電子情報通信学会パターン認識・メディア理解研究専門委員会
情報処理学会コンピュータビジョンとイメージメディア研究会
MIRU2022　第 25 回 画像の認識・理解シンポジウム

Structured Pooling を用いた複数の人物骨格と物体輪郭からの人物行動認識

八馬 遼[1,a)]　佐藤 文彬[1,b)]　関井 大気[1,c)]

概要

本稿では，骨格ベース行動認識の従来研究が持つ 3 つの課題，認識対象行動の拡大，骨格検出誤りの影響の低減，アノテーションコストの抑制，を同時に解決する手法を提案する．具体的には，複数人物の関節点に加え物体輪郭の端点を入力として用いて，各点を時空間上の 3D 点群として扱うことで，点群処理 DNN を用いて従来法では認識が困難であった行動を頑健に識別する．Permutation-Invariant な構造を DNN に採用し，各点が属するインスタンスやフレームなどの事前知識を下に段階的に特徴を集約する Structured Pooling を新たに導入することで，複数物体から得られた点群の特徴量を効率的にモデル化する．インスタンスごとに集約された特徴量を用いて，Multiple Instance Learning に基づく弱教師あり学習の枠組みにしたがい，密なアノテーションなしにインスタンスごとに行動を識別する．Kinetics-400 データセットを用いた実験では，各課題に対する提案法の有効性を検証し，また，本手法が弱教師あり時空間行動検出に適用可能であることを示す．

1. 背景

動画像に映った人物の取る行動を認識する技術は，ロボティクスや監視カメラなど様々な応用で重要な役割を果たす．行動認識は人物を撮影した動画像，または，動画像から検出した人物ごとの骨格情報を入力とする場合それぞれでアプローチが異なる．前者の外観ベースの手法 [6] は，動画像を直接入力として用いて行動ラベルを推定するため，比較的微小な動作に対しても行動を認識できる一方，Deep Neural Networks（DNN）の学習時と見えが異なる人物やシーンに対して頑健性が低下（過学習）する [5]．一方，後者の骨格ベースの手法 [15] は，複数人物骨格検出 [2] により動画像から検出した骨格情報のみを入力として用いるため，シーンや人物の見えの変化に比較的頑健である．

1 コニカミノルタ株式会社 技術開発本部 FORXAI 開発センター AI 技術開発部第 1 グループ
a) ryo.hachiuma@konicaminolta.com
b) fumiaki.sato1@konicaminolta.com
c) taiki.sekii@konicaminolta.com

骨格ベース行動認識の従来研究の多くは，以下に述べるように，問題設定や用いる DNN モデルに関していずれかの課題があり，利用できるシーンの拡大や性能の改善に際して拡張性に乏しい．

課題 1．認識対象行動に対する制限

DNN の入力を 1〜2 人程度の骨格情報のみに制限しているため，それ以上の人数での行動や人物以外の物体に関する情報が必要な行動を認識することは不良設定問題となる．一方，幅広い応用を想定した場合，多様な行動クラスを認識できることが望ましい．

課題 2．骨格検出誤りの影響

多くの従来法は，骨格情報と行動の関係をモデリングするのに Graph Neural Networks（GNN）[15] を用いる．これは，時系列に検出される同一人物の関節点の間で特徴量を正確に伝搬できることを前提としている．そのため，関節の未検出・誤検出が発生したり，人物追跡に失敗した場合，この前提が成り立たず認識精度が低下する．

課題 3．アノテーションに要する人的コスト

動画全体を特定の行動に分類する従来研究 [17] とは異なり，実際には，複数人物が違う行動を取る複雑なシーンも多い．その場合，人物ごとの行動を識別する必要があるが，DNN の学習に際して各フレームで人物ごとの行動ラベルが必要となるため，アノテーションに人的コストがかかる．

本稿では，骨格ベース行動認識における応用や性能面の拡張性を高めることを目的として，Structured Pooling（StructPool）と呼ぶ新たな DNN モデルを導入することにより，先述の 3 つの課題を同時に解決する手法を提案する．まず，動画像から検出された骨格の関節点に加え物体輪郭上の端点（以下，関節点と合わせて Keypoint または KP と総称）を検出し，StructPool の入力として用いることで，外観の過学習を抑えながら入力として用いる情報量を拡充する．また，入力の KP を点群データとして扱い，順不同の点群の特徴量を疎に集約する PointNet [9] をベースに，各点の属する人物・物体の検出結果（インスタンス）やフレームなどの事前知識を下に，Max-Pooling を用いて特徴量を段階的に集約する．これにより，KP 間で特徴量を伝搬することなしに，KP 同士の関係を効率的に

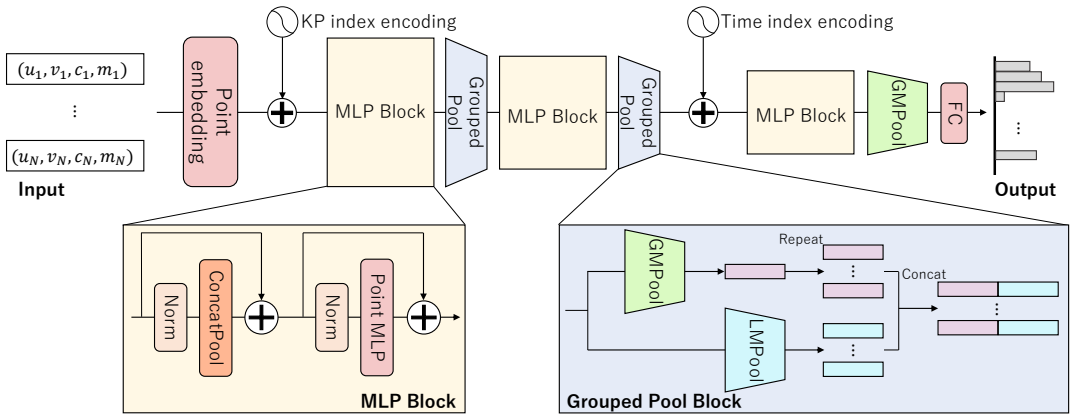

図 1: Structured Pooling のアーキテクチャ

モデリングすることで課題 1 と 2 に同時に対処する．最後に，Multiple Instance Learning [1]（MIL）に基づく弱教師あり学習の枠組みを StructPool に導入し，動画単位の行動ラベルのみ学習に用いながらも，Grad-CAM [11] を用いてインスタンスごとに行動を認識することで，課題 3 に対処する．実験では，行動認識の大規模データセットの 1 つである Kinetics-400 [3] を用いて，各課題に対する提案法の有効性を検証した．

2. 提案法

StructPool のアーキテクチャを図 1 に示す．入力の動画像に複数人物骨格検出および物体端点検出を適用し関節点と物体端点（以降，KP と総称）を検出する．DNN に入力される各 KP は，画像座標 (u, v)，検出時の確信度 c，物体カテゴリ番号 m （例えば，0 は “人”，1 は “車” など）を結合したベクトル (u, v, c, m) で表される．各 KP の種類の扱いについては後述する．

Point embedding 層は，各 KP の入力ベクトルを MLP（Multi-Layer Perceptron）により特徴ベクトルに変換する．KP index encoding は，この特徴ベクトルに各 KP のカテゴリ番号を Sinusoid position encoding [14] を用いて変換した特徴量を加算する．KP のカテゴリとは，例えば，関節点であれば 0 は “左肩”，1 は “右肩”，また物体端点であれば 0 は “左上”，1 は “右上”，といった KP の種類を表す．MLP Block で KP 間の関係性を下に変換された特徴ベクトルは，Grouped Pool Block により，同じインスタンスまたは同じフレームに属するグループごとに集約される．Time index encoding は，KP index encoding と同様に入力の特徴ベクトルに時刻を変換した特徴量を加算する．特徴ベクトルは，最終的に Global Max-Pooling（GMPool）により集約され，行動ごとのクラス確率の計算に用いられる．

2.1 Grouped Pool Block

Grouped Pool Block は，同じインスタンスに属する各 KP の特徴ベクトルを，または，同じフレームに属する各インスタンスの特徴ベクトルをグループに分ける．続いて，グループごとの局所的な Max-Pooling（Local Max-Pooling または LMPool）により特徴ベクトルをグループ数分に集約・削減する．最初の Grouped Pool Block では動画内のインスタンス数分の特徴ベクトルが出力され，次の Grouped Pool Block ではフレーム数分の特徴ベクトルが出力される．

Grouped Pool Block の処理は，次式のように表される．

$$Y = \mathrm{Concat}\left(\mathrm{LMPool}\left(X\right), \mathrm{Repeat}\left(\mathrm{GMPool}\left(X\right)\right)\right), \tag{1}$$

ただし，X，Y はそれぞれ入力・出力の特徴ベクトルを並べた行列，$\mathrm{Concat}(\cdot,\cdot)$ は複数の特徴ベクトルを結合する操作，$\mathrm{Repeat}(\cdot)$ は特徴ベクトルを複製する操作である．

2.2 MLP Block

MLP Block は，Poolformer [16] に倣い，最初の Residual Block において特徴ベクトル間の関係性を疎にモデリングする．続く Block は Poolformer と同じく特徴ベクトルごとに MLP 処理を適用する．

1 つ目の Residual Block の処理は次式のように表される．

$$Y = \mathrm{ConcatPool}(\mathrm{Norm}(X)) + X, \tag{2}$$

ただし，$\mathrm{Norm}(\cdot)$ は正規化層であり，$\mathrm{ConcatPool}(\cdot)$ は次式にしたがう学習可能な層である．

$$\begin{aligned}&\mathrm{ConcatPool}(X)\\ &= \sigma\left(W \cdot \mathrm{Concat}\left(X, \mathrm{Repeat}\left(\mathrm{LMPool}\left(X\right)\right)\right)\right),\end{aligned} \tag{3}$$

ここで，$\sigma(\cdot)$ は活性化関数，W は学習可能な重みである．

2.3 弱教師あり時空間行動検出

従来研究 [1] では，MIL に基づく弱教師あり学習の枠組みで，動画単位に付与された行動ラベルのみ教師データに用いて，インスタンスごとの行動を求める．本研究でも同様に MIL の枠組みを踏襲しつつ，従来研究と異なる骨格

表 1: Kinetics-400 データセットを用いた従来法との行動分類精度の比較結果．Total FPS は KP 検出を含むシステム全体の処理速度．

Method	Top-1 Acc. (%)	KP Detector	COCO AP_{kp} (%)	Runtime (ms)	Total FPS
ST-GCN [15]	30.7	OpenPose	56.3	4.0	85.4
2s-AGCN [12]	36.1			27.6	84.8
MS-G3D [8]	38.0			28.2	84.8
Ours w/o objects	38.9			9.8	85.2
MS-G3D [8]	45.1	HRNet	74.6	28.2	8.8
PoseConv3D [4]	47.7			960.0	8.5
Ours w/o objects	50.3			9.8	8.8
Ours w/o objects	43.1	PPNv2	36.4	9.8	1913.3
Ours w/ objects	52.3			11.2	1896.3

ベース行動認識の文脈において，インスタンスごとに行動を認識する弱教師あり時空間行動検出手法を提案する．

推論時，StructPool は入力の KP 群の特徴量を途中インスタンス単位に集約する．したがって，Grad-CAM [11] といった，DNN の出力への入力値の寄与を可視化する手法と StructPool を組み合わせることによって，各インスタンスが出力の行動クラスに寄与した度合いを定量化できる．具体的には，推論時に各行動のクラス確率を求めた上で，Grad-CAM を用いてインスタンスごとに特徴量の寄与度を算出する．任意の行動クラスに対して求めた寄与度とクラス確率をしきい値処理することによって，各インスタンスがその行動を取っているか判定する．

3. 評価実験

3.1 データセット

行動認識の大規模データセットの 1 つである Kinetics-400 [3] を用いて提案法の有効性を検証した．Pose Proposal Networks [10] [*1]（PPNv2）を Kinetics-400 の動画像に適用することにより KP を生成した．PPNv2 は MS-COCO データセット [7] を用いて関節点・物体輪郭点を同時に検出するよう学習され，骨格の定義は OpenPose [2] と同様である．物体輪郭点は 8 種類の輪郭上の端点により定義した．動画はすべて 30 [FPS] に変換した上で，320×224 [px^2] の解像度にリサイズし KP 検出に用いた．学習時と推論時両方において，各フレームで検出されたバウンディングボックスの確信度上位 2 人分の骨格，および上位 1 つの物体輪郭を入力として用いた．

3.2 学習・評価の設定

StructPool の学習には，学習率が 0.12，重み減衰項が 0.00005 の Stochastic Gradient Descent を用い，交差エントロピーを損失関数として 150 エポック学習した．その際，学習率はエポックが進むにつれて線形に減衰させた．また，Point embedding 層の出力ベクトルを 256 次元に，MLP Block における入力ベクトルの次元を $256, 512, 1024$ にそれぞれ設定した．各 MLP Block の繰り返し回数はすべて 2 とした．従来法 [4] と同様，入力する KP の画像座標を KP 間の相対座標に置き換えた DNN を用いて，アンサンブル学習をおこなった．処理速度の計測には，Intel i7-10700K CPU，32GB RAM，および GeForce RTX 3080 Ti GPU を用いた．

[*1] 国際出願番号 PCT/JP2020/040222

3.3 従来法に対する比較実験

骨格ベース行動認識の従来法と提案法の行動認識精度（以下，精度）を比較した結果を表 1 に示す．同表より，PPNv2 により検出された関節点と物体端点両方を入力として用いる提案法（w/ objects）は，関節点のみ用いる場合（w/o objects）に対して精度（Top-1 Acc.）が 9.2 [%pt] 向上している．また，提案法の行動クラスごとの精度を評価したところ，“*grooming dog*” や “*using computer*” といった行動クラスでは，物体端点なしではそれぞれ 36 [%], 35 [%] であった精度が，物体端点の導入により 36 [%pt] 以上向上していた．一方，5 [%pt] 以上精度が劣化した行動クラスは全体の 9 [%] に留まる．以上を踏まえると，物体端点の導入は，外観ベースの従来法の課題 [6] であった外観の過学習なしに，課題 1（1 章参照）で述べた認識可能な行動の拡大を実現できることがわかる．

また，同表より，KP 検出に OpenPose [2], [15] または HRNet [4], [13] を用いた場合の両方において，提案法は従来法 [4], [8] より高精度でありながら，それぞれ 1/3・1/98 の処理時間（Runtime [ms]）を実現している．加えて，KP 検出を含むシステム全体の処理速度（Total FPS）を改善することを目的に，KP 検出精度（COCO AP_{kp}）が著しく低い PPNv2 を用いる場合においても，提案法（w/o objects）は，OpenPose を用いる MS-G3D [8] を超える精度（+5.1 [%pt]）と 1/3 の処理時間を実現している．したがって，StructPool は課題 2 で述べた KP 検出精度の劣化に対して頑健化しつつ，KP 同士の関係を効率的にモデリングできることがわかる．

表 2: Kinetics-skeleton データセット [15] を用いた KP 検出誤り率に対する行動分類精度の比較結果.

	KP detection error ratio (%)					
Method	0	10	30	50	70	90
MS-G3D [8]	38.0	31.6	23.8	19.1	13.9	9.3
Ours	38.9	37.6	36.1	33.8	31.3	27.9

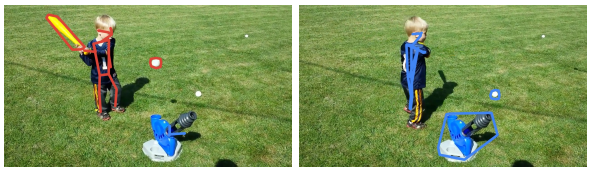

図 2: 弱教師あり時空間行動検出結果．赤は出力の行動ラベル（“*hitting baseball*”）に対する寄与度がしきい値以上の KP を持つインスタンス.

3.4 KP 検出誤りに対する頑健性の比較実験

また，KP の未検出・誤検出や人物追跡の誤りといった KP の検出誤りに対して GNN を用いる手法 [8] と提案法の精度を比較した結果を表 2 に示す．同表では，KP の検出誤りの頻度を変えた場合の精度を比較している．具体的には，関節点の未検出については，一定の頻度で関節座標と確信度を 0 に置換し，誤検出についても一定の頻度で正規分布にしたがうノイズを関節座標に付与した．例えば，KP 検出誤りの頻度が 50 [%] の場合には，全関節点の 50 [%] で未検出と誤検出を疑似的に発生させ，また，300 フレームの動画の中で 75 フレーム間隔を空けて人物 ID をランダムに入れ替えた．同表より，従来法 [8] に比べ，提案法は KP 検出誤りの増加に伴う精度劣化が少ないことから，課題 2 で述べた KP 検出誤りに対して高い頑健性があることがわかる.

3.5 弱教師あり時空間行動検出の定性的評価

本節では，各インスタンスの行動クラス推論に対する寄与度を可視化することで，弱教師あり時空間行動検出の実現可能性を検証する．KP ごとの寄与度を可視化した結果を図 2 に示す．同図では，各画像の動画が分類された “*hitting baseball*” クラスに対する各 KP の寄与度を可視化しており，子供がバットでボールを打つフレーム（左図）で，ボールを打つ動作に関係する物体（バットとボール）と人が検出されている．ボールを打った後，ボールがカメラの画角から外れ，バットが体に隠れたフレーム（右図）では，左図同様の検出はされなかった．同図より，提案法は動画単位に付与された教師ラベルのみ学習に用いて，インスタンスごとの行動を識別できることがわかる.

4. まとめ

本稿では，骨格ベース行動認識の従来研究が持つ，(a) 認識対象行動の拡大，(b) KP 検出誤りの影響の低減，(c) アノテーションコスト，の 3 つの課題を同時に解決することを目的として，関節点だけでなく物体端点を入力として用いて，各点の属するインスタンスやフレームなどの事前知識を下に，特徴量を段階的かつ効率的に集約する StructPool を提案した．実験では，先述の課題 a〜b に対する提案法の有効性を検証し，また，課題 c に対して提案法が弱教師あり時空間行動検出に適用可能であることを示した．今後の課題として，提案した弱教師あり手法を定量的に評価し，特長を分析することがあげられる.

参考文献

[1] Arnab, A., Sun, C., Nagrani, A. and Schmid, C.: Uncertainty-Aware Weakly Supervised Action Detection from Untrimmed Videos, *ECCV* (2020).

[2] Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y.: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, *CVPR* (2017).

[3] Carreira, J. and Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, *CVPR* (2017).

[4] Duan, H., Zhao, Y., Chen, K., Shao, D., Lin, D. and Dai, B.: Revisiting Skeleton-based Action Recognition, *CVPR* (2022).

[5] Gupta, P., Thatipelli, A., Aggarwal, A., Maheshwari, S., Trivedi, N., Das, S. and Sarvadevabhatla, R. K.: Quo vadis, skeleton action recognition?, *IJCV*, Vol. 129, No. 7, pp. 2097–2112 (2021).

[6] Hara, K., Kataoka, H. and Satoh, Y.: Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?, *CVPR* (2018).

[7] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, *ECCV* (2014).

[8] Liu, Z., Zhang, H., Chen, Z., Wang, Z. and Ouyang, W.: Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition, *CVPR* (2020).

[9] Qi, C. R., Su, H., Mo, K. and Guibas, L. J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, *CVPR* (2017).

[10] Sekii, T.: Pose Proposal Networks, *ECCV* (2018).

[11] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, *ICCV* (2017).

[12] Shi, L., Zhang, Y., Cheng, J. and Lu, H.: Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition, *CVPR* (2019).

[13] Sun, K., Xiao, B., Liu, D. and Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation, *CVPR* (2019).

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *NeurIPS* (2017).

[15] Yan, S., Xiong, Y. and Lin, D.: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, *AAAI* (2018).

[16] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J. and Yan, S.: MetaFormer is Actually What You Need for Vision, *CVPR* (2022).

[17] Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J. and Gan, C.: Graph Convolutional Networks for Temporal Action Localization, *ICCV* (2019).