時空間のアテンション情報を考慮した Graph Convolutional Networks による動作認識に関する研究

TP19008 白木克俊　　指導教授：山下隆義

1. はじめに

骨格データは人間の動きを直接捉えることができることや, 認識時における環境や視点の変化に対して頑健な利点がある. 骨格データを用いた Graph Convolutional Networks (GCN) による従来の動作認識では, 人間の骨格パターンでグラフ構造を事前に定義するため, 動作特有の関節間の関係性を考慮できない. また, 認識における関節の重要度も動作ごとに異なることが予想される. 本研究では, 関節の重要度と関係性を考慮して動作認識を行う Spatial Temporal Attention Graph Convolutional Networks (STA-GCN) を提案する. STA-GCN は, フレームごとの関節の重要度と, 動作ごとの関節間の接続関係を獲得する. 重要度を特徴マップに重み付けし, 接続関係を用いてグラフ畳み込み処理を行うことで, 重要度と関係性を考慮しつつ認識を行う. また, マルチモーダル学習において Mechanics-stream 構造を導入し, 高精度化を実現する.

2. 骨格データからの GCN を用いた動作認識

骨格データからの動作認識に, GCN を用いた手法である Spatial Temporal GCN (ST-GCN) [1] がある. ST-GCN は, 同一フレーム内の関節を結ぶ空間グラフと, 隣接フレームの同一関節を結ぶ時間グラフの 2 つのグラフ構造 (図 1) で骨格データを表現することにより高い認識精度を達成した. しかしながら, ST-GCN は, 空間グラフの接続関係 E_S を骨格パターン H で定義しているため, 動作特有の関節間の関係性を考慮できない. また, 動作や時間ごとに変化する関節の重要度も表現できないという問題がある.

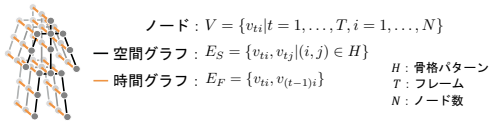


図 1: ST-GCN における骨格データのグラフ表現

3. 提案手法

関節の重要度と関係性を考慮して動作認識を行う Spatial Temporal Attention GCN (STA-GCN) を提案する. STA-GCN は関節の重要度である Attention node と, 関節間の関係性である Attention edge を獲得する. 獲得した Attention を考慮しつつ認識を行う. また, マルチモーダル学習において Mechanics-stream 構造を提案する. Mechanics-stream 構造は, 各モーダルが持つ力学的特性や値のスケールの違いにもとづいてネットワークを構築する.

3.1 STA-GCN のネットワーク構造

図 2 に STA-GCN のネットワーク構造を示す. STA-GCN は, Feature extractor, Attention branch, Perception branch の 3 つのモジュールで構成する. 各モジュールに含まれる STGC-block では, 空間グラフと時間グラフのグラフ畳み込み処理を行う. Feature extractor には 2 つのモーダルを入力し, 人間の骨格パターン $\mathbf{A}^{skeleton}$ を空間グラフとする複数の STGC-block により各モーダルの特徴マップを獲得し, 結合する. Attention branch は, 関節の重要度を表す Attention node と, 関節間の重要な関係性を表す Attention edge を特徴マップから獲得する. Perception branch は, Attention node と Attention edge を考慮しつつ, 最終的なクラス確率を出力する. 学習は, 2 つのブランチからの出力に対してクロスエントロピー誤差で損失を求め, 損失の和をネットワークの学習誤差として用いる.

3.2 Spatial Temporal Attention Graph

Attention node $M(\mathbf{X}_{out}^{FE})$ は関節 × フレームの 2 次元マップであり, 動作特有の関節の重要度をフレームごとに表現できる. 獲得した Attention node は Attention 機構を用いて Feature extractor からの特徴マップ $\mathbf{X}_{out}^{FE}$ へ反映し, Perception branch の入力とする. 式 (1) に Attention 機構を示す.

$$\mathbf{X}'_{out} = M(\mathbf{X}_{out}^{FE}) \cdot \mathbf{X}_{out}^{FE} \tag{1}$$

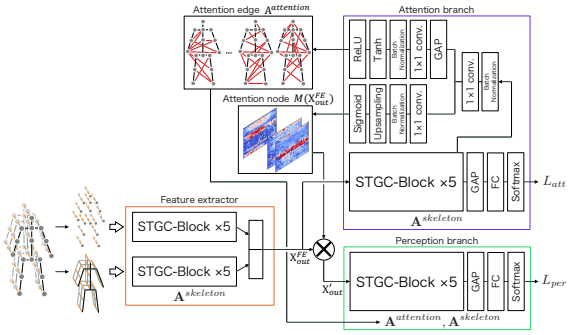


図 2: Spatial Temporal Attention GCN

Attention edge は, 動作ごとの関節間の重要な関係性を表す隣接行列である. 獲得した Attention edge は Perception branch に与える. Perception branch では, Attention edge $\mathbf{A}^{attention}$ と人間の骨格パターン $\mathbf{A}^{skeleton}$ によるグラフ畳み込み処理を行う. Perception branch における入力特徴 $\mathbf{X}_{in}$ に対する空間グラフ畳み込み処理を式 (2) に示す. $\mathbf{M}^{skel}, \mathbf{W}^{skel}, \mathbf{W}^{att}$ は重み行列である. Q は hop 数であり Q 個離れた関節までを結ぶ. K は動作ごとに生成する Attention edge の数である. 本研究では $Q = 3, K = 4$ とする.

$$\mathbf{X}_{out} = \sum_{q}^{Q} \mathbf{M}_q^{skel} \circ \hat{\mathbf{A}}_q^{skel} \mathbf{X}_{in} \mathbf{W}_q^{skel} + \sum_{k}^{K} \hat{\mathbf{A}}_k^{att} \mathbf{X}_{in} \mathbf{W}_k^{att} \tag{2}$$

Attention node と Attention edge によって時間的, 空間的な特徴を強調でき, 2 つの Attention を組み合わせたグラフ表現を Spatial Temporal Attention Graph (Attention graph) と呼ぶ.

3.3 Mechanics-stream 構造

Mechanics-stream 構造 (図 3) は, マルチモーダル学習におけるネットワーク構造である. 本研究では, 関節座標から関節速度, 関節加速度, 骨情報, 骨速度, 骨加速度を算出しネットワークの入力とする. 骨情報は, 関節間の距離であり関節間の方向も表現する. 座標や骨情報は空間的な位置の情報, 速度や加速度は時間的な移動量の情報であり, 各モーダルが持つ特性には違いがある. また, 座標や速度, 加速度は値のスケールが異なるため, 同じネットワークでの学習が困難である. そこで, 各モーダルが持つ力学的特性や値のスケールの違いにもとづいた Mechanics-stream 構造を提案する. Mechanics-stream 構造は 3 つのネットワークを用意し, 特性や値のスケールが近いモーダルを入力とする. 各ネットワークは独立して学習を行い, 各ネットワークから得られるクラス確率をクラスごとに合計することで最終的なクラス確率とする.

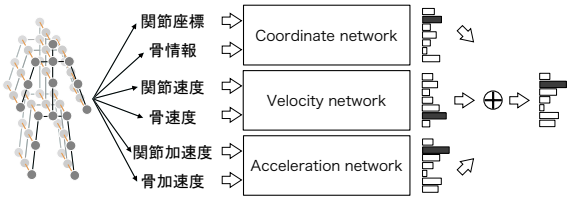
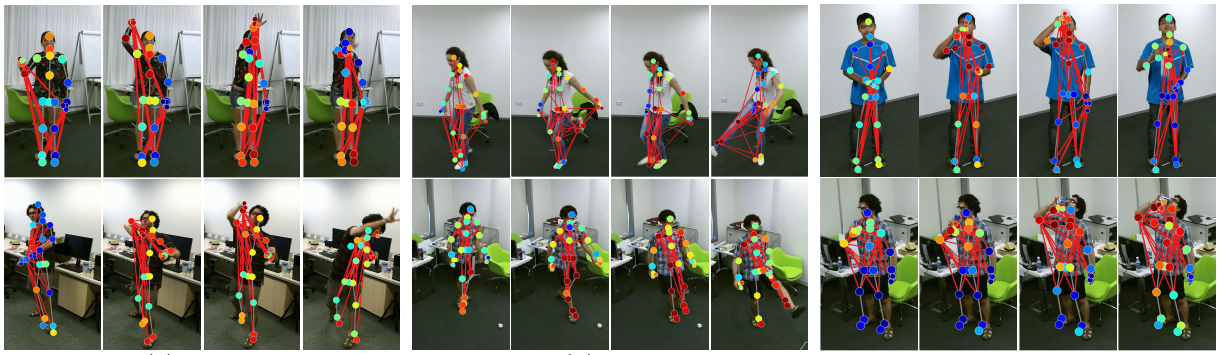


図 3: Mechanics-stream 構造

4. 評価実験

提案手法の有効性を確認するために評価実験を行う. データセットには NTU-RGB+D [2] と NTU-RGB+D120 [3] を用いる. NTU-RGB+D は関節数 25 の 3 次元座標 (X, Y, Z) を持つ動作認識用のデータセットであり, 動作クラス数は 60 である. 評価方法は, 40 名の被験者のデータを学習と検証で分ける Cross subject (x-sub) と, 3 方向から撮影したデータを学習と検証に分ける Cross view (x-view) があ



(a) Throw (b) Kick (c) Drink

図 4: **Attention graph の可視化**: 関節の色は Attention node にもとづいており, 赤いほど重要度が高く, 青いほど重要度が低い. 赤い線は Attention edge を示しており重みの大きい上位 30 本のみを描画している.

表 3: **モーダルの組み合わせの違いによる認識精度 [%]**: 単一のネットワークの認識精度 (Ind. net) と stream 構造を構築したときの認識精度 (X-stream). 評価は NTU-RGB+D の Cross subject.

Input data	w/o Attention		w/ Attention	
	Ind. net	X-stream	Ind. net	X-stream
Joint (coordinate, velocity)	86.0	87.1	86.2	87.2
Bone (coordinate, velocity)	86.1		86.3	
Joint (coordinate, velocity, acceleration)	86.7	87.8	85.7	86.7
Bone (coordinate, velocity, acceleration)	86.7		85.8	
Coordinate (joint, bone)	87.5	89.3	88.6	**90.1**
Velocity (joint, bone)	85.6		87.0	
Coordinate (joint, bone)	87.5	89.1	88.6	89.4
Velocity (joint, bone)	85.6		87.0	
Acceleration (joint, bone)	74.6		77.2	

表 1: 従来手法との認識精度の比較 [%]

NTU-RGB+D			NTU-RGB+D120		
Methods	x-sub	x-view	Methods	x-sub	x-setup
ST-GCN	81.5	88.3	2s-ALSTM	61.2	63.3
AS-GCN	86.8	94.2	BPEM	64.6	66.9
2s-AGCN	88.5	95.1	SkelMotion	67.7	66.9
AGC-SLTM	89.2	95.0	TSRJI	67.9	62.8
DGNN	89.9	**96.1**	ST-GCN	72.8	75.4
STA-GCN	**90.1**	95.8	STA-GCN	**83.9**	**86.5**

表 2: **Attention graph の従来手法への適用結果 [%]**: 評価は NTU-RGB+D の Cross subject.

Attention edge	Attention node	ST-GCN	AS-GCN	2s-AGCN
×	×	81.5	86.8	88.5
✓	×	84.1	86.9	89.2
×	✓	82.8	86.8	89.1
✓	✓	**84.8**	**87.0**	**89.3**

る. NTU-RGB+D120 は NTU-RGB+D に 60 の動作クラスを追加した大規模なデータセットである. 評価方法は, Cross subject と, カメラの高さや被験者との距離によって割り当てられた ID を用いてデータを学習と検証に分ける Cross setup (x-setup) がある.

4.1 従来手法との認識精度の比較

表 1 に, 従来手法との認識精度の比較結果を示す. どちらのデータセットにおいても提案手法の認識精度は, 従来手法と比較して認識精度の向上, または同等の認識精度を達成した. STA-GCN は, 動作特有の特徴を強調することで類似した動作でも異なる特徴を獲得できるため, 認識精度の向上に貢献したといえる.

4.2 Attention graph の可視化

STA-GCN により獲得した Attention graph を図 4 に示す. 投げる動作 (図 4(a)) は, エッジが右腕に集中している. 右腕の重要度は動作中に徐々に高くなり, 投げ終わると重要度が低くなる. 蹴る動作 (図 4(b)) は, 蹴り出す脚に対してエッジが集中する. このことから, 動作特有の Attention edge を獲得したといえる. 飲む動作 (図 4(c)) は投げると同様に右腕を重要視している. しかしながら, 飲む動作の脚の重要度は一貫して低いのに対し, 投げる動作は体重移動を行うため脚の重要度が高くなることから, 動作特有の Attention node を獲得したといえる.

4.3 Ablation study

Attention graph の従来手法への適用: 表 2 に, Attention graph を従来手法へ適用したときの認識精度を示す. 表 2 から, Attention edge, および Attention node のみを適用したほぼ全ての場合で認識精度の向上を確認できる. また, 全ての従来手法において Attention edge と Attention node の両方を適用した場合に最も認識精度が高くなる. この結果から, Attention graph は認識精度の向上に貢献しており, 両方を同時に適用した場合に最も効果的である.

モーダルの組み合わせの違いによる認識精度: 表 3 に, 入力モーダルの違いによる認識精度を示す. 表 3 から, 座標と速度を同じネットワークで学習するより, 別々のネットワークで学習した方が認識精度が高くなる. このことから, Mechanics-stream 構造の有効性を確認できる. しかしながら, 加速度を入力する場合では認識精度が低下した. 加速度は値のスケールが小さく, 認識のための十分な特徴が含まれていないことが考えられる.

5. おわりに

本研究では, 関節の重要度と関係性を考慮して動作認識を行う Spatial Temporal Attention GCN を提案した. また, マルチモーダル学習におけるモーダルの特性の違いにもとづいてネットワーク構築する Mechanics-stream 構造を提案した. 評価実験では, 提案手法が従来手法を超える認識精度を達成し有効性を確認した. また, 動作特有の Attention graph が獲得することを確認した. 今後の課題として, 動作認識以外のタスクへの応用などがあげられる.

参考文献

[1] S. Yan, *et al.*, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition", AAAI, 2018.

[2] S. Amir, *et al.*, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis", CVPR, 2016.

[3] L. Jun, *et al.*, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding", TPAMI, 2019.

研究業績

[1] K. Shiraki, *et al.*, "Spatial Temporal Attention Graph Convolutional Networks with Mechanics-Stream for Skeleton-based Action Recognition", ACCV, 2020.
(他 3 件)