G-5

# Set Autoencoder および Transformer を用いた骨格ベース行動認識モデル構築の検討

**Construction of Skeleton-based Action Recognition Model using Set Autoencoder and Transformer**

○岡﨑丈二[1]，香取照臣[2]
*Joji Okazaki[1], Teruomi Katori[2]

Abstract: At university, some students hesitate to enter a laboratory, sometimes, they loiter in front of a door. In this study, we equip a security camera in front of a laboratory, detect student behavior, and construct a system which notifies members when a visitor arrives. In this paper, we propose a skeleton-based action recognition model which can adapt to multi-person action by combining a pose estimation model, a Set Autoencoder, and a Transformer. We evaluate the proposed method using the UCF101 dataset.

## １．はじめに

大学教育の現場では，研究室の場所がわからないことや，訪問が失礼に当たらないかを懸念することなどにより，学生が研究室への入室をためらい，研究室前付近をうろつく行動を起こすことがある．このような場面では，主に 1 人で悩んでいることが多く，他人からの声掛けによって懸念が解決，または緩和されることがある．そこで，研究室前の廊下に監視カメラを設置し人物の行動を検出することで，研究室にいる人物から積極的な声掛けに繋げることを目的とした研究を行っている．

先行研究では，時空間画像での左右の折り返し回数[1]や，オプティカルフローを用いた人物の移動軌跡をもとにためらい行動の検出を行っていた[2]．

行動認識のアプローチとして，RGB 画像による手法と骨格情報による手法が挙げられる．RGB 画像による手法は人物の外観や背景などの資格情報を用いて行動を認識するものである．しかし，人物の体系や服装，環境光の影響などの行動認識に不必要な要素が多く含まれている．そのため，人物の動きにのみ注目した要素を抽出する必要がある．

一方で，骨格情報による手法は人間の関節座標を用いて行動認識を行うため，外観の変動などに比較的頑健であるが，関節座標同士の対応付けなどにより，複数人物への対応が困難である．

また，認知心理学の観点からも，人間の行動認識は脳の背側経路を使用することで関節点からでも素早く正確に認識することができること[3]から，骨格情報による手法は理論的に妥当性を持つと考えられる．

本論文では，姿勢推定モデル，Set Autoencoder，Transformer による複数人物の骨格ベースの行動認識モデルを構築し，UCF101 による性能評価について述べる．

## ２．骨格情報による行動認識モデル

### ２．１　姿勢推定モデル

姿勢推定モデルとは，人物の骨格情報を予測する畳み込み深層学習モデルである．本論文では，Ultralytics が公開した 17 点の関節点を出力する YOLOv8 Pose を使用し，連続フレーム間の関節点の差分をとることで移動ベクトルとしてモデルの入力に適用する．

### ２．２　Set Autoencoder

Set Autoencoder とは，順序を持たない要素集合を対象としたオートエンコーダモデルである[4]．入力集合の各要素から集約処理などによって潜在ベクトルを生成することで，要素数が変動する集合に対しても固定次元表現が得られる．本手法では，各フレームに映る複数人物の骨格情報の移動ベクトルを潜在ベクトルに変換するために適用する．

### ２．３　Transformer

Transformer とは，自己注意機構を用いて系列データを効率的に処理できる深層学習モデルである[5]．従来の系列処理モデルと異なり，系列全体の関係を捉えることができるため，長期の依存関係を効果的に学習できる．本手法では，Set Autoencoder によって得られた潜在ベクトルの系列データを Transformer に入力することで，行動分類ラベルを出力するために適用する．

## ３．骨格ベースモデルの学習

本論文では，姿勢推定モデル，Set Autoencoder，Transformer を使用した行動認識モデルを提案する．

動画の各フレーム$t$において，姿勢推定モデルを適用し複数人の骨格座標を取得し，$t-1$の骨格座標都の差分処理を行うことにより，骨格移動ベクトルを算出する．その後，複数人数分の骨格移動ベクトルを Set Autoencoder に入力し$N$次元の潜在ベクトルへ変換する．その後，動画の全フレームにおいて出力された潜在ベ

１：日大理工・院(前)・情報　２：日大理工・教員・情報

クトルを Transformer へ入力することで分類ラベルを出力する．

なお，Set Autoencoder および Transformer は一括で学習を行う End-to-End 学習を使用し，データセットは 101 種類の行動カテゴリを持つ UCF101 を使用する．Figure 1 に骨格ベースの行動認識モデルの概念図を示す．

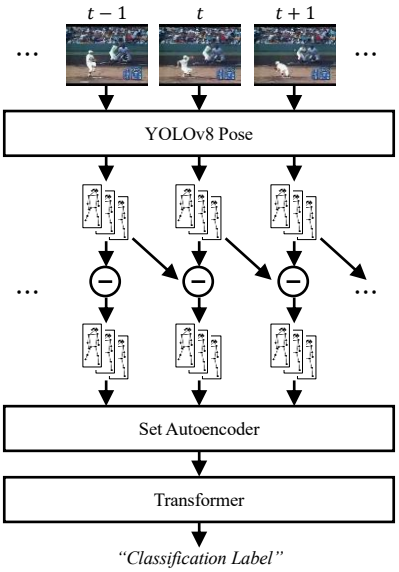


**Figure 1.**　A Schematic diagram of Skeleton-based model

## ４．結果と考察

Set Autoencoder および Transformer を End-to-End で学習した結果，テストデータに対する Top-1 精度は 0.514であった．UCF101 のランダム分類の Top-1 精度が約0.001であることを踏まえると，テストデータに対する top-1 精度はこれを大きく上回るため，本手法がデータから一定の認識能力を獲得したといえる．

Figure 2 に Set Autoencoder および Transformer の損失曲線を，Figure 3 に混合行列を示す．Figure 2 より，Set Autoencoder の損失は学習とともに単調減少し，収束している．一方，Transformer は後半で損失の減少が頭打ちとなり，過学習の兆候が表れている．また，Figure 3 より，モデルは全体的に多くのクラスを認識しているが，クラスごとに精度のばらつきがあり，YOLOv8 Pose による欠損値の不均一が影響していると考えられる．

以上の結果から，複数人物の骨格を用いた行動分類に活用できる．

## ５．まとめ

本論文では，Set Autoencoder および Transformer を用いた骨格ベースモデルを用いて UCF101 の分類を行い，一定の認識能力を獲得できた．今後は，骨格ベースモデルの精度向上および，RGB ベースモデルの作成を行い，2 モデルを結合することにより，行動認識モデル全体の拡充を行う．

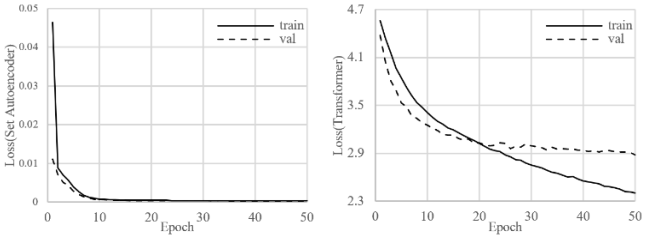


**Figure 2.**　Loss curve of Set Autoencoder

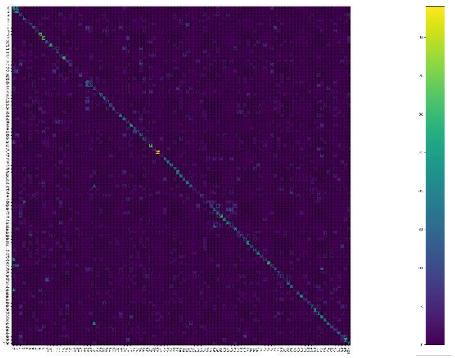

**Figure 3.**　Loss curve of Transformer

## ６．参考文献

[1] 海老原優太，香取照臣，泉隆:「時空間画像処理での滞在時間と折り返し回数によるドア前徘徊行動の自動検出」，電気学会知覚情報・次世代産業システム合同研究会，PI-21-12/IIS-21-25，pp.13-18 (2021-3)

[2] 篠原巧，香取照臣:「オプティカルフローによるドア前におけるためらい行動検出の検討－移動軌跡と滞在時間を用いた複数人物の検出－」，電気学会知覚情報・次世代産業システム合同研究会，PI-23-031/IIS-23-036，pp.11-16 (2023-3)

[3] Gunnar Johansson: “Visual perception of biological motion and a model for its analysis”, Perception & Psychophysics, Vol. 14, No. 2, pp.201-211 (1973)

[4] Martin Probst, Babak Alipanahi, Matthew G. Heller: “Set Autoencoder: Unsupervised Representation Learning for Sets”, International Conference on Learning Representations (2018)

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: “Attention Is All You Need”, Advances in Neural Information Processing Systems, Vol. 30, pp. 5998–6008 (2017)