［研究助成（A）］

エッジ AI デバイスを用いた監視システム実現のための行動分類

Behavior Classification for Realization of Surveillance Systems Using Edge AI Devices

2221021

研究代表者　長野工業高等専門学校　工学科　准教授　力丸彩奈

［研究の目的］

近年，自動車犯罪や車上狙いといった街頭犯罪の認知によりそれらの犯罪数は増加傾向にある。その対策として，ショッピングモールや駅などの公共施設に多くの防犯カメラや監視カメラが設置され，犯罪の抑制や事件発生後の捜査に活用されている。これらのカメラを事件発生前に活用することで犯罪を未然に防ぐことも可能になるが，防犯カメラの台数は多く，収集されるデータは大量になるため限られた人手ですべてを監視することは不可能である。特に不審な行動をする人物を見つけることは非常に難しい。本研究の目的は，防犯カメラ映像から自動で不審人物を発見する監視システムの実現である。

行動解析を可能にする技術として行動認識技術がある。カメラやセンサなどから得られるデータから人の行動の特徴を抽出し，それをもとに対象者の行動を分析する認識技術である。行動認識は応用の範囲も広く，防犯や介護を目的とした活用が期待できる。防犯用途では町中や施設にカメラを設置し，疑わしい動きをする人物を自動で検出することで犯罪の抑制が期待でき，介護分野においては施設内での異常行動の発検出や身体機能の継続的な観察を行うことで，事故防止・介護士の負担軽減が期待できる。本研究の最終目的は，図１に示すような行動認識技術を用いて日常行動の中に潜む危険行動を検出するシステムの構築であるが，本報告ではその前段階として，日常的な行動が行動の種類ごとに分類可能か検討を行う。

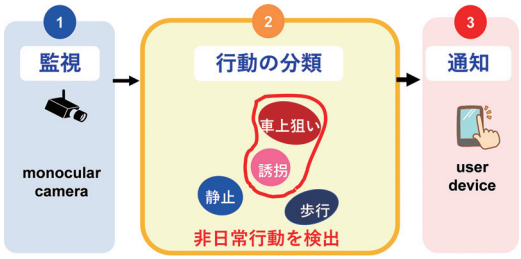

図１　本研究で想定するシステム概要

行動の分類には大量データの解析手法である機械学習を適用する。あらかじめ学習用データを学習することで未知データの解析が可能であり，製品検査など官能検査の自動化を可能にする技術として注目されている。その多くは「教師あり学習」であり，学習用データとそれに対応する分類種の名前（ラベル）を付けて学習を行うため，事前にデータの一つ一つにラベル付けを行うことが必須である。近年では教師あり機械学習アルゴリズムを用いた行動認識も注目され，特定の行動に対して高い認識精度を実現している[1]。しかし，行動は種類が多いため全ての行動に名前を付けることは困難である。また，ラベルを付けていない種類の行動に対しては判別が難しい。人の行動には曖昧なものや，動作の途中など名前の存在しない行動も数多くある。そこで本研究ではラベルを必要としない

「教師なし学習」を用いる。特定の行動を認識するのではなく，日常的な行動かどうかを分類することで様々な種類の行動に対応できるシステムを目指す。

［研究の内容，成果］

1．行動データの取得

本研究では行動データとして人の骨格情報を用いる。行動データは深度カメラやセンサを介して取得する方法[1]が多く，この方法では大型機器の設置やセンサ装着が事前に必要であり，対象者も限定される。それに対し，骨格データはすでに撮影された映像からの取得が可能である。骨格情報の抽出には機械学習による骨格位置推定手法を用いる。

機械学習による骨格位置推定手法には Top-down 型と Bottom-up 型の 2 種のアプローチが存在する。単一画像に複数人が含まれている場合，Top-down 型では撮影した画像に対して人間の検出を行い，各人物ごとに骨格推定を行う。一方 Bottom-up 型では，画像に存在する骨格点すべてに対し推定を行い，得られた骨格点について各人物ごとに結合させる。Top-down 型と比較して Bottom-up 型では複数人の骨格抽出時に処理速度が落ちにくいという利点が存在することから，本研究では，Bottom-up 型に属する Cao らの人物姿勢推定手法[2]を用いて骨格の抽出を行う。Cao らの方法では VGG-19[3]の一部を用いて特徴量の抽出を行っている。VGG-19 は物体認識に用いられる畳み込みニューラルネットワークであり，畳み込み層が 16 層，全結合層が 3 層，プーリング層が 5 層で構成されている。Cao らの人物姿勢推定手法ではまず，図 2 に示す構造の畳み込みニューラルネットワーク（Convolutional Neural Network: CNN）により解像度を 1/8 にまで圧縮した特徴マップ（Future Map）を生成する。図中の C は畳み込み層，P はプーリング層を表す。その後の骨格抽出部分では，骨

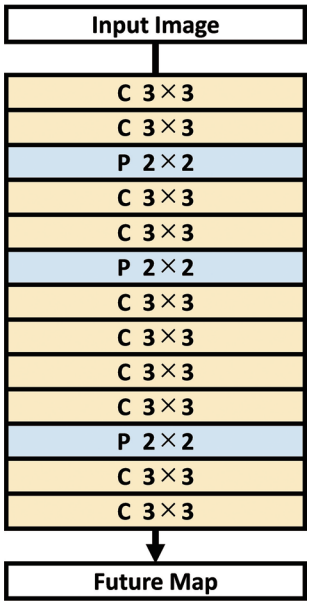

図 2　VGG-19 を用いたネットワーク構造

格推定を行うネットワーク・骨格結合の可能性推定を行うネットワークの二つに分岐するネットワーク構造を持つ。骨格推定を行うネットワークでは，関節位置の予測結果が得られ，骨格結合の可能性推定を行うネットワークでは，関節間の領域の全ピクセルに対し，方向ベクトルが定義される。これらのネットワークでの学習を繰り返すことで精度が向上し，複数人の同一骨格に対して互い違いにならないように骨格点とその結合が考慮され，マッチングに基づいた複数人の骨格点の推定を行う。取得する骨格点の種類について，実際に取得した骨格データを画像に重ねた結果を図 3 に示す。

使用したモデルは COCO2016[4]のデータセットに基づく学習モデルである。このモデル

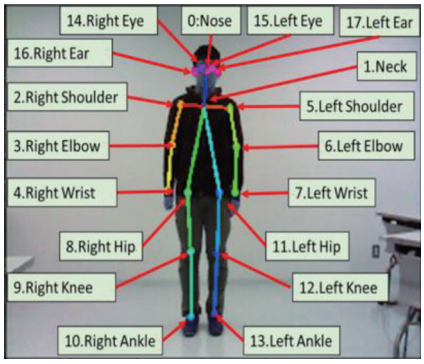

図 3　映像から抽出した 18 骨格点

においては，目，鼻，耳，首の付け根，肩，ひじ，手首，臀部，膝，足首の計 18 個の骨格位置についてｘ，ｙ座標と尤度を取得することができる。

近年の機械学習の発達に伴い，様々な分野での研究が行われている。リアルタイムでの実行処理を行う場合，高い性能を有するハードウェアが必要になるが，今後の IoT デバイスの増加に伴い，データのやり取りの増加による通信障害や消費電力の増加による電力不足という問題が予想される。そのため，通信遅延が発生しないこと，かつ低消費電力で実行することが求められる。このような需要に対し，近年では機械学習モデルを実行するための高コストパフォーマンスを持つエッジ AI デバイスが製品化されつつある。リアルタイムで推論を行うためにもエッジ AI デバイスは欠かせないものであり，それらを用いることで通信コストの削減や省スペースも可能となる。本研究でもエッジ AI デバイスを用いた通信量の削減を提案する。最終目的である監視システムでは，骨格抽出は映像取得後にカメラ側で行うエッジ処理を想定している。単眼カメラを用いて撮影し，その画像に対してエッジ AI デバイス上で骨格抽出処理を行い，得られた骨格情報をサーバへと送る。サーバ上において行動の認識と異常行動の認識を行い，異常行動となった場合に，監視員等のユーザ端末に対してアラートを通知するシステムを想定する。従来のシステムでは，サーバへ画像自体をリアルタイムで送信するため，対象となる施設に十分な帯域を持つネットワーク設備が必要であり，また，特徴量の生成に機械学習による推論を行う場合，通信とサーバにおける処理の負荷が増大し，リアルタイム性を損なうことが考えられる。しかし本研究で提案するシステムでは情報量の少ない骨格データをサーバへ送信するため，通常のカメラ画像による監視と比較して通信回線への負荷の減少が見込まれる。画像をそのままサーバへ送信する場合，1 画像を 640×480pixel とした際に 921 KB 程となるが，骨格点では 0.144 KB となり，1/6400 程度のデータ量削減が見込まれる。従来システムのサーバ負荷を分散させることもでき，サーバにおける処理を複雑化させるなどの拡張性を確保できる。これらの利点により，施設内で複数台の同時動作時においてもシステムの安定稼働が期待できる。

2. 行動データの分類

本研究では日常行動の分類方法として自己組織化マップ（Self-Organizing Map: SOM）[5] による分類を行う。SOM はニューラルネットワークに基づいた位相保存写像であり，多次元データのパターン認識や分類を可能とする。従来のクラスタリングでは不可能であった大規模なデータのクラスタリングや非線形データの自然なクラスタリングが可能であることから，本研究では SOM による分類方法を採用する。データの圧縮にも有効であり，複雑かつ高次元のデータを相互間の位相関係を保存しながら低次元へマッピングを行うため，データ間の差異が可視化できる。SOM は機械学習アルゴリズムのうち教師なし学習に分類される。通常，SOM によって学習を行った出力層は彩色を施して用いる。本研究では U-matrix 法 [6] によって出力層の各ユニット間の距離情報の 2 次元への視覚化を行う。

3. 取得した行動種類と分類結果

本研究では図 4 に示す 4 種類の行動について

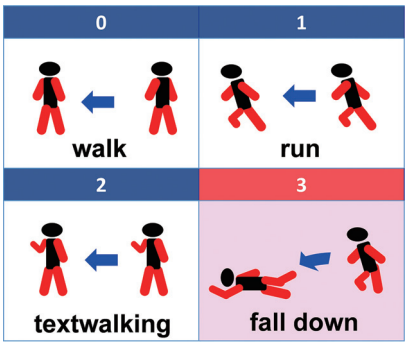

図 4　本研究で取得した行動イメージ

分類を行う。

0. 歩く
1. 走る
2. スマートフォンを所持して歩く
3. 転倒する（歩行 2 歩目で転倒）

これら 4 種類の行動について 30 フレーム分の映像を取得した。それを均等に間引いた 10 フレームから骨格データを取得し，分類のための行動データとした。各フレームから取得する 18 個の骨格位置(x,y)，間引き後の連続する 2 フレーム間の差分(v_x,v_y) を入力データとする。差分(v_x,v_y) を各骨格点の移動速度をとする。本実験では 3 番の転倒を非日常行動と想定している。

4 種類の行動について各行動 1 回のデータを取得し，SOM による分類を行った結果を図 5 に示す。図中の点で示される部分にデータが分布し，その隣の番号は入力データ番号である。この番号は図 3 に示した行動番号に対応している。黄色や赤で示す線は各データの境界線を示し，緑から赤に近づくほど他のデータと違いが大きいことを意味する。図 5 の出力結果では 3 番データを囲む境界線が少し赤くなっていることから，転倒行動が他の行動と異なる行動であることを示す結果となった。また，SOM による出力ではデータの類似度が高いほどデータは近くに配置される。0 番，2 番の行動が近くに配置されていることから，「歩く」「スマートフォンを所持して歩く」の行動は似ている行動であることを示す。これらの結果は，入力した 4 種類の行動が正しく分類されていることを示している。

次に，入力データ数を増やしての実験をおこなった。4 種類の行動について 3 回ずつ撮影し行動データを作成した。分類結果を図 6 に示す。図中のデータ番号については，0・1・2 が「歩く」，3・4・5 が「走る」，6・7・8 が「スマートフォンを所持して歩く」，9・10・11 が「転倒する」を示している。図 5 と比較すると入力データ数が増加したため，境界線も複雑になった。図の左上には「転倒する」行動である 3 つのデータが近くに配置され境界線も赤く示されたことから，この 3 つは似た行動であり他のデータと大きく異なる結果となった。これは，非日常行動として想定した転倒が他と違う行動として分類できたことを意味する。

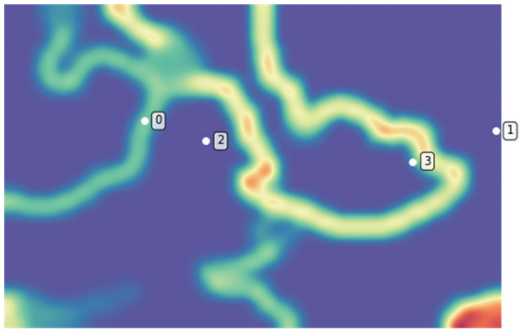

図 5　各行動 1 回撮影時の分類結果

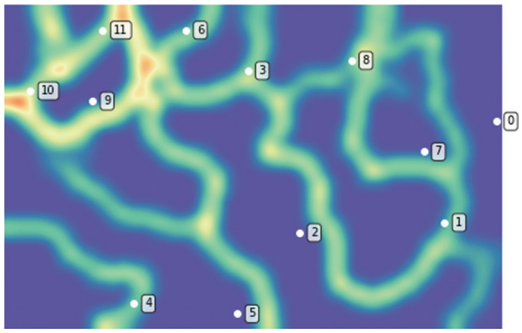

図 6　各行動 3 回撮影時の分類結果

一方で，他のデータを行動の種類ごとに見ると 3，4，5 は同一行動であるにも関わらず，離れて出力されたデータもある。これは転倒以外の行動が比較的似ていたため，このような結果になったと考えられる。データ数や行動種類を増やすことで改善できる可能性がある。

4. エッジ AI デバイスの選定

本研究で想定するシステムではエッジ AI デバイスを用いて骨格抽出を行う。エッジ AI デバイスの候補として複数のデバイスを選定した。図 7 にその一例を示す。左から Nvidia Jetson Xavier NX [8], Xilinx KV260 [9], Latte Panda v1 [10] である。Jetson Xavier NX は GPU が実装されているため，推論モデルの実行に

GPU を用いるため，サーバ等で作成した機械学習モデルをそのまま実装できるメリットがある。デメリットとしては消費電力が大きいことや高価であることが挙げられる。Xilinx KV260 は FPGA（Field-Programmable Gate Array）であるため，使用者が自由に処理プログラムを書き換えることが可能であり，CPU，GPU と比較すると汎用性が高く，消費電力が低い。しかし高度なプログラミング能力が必要であり，導入のハードルが高いことがデメリットである。また，Latte Panda は Windows を搭載したシングルボードコンピュータである。扱いやすく，誰でも導入できるが，短時間の使用で発熱しやすく長期間の稼働に対し不安が残る。

［今後の研究の方向，課題］

本研究では，4 種類の行動について行動データを作成し分類を行った。SOM による分類では，非日常行動として想定したデータの検出が可能であった。他の 3 種類の行動については正確に分類が行われなかったが，日常行動として想定していたため，日常・非日常行動の分類は正しく行われたと言える。しかし，最終的なシステムとしての精度を考慮すると，似た行動であっても行動の種類ごとに分類できることが望ましい。今後は行動の種類を増やしての分類や，撮影時間を変更してのデータの作成を行う。

また，本報告で選定したエッジ AI デバイスへの骨格抽出処理の実装は今後行う予定である。骨格抽出精度，消費電力，実装のしやすさや耐久性の観点からどのようなデバイスが適切か検討を行う。今後は図 8 に示すように，エッジ AI デバイスを処理フローに取り組んだシステム構築を目指す。

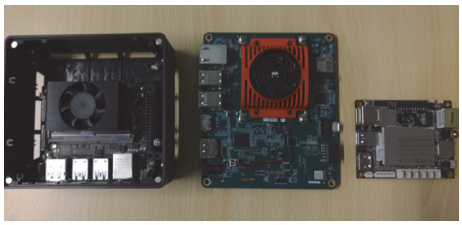

図 7　エッジ AI デバイス

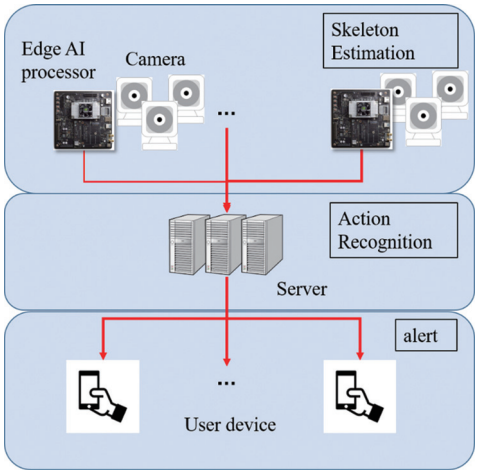

図 8　エッジ AI デバイスを用いた処理フロー

［成果の発表，論文など］

力丸彩奈：非日常行動検出のための教師なし機械学習による行動分類，電子情報通信学会総合大会，D-12-43，（2023）

［参考文献］

［1］ Minh, T. L., et al., A Fine-to-Coarse Convolutional Neural Network for 3D Human Action Recognition. 29th BMVC. 2018. p. 227.

［2］ 武田紳吾，Paula Lag，大北剛，井上創造，出野義則“モーションキャプチャを用いた行動認識におけるマーカー身体対応付け作業の削減，”マルチメディア，分散，協調とモバイル（DICOMO）シンポジウム，pp. 1280-1286（2019）

［3］ Cao, Z., et al., Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. CVPR. 2017. pp.7291-7299.

［4］ Simonyan, K. and Zisserman, A.: “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Published as a conference paper at ICLR, arXiv: 1409.1556（2015）

［5］ COCO dataset: https: //cocodataset. org（Accessed: 2023/1/18）

［6］ T. Kohonen, et al., Self-organized formation of topologically correct feature maps, Biological Cybernetics, 1982, 43（1）pp. 59-69.

［7］ A. Ultsch and H. P. Siemon. Kohonen’s self organizing feature maps for exploratory data analysis. INNC- 90. 1990 pp. 305-308.

［8］ Jetson Xavier NX: https://www.nvidia.com/en-

us/autonomous-machines/embedded-systems/jetson-xavier-nx/（Accessed: 2023/5/18）
［9］ KV260: https://www.xilinx.com/products/som/kria/kv260-vision-starter-kit.html（Accessed: 2023/5/18）
［10］ LattePanda v1: https://www.lattepanda.com/lattepanda-v1（Accessed: 2023/5/18）