%	79.35%	95.08%
Campus+Shelf	63.72%	75.71%	98.27%

* ModelのTrain Data: 80% of 「Model」 Dataset

** Test Data: 20% of 「Model」 Dataset

実験結果より、本研究において提案したモデルが作成したデータセットに対して、単一および複数人の動作識別を正しく推定できたことを確認した。

5 おわりに

本研究では、動画画像中の単一人物および複数人の2次元動作情報を3次元情報に変換して、それぞれの動作識別を行う手法を提案した。

今後の課題として、提案手法の性能を様々なデータセットを用いて検証すると共に、他の類似する手法との比較を行うつもりである。

参考文献

- [1] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7792–7801, 2019.
- [2] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12026–12035, 2019.