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



# 事前学習済みDNNを用いたゼロショット異常行動認識

佐藤 文彬[1,a)]　八馬 遼[1,b)]　関井 大気[1,c)]

## 概要

骨格ベース異常行動認識において応用や性能面で拡張性の低下を引き起こすドメインシフトや正常データ不足の課題を同時に解決することを目的として，本研究では，特定行動認識の大規模データセットで事前に学習された DNN を用いて，異常行動の観測データ・教師データなしに，ユーザーがテキスト入力した説明文に基づき異常を具体化し識別するゼロショット学習手法を提案する．学習時は，DNN の出力空間上で重みの更新なしに正常データの分布をモデル化し，推論時の異常度の計算に用いる．また，異常行動の説明文の特徴量に基づき異常度を重み付けすることにより，正常行動に関する情報を間接的に補い正常を異常と誤るのを抑制する．実験では，暴力行動分類のデータセットである RWF-2000 を用いて，教師あり学習に基づく従来研究と提案法の性能を比較し，先述の課題それぞれに対して提案法の有効性を検証した．

## 1. はじめに

カメラを用いて人物が取る異常な行動を認識する異常行動認識は，事故や犯罪を予防するために必要不可欠な技術として期待される．異常行動認識の従来研究は，外観特徴 [5] または骨格情報 [8] を用いるアプローチに大別される．前者は，カメラ映像に Deep Neural Networks（DNN）を適用して得られる，人物の外観に関する特徴量を用いて，異常行動発生の有無を識別する．一方，後者は，各フレームに複数人物骨格検出を適用して得られる骨格情報のみを用いるため，前者よりも学習時と異なる認識対象の見えの変化に頑健である [15].

加えて，従来研究は，カメラ映像の各フレーム [8] または動画単位 [3] で異常行動発生の有無を判定する．また，学習には教師あり学習 [3] または教師なし学習 [8] が用いられる．以上を踏まえ，本研究では，教師データ作成の人的コストが少なく，また，認識対象の異常行動に対する限定が比較的少ない特長を鑑み，人物の正常な行動を撮影した動画（正常データ）から検出した骨格情報を教師なし学習に用いて，動画ごとに異常の有無を識別する．ただし，異常行動の定義はユーザーから与えられ，正常行動は異常行動以外の行動すべてを指すものとし，正常データの学習時に観測されなかった行動を例外行動と呼ぶ.

近年，本研究と同様の問題設定の手法 [8] が提案されているが，以下に述べるように，利用シーンの拡大や性能の改善に際して拡張性が低い.

**課題 1. ドメインシフト**

経時的な変化を含む複数シーンの間で正常データの特徴や分布が異なる（ドメインシフトが発生する）場合，シーンごとに正常データを収集し DNN を学習する必要があり，計算資源や学習時間，人的・エネルギーコストがかかるため，利用できるシーンが制限される.

**課題 2. 正常データの不足**

多様な正常データを得られない場合，広い範囲の行動を異常と定義せざるを得ない．したがって，対象の異常行動をユーザーが意図的に限定できることが望ましい．例えば，歩行シーンのみ正常データとして観測できる状況において，ユーザーから指定された暴力行為のみ異常として検知し，走行シーンを正常と判定するケースである.

これらの課題を同時に解決することを目的として，本稿では，時系列の骨格情報を入力として事前に学習された DNN を用いて，異常行動の観測データ・教師データなしに，ユーザーがテキスト入力で指示した異常行動を識別するゼロショット学習手法を提案する．具体的には，課題 1 のドメインシフトの影響を低減するため，比較的大規模な特定行動認識のデータセット（Kinetics-400 [2] など）を用いて DNN を骨格特徴抽出器として事前に学習する一方，正常データは DNN の重み更新に用いず，骨格特徴の特徴空間上で正常データの分布をモデル化するのに用いる.

また，従来法 [16] の多くは，Graph Neural Networks（GNN）などを用いて同一人物の時系列の関節間で特徴量を伝搬できることを前提としているが，シーンの経時的変化により関節の未検出や誤検出，人物追跡の誤りが発生した場合，この前提を満たせず頑健性が低下する．そこで本研究では，このようなドメインシフトに対して頑健性を高

[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)] fumiaki.sato1@konicaminolta.com
[b)] ryo.hachiuma@konicaminolta.com
[c)] taiki.sekii@konicaminolta.com

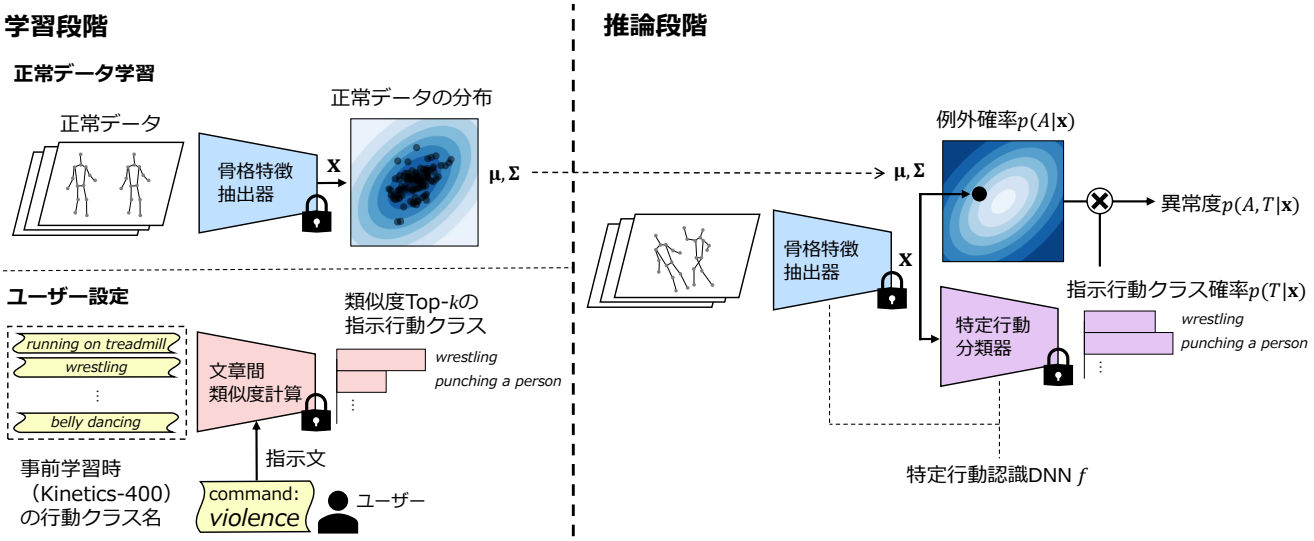


図 1: 提案法の概要

めるため，DNN の設計に PointNet [9] をベースとした構造を導入し，関節間の特徴量の伝搬が不要な骨格特徴抽出器を提案する．

課題 2 に対して，ユーザーから認識対象の異常行動をテキスト形式で受け付け，MPNet [12] により変換された文章特徴量に基づき異常度を重み付けすることで，正常行動に関する情報を間接的に補い，正常を異常と誤るのを抑制する．実験では，骨格特徴抽出器の事前学習に Kinetics-400 [2] を，異常行動認識精度の評価に暴力行動分類データセットである RWF-2000 [3] をそれぞれ用いて，課題 1〜2 に対する提案法の有効性をそれぞれ検証する．

## 2. 提案法

提案法の概要を図 1 に示す．推論段階では，はじめに動画像から検出した人物ごとの骨格情報を DNN に入力し，骨格特徴 $\mathbf{x}$ を抽出する．骨格特徴 $\mathbf{x}$ に対する異常度を，$\mathbf{x}$ が正常データに含まれず（例外である），かつユーザーが指示した行動であることを表す同時確率

$$p(A, T|\mathbf{x}) = p(A|\mathbf{x})p(T|\mathbf{x}) \tag{1}$$

により近似する．ただし，$A$，$T$ は 2 値の確率変数である．次節以降で右辺の各項について学習方法とともに詳述する．

DNN の設計には後述する PointNet [9] ベースの構造を採用し，あらかじめ特定行動分類のデータセット（Kinetics-400 など）を用いて事前に学習されているものとする．検出した関節点ごとの入力ベクトルは，画像座標，時刻，確信度，関節の種類を表す ID，および属する骨格の重心の画像座標を結合した 7 次元から成る．

### 2.1 例外確率

式 (1) の右辺第一項 $p(A|\mathbf{x})$ を，$\mathbf{x}$ が正常データに含まれない確率（例外確率）として，マハラノビス距離を用いて次式のように近似する．

$$p(A|\mathbf{x}) \sim \min\left(1.0, w_1\sqrt{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}^{w_2}\right), \tag{2}$$

ただし，$(w_1, w_2)$ は調整可能なパラメータである．$p(A|\cdot)$ は学習時に正常データの骨格情報の平均ベクトル $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{\Sigma}$ を計算することによって決定される．

なお，Rippel ら [10] は，異常検知の文脈において，事前学習済み DNN を用いて，DNN の重み更新なしに正常データの分布を求め，教師なし異常検知に活用している．提案法は，Rippel ら [10] の手法を新たに行動認識に応用し，例外確率 $p(A|\mathbf{x})$ を設計した．実験では，$p(A|\mathbf{x})$ のみを $\mathbf{x}$ の異常度として用いた場合においても，DNN の重み更新なしで教師なし異常行動認識を実現できることを示す．

### 2.2 指示行動クラス確率

式 (1) の右辺第二項 $p(T|\mathbf{x})$ は，$\mathbf{x}$ がユーザーの意図した行動の特徴量である確率であり，$\mathbf{x}$ と，学習段階にユーザーがテキスト入力する文章（以降，指示文）を MPNet により変換した特徴量 $\mathbf{y}_{\mathrm{in}}$ を用いて次式のように近似する．

$$p(T|\mathbf{x}) \sim \min\left(1.0, w_3\mathrm{KineticScore}(\mathbf{x}|\mathbf{y}_{\mathrm{in}})^{w_4}\right), \tag{3}$$

ただし，$(w_3, w_4)$ は調整可能なパラメータである．

$\mathrm{KineticScore}(\mathbf{x}|\mathbf{y}_{\mathrm{in}})$ を次のように定義する．本章冒頭で述べたように，骨格特徴を用いて行動を分類する DNN，$f$，は正常データに関係なく事前に学習されており，事前学習に用いたデータセットで定義された $S$ 個の行動クラスごとの分類スコア $(q_1, \cdots, q_S)$ を出力するものとする．また，$S$ 個の行動クラス名をそれぞれ MPNet を用いて特徴量 $\{\mathbf{y}_1, \cdots, \mathbf{y}_S\}$ に変換しておく．学習段階では，まず指示文を MPNet を用いて特徴量 $\mathbf{y}_{\mathrm{in}}$ に変換する．次に，$\{\mathbf{y}_1, \cdots, \mathbf{y}_S\}$ のうち，$\mathbf{y}_{\mathrm{in}}$ とコサイン距離の近い上位 $k$ ク

ラス（以下, 指示行動クラス）を特定する．ただし, $k$ は調整可能なパラメータである．推論段階で KineticScore($\mathbf{x}|\mathbf{y}_{\text{in}}$) を計算する際は，はじめに指示行動クラスに対する $\mathbf{x}$ の分類スコア $(q_1, \cdots, q_k)$ を求める．次に, KineticScore($\mathbf{x}|\mathbf{y}_{\text{in}}$) を指示行動クラスの分類スコアの最大値として次式のように求める．

$$\text{KineticScore}(\mathbf{x}|\mathbf{y}_{\text{in}}) = \max(q_1, \cdots, q_k). \quad (4)$$

### 2.3 DNN アーキテクチャ

本研究では，DNN の設計のベースに PointNet [9] を用いる．はじめに, 元の PointNet の構造から 3 次元点群処理向けの機構である T-Net を除外する．次に，Max-Pooling より浅い層の MLP（Multi-Layer Perceptron）を 6 層のボトルネック構造の残差ブロック [6] に置き換える．

## 3. 評価実験

### 3.1 データセット

DNN の事前学習に大規模行動分類データセット Kinetics-400 [2] を，また，ゼロショット異常行動認識の精度評価に暴力行動分類のデータセット RWF-2000 [3] をそれぞれ用いて提案法の有効性を検証した．Kinetics-400 は 400 クラスの行動分類タスクを対象として，また，RWF-2000 は暴力・非暴力 2 クラスの行動分類タスクを対象として，それぞれ YouTube [*1]から映像を収集したデータセットである．

両データセットにおいて，行動クラスがアノテーションされた動画の各フレームに Pose Proposal Networks [11]（PPN）を適用することにより, 時系列の骨格情報を作成した．PPN の DNN 構造には Pelee [14] を採用し，骨格の定義は OpenPose [1] と同様である．動画はすべて 30 [FPS] に変換した上で，$320 \times 224$ [px$^2$] の解像度にリサイズし骨格検出に用いた．

### 3.2 学習・評価の設定

各フレームにおいて，PPN で出力されるバウンディングボックス検出の確信度上位の骨格のうち，DNN の事前学習，異常行動認識の学習・評価それぞれにおいて，2 人分，5 人分の骨格を用いた．DNN の事前学習では，学習率が 0.18, 重み減衰項が 0.0001 の Stochastic Gradient Descent を採用し，交差エントロピーを損失関数として 100 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．提案法の学習時間の計測には，Intel i7-9700K CPU，32GB RAM，および GeForce RTX 2080 Ti GPU を用いた．

RWF-2000 を用いた提案法の評価では，非暴力行動を正常行動と定義し，暴力行動を異常行動と定義した．学習用に準備されている非暴力行動の動画のみを 2.1 節で述べた平均ベクトル $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{\Sigma}$ の計算に用い，事前学習を除き DNN の重みは更新しない．したがって，教師あり学習を用いる従来法とは異なり，提案法では，暴力行動の動画は異常サンプルとして評価にのみ用いられるため，学習において異常行動（暴力行動）に関する観測データ・教師データは一切利用されない設定（ゼロショット学習）である．指示文には，後述する暴力を表す文章を複数パターン評価し，最も精度が高かった “*punch or kick related to fighting*” を検証に利用した．このとき，2.2 節で述べた指示行動クラスとして，$k = 5$ クラス（“*punching person (boxing)*”, “*drop kicking*”, “*punching bag*”, “*side kick*”, “*sword fighting*”）が選択された．

表 1: RWF-2000 データセットを用いた従来法との暴力行動分類精度の比較結果．従来法は教師あり学習を利用．

| Method | Acc. (%) | Pose det. | $AP_{kp}$ | Train. time |
|---|---|---|---|---|
| PointNet++ [13] | 78.2 | | | |
| DGCNN [13] | 80.6 | RMPE [4] | 61.8 | - |
| SPIL [13] | 89.3 | | | |
| ConvLSTM [7] | 89.8 | - | - | 30h |
| Only Kinetics | 73.5 | | | 0 |
| Ours w/o text | 73.3 | PPN [11] | 36.4 | 15s |
| Ours w/ text | 79.0 | | | 15s |

### 3.3 従来法に対する比較実験

従来法と提案法の暴力行動認識精度（以下，精度）を比較した結果を表 1 に示す．同表より，指示文を利用する提案法（w/ text）は，教師あり学習を用いる従来法のうち，PointNet++や DGCNN と同等精度（Acc. 参照），かつ学習時間（15 [s]）は著しく短かった．一方, 2.2 節で述べた指示行動クラス確率のみを用いた場合（Only Kinetics）には，提案法（w/ text）の精度には及ばないことがわかる．加えて，2.2 節で述べた指示行動クラス確率を利用しない場合（w/o text）においても，教師あり学習を用いる PointNet++と約 5 [%pt] の精度差に達している．以上の結果を踏まえると，提案法は課題 1 で述べたドメインシフト（1 章参照）に対してシーンごとの DNN の学習なしにゼロショット異常行動認識を実現できることがわかる．また，提案法において，指示文を用いない場合（w/o text）と用いる場合（w/ text）で約 6 [%pt] 精度が向上していることから，課題 2 に対して，指示文が認識対象の暴力行動を具体化し正常行動に関する情報を補うことで，正常行動を異常と識別することを抑制できていることがわかる．

### 3.4 骨格検出誤りに対する頑健性の比較実験

関節点の未検出・誤検出や人物追跡の誤りといった経時的変化で発生しやすい骨格検出の誤りに対して提案法の頑

[*1] youtube.com

表 2: RWF-2000 データセットを用いた骨格検出誤りに対する暴力行動分類精度の比較結果.

| Method | Pose detection error ratio (%) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 10 | 30 | 50 | 70 | 90 |
| ST-GCN [16] | 82.0 | 77.0 | 69.5 | 64.3 | 62.3 | 60.5 |
| Ours | 79.0 | 76.0 | 75.5 | 76.0 | 74.3 | 74.8 |

表 3: RWF-2000 データセットを用いた指示文ごとの暴力行動分類精度の比較結果.

| Language direction | Acc. (%) |
|---|---|
| n/a | 73.3 |
| *violence* | 75.3 |
| *violent event* | 78.3 |
| *fighting event* | 77.0 |
| *punch or kick related to violence* | 78.5 |
| *punch or kick related to fighting* | 79.0 |

健性を検証した結果を表 2 に示す．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，骨格検出誤りの頻度が 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，150 フレームの動画の中で 38 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，GNN を用いる教師あり学習手法 [16]（著者らによる実装）に比べ，提案法は骨格検出の誤りの増加にともなう精度低下の幅が小さいことから，課題 1 で述べた経時的変化などのドメインシフトに対して比較的頑健であることがわかる．

### 3.5 指示文の表現のバラつきに対する比較実験

前節の提案法（w/ text）の検証に用いた 5 パターンの指示文ごとの精度を示す．指示文から求めた指示行動クラス確率を用いない場合（w/o text）よりも総じて精度が向上しており，指示文のバラつきが一定範囲内であれば，指示行動クラス確率が正常行動を異常と識別することを抑制できることがわかる．

## 4. まとめ

骨格ベース異常行動認識において応用や性能面で拡張性の低下を引き起こすドメインシフトや正常データ不足の課題を同時に解決することを目的として，本研究では，特定行動認識の大規模データセットで事前に学習された DNN を用いて，異常行動の観測データ・教師データなしに，ユーザーがテキスト入力した説明文に基づき異常を具体化し識別するゼロショット学習手法を提案した．実験では，暴力行動分類のデータセットである RWF-2000 を用いて，教師あり学習に基づく従来研究と提案法の性能を比較することにより，先述の課題それぞれに対して提案法の有効性を検証した．

実験において事前学習に用いた Kinetics-400 データセットと，評価に用いた RWF-2000 データセットの行動の類似点を検証していないことから，今後の課題として，多様なデータセットの組み合わせで提案法の有効性と汎化性を検証することがあげられる．

### 参考文献


[1] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. and Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, *PAMI*, Vol. 43, No. 1, pp. 172–186 (2021).

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

[3] Cheng, M., Cai, K. and Li, M.: RWF-2000: An Open Large Scale Video Database for Violence Detection, *ICPR* (2020).

[4] Fang, H.-S., Xie, S., Tai, Y.-W. and Lu, C.: RMPE: Regional Multi-Person Pose Estimation, *ICCV* (2017).

[5] Feng, J.-C., Hong, F.-T. and Zheng, W.-S.: MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection, *CVPR* (2021).

[6] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *CVPR* (2016).

[7] Islam, Z., Rukonuzzaman, M., Ahmed, R., Kabir, M. and Farazi, M.: Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM, *IJCNN* (2021).

[8] Morais, R., Le, V., Tran, T., Saha, B., Mansour, M. and Venkatesh, S.: Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos, *CVPR* (2019).

[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] Rippel, O., Mertens, P. and Merhof, D.: Modeling the Distribution of Normal Data in Pre-Trained Deep Features for Anomaly Detection, *ICPR* (2021).

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

[12] Song, K., Tan, X., Qin, T., Lu, J. and Liu, T.-Y.: MPNet: Masked and Permuted Pre-training for Language Understanding, *NeurIPS* (2020).

[13] Su, Y., Lin, G., Zhu, J. and Wu, Q.: Human Interaction Learning on 3D Skeleton Point Clouds for Video Violence Recognition, *ECCV* (2020).

[14] Wang, R. J., Li, X. and Ling, C. X.: Pelee: A Real-Time Object Detection System on Mobile Devices, *NeurIPS* (2018).

[15] Weinzaepfel, P. and Rogez, G.: Mimetics: Towards Understanding Human Actions Out of Context, *IJCV*, Vol. 129, No. 5, pp. 1675–1690 (2021).

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