動画像内における複数人の動作識別への取り組み

理学専攻・情報科学コース　Yiwen Fang

1　はじめに

近年，OpenPose [1] などを用いた 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

本研究では，Dong らの手法 [2] に基づき，多角的なカメラで撮った動画像によって多人数の 3 次元姿勢推定を行い，Shi らの手法 [3] に基づき，骨格情報を用いて人の動作識別を行う手法を提案する。以下に，それら先行研究の説明をする。

2.1　動画像内の複数人の 3 次元姿勢推定

Dong ら [2] 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。そして, 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 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。図 1 を参照します。

2.2　3 次元姿勢推定に基づく動作識別

Shi らの研究を，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トポロジー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 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 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。

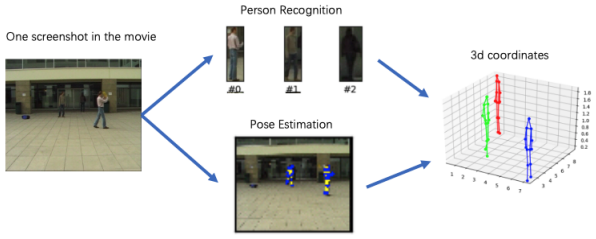

図 1: Campus データセットから取り出す 3 次元姿勢の構築

3　提案手法

本研究では，上記 2 つの先行研究 [2] [3] 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 2 に示す。

3.1　データセットの前処理

先行研究 [2] を使って，複数のカメラで撮影した動画像を複数の画角間で対応させることで，そのシーン内の人数を自動的に識別し，異なる画角での同一人物の 2 次元姿勢を与えると，それを対応する三次元姿勢を構築することができた。また，人体の関節数を 17 個に設定しました。これらを踏まえた上で，任意の時間間隔における全ての人の関節座標データを抽出した。

Campus と Shelf[1]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 人が交流するシーンと室内で 4 人が棚を作るシーンを撮影している。Campus と Shelf から取り出した各フレームの画像内の人物に対して，人の 3 次元関節座標データを抽出した。そして，複数人の動作識別を行うために，すべての関節データを正規化し，共通の空間に格納します。最後に，単一人物および複数人の動作識別を学習するため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のデータセット構築とします。 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')。

3.2　単一人物および複数人の動作識別

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 [3] である 2s-AGCN モデルを適用することで，多角的に撮影された複数の動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。

人体の自然な骨格構造を異なるトポロジーで表現し，単一人物および複数人の動作の分析精度を向上させることを目指します。ネットワークトポロジーには，単一人物の 3 次元関節構造の表現するもの（独立トポロジー）および複数人の 3 次元関節構造を表現するグローバルトポロジーがある。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

[1]https://campar.in.tum.de/Chair/MultiHumanPose

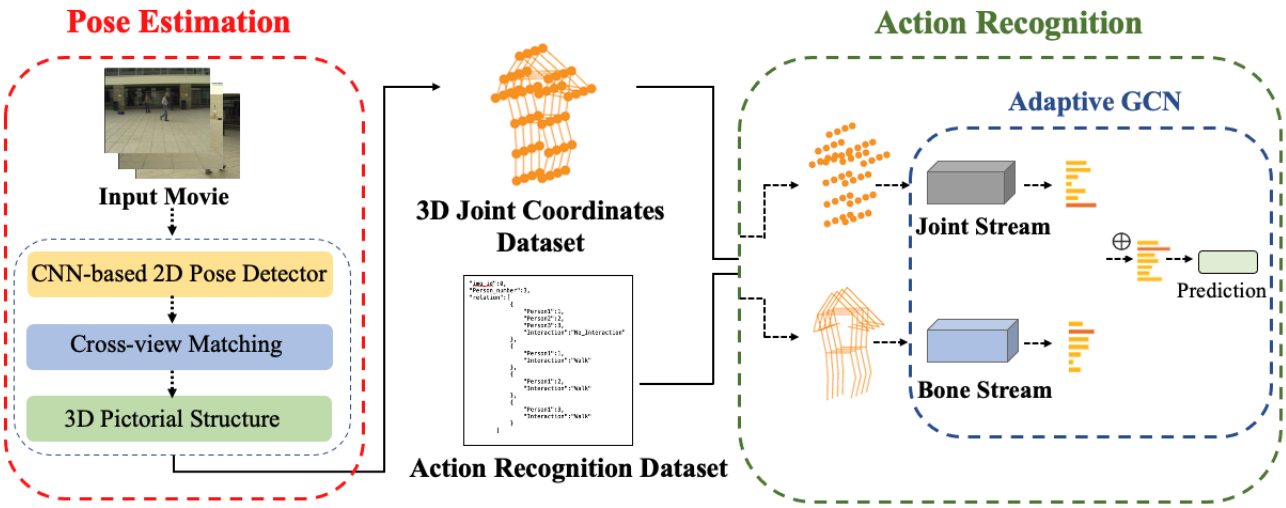

図 2: 提案手法の概要図

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。また, 関節と骨格情報を入力して,Joint Stream と Bone Stream 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。例えば，単一人物の動作「Walk」の場合，Joint Stream は足首の関節に，Bone Stream は太ももとふくらはぎの骨に注目して，動作を学習した。最後に，2 つのネットワークの学習結果を組み合わせ，Softmax 関数を用いて，単一人物および複数人の動作を予測する。

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 [2] [3] 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人の 3 次元関節座標データに対応する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きを識別した。

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。

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

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。例えば，3 行目の 1 列目では,「Campus + Shelf」の 80%を訓練データとして，残りの「Campus」の 20%を評価データとして選択したところ，63.72%の精度となった.

表 2: 複数人物の動作識別精度

Model* \ Test Dataset**	Campus	Shelf	Campus+Shelf
Camupus	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, 識別する手法を提案した。実験を通じて，提案手法は単一人物の動作識別を向上させるだけでなく，複数人の動作を識別することも可能であることを確認した。

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] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 43, No. 1, p. 172–186, jan 2021.

[2] 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.

[3] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. 2019.