H-010

人の骨格情報を用いた映像からの不審行動検知に関する検討
A Consideration on Detecting Suspicious Behavior from Video Images Using Human Skeleton Information

田　俊逸† 亀山　渉‡
Junyi TIAN Wataru KAMEYAMA

1. はじめに

近年、公共施設に多く設置された防犯カメラの映像を想定し、犯罪行為や不審行動の検知に関する研究が行われている。不審行動を検知する研究として、オプティカルフローを利用した方法等が提案されているが、映像中の輝度等の変化はオプティカルフロー検出に大きな影響を与えるため、一般的な防犯カメラでは得られない高品質な映像が必要である。また、不審行動を全て網羅するデータセットは存在しないため、全ての不審行動の特徴量を学習して抽出するのは困難である。

そこで本稿では、高品質でない映像でも比較的検出が容易な人の骨格情報を用いて行動の特徴量を学習し、不審ではない行動（以下、正常行動と呼ぶ）及び不審行動の特徴量の比較を教師なし学習によって行うことで不審行動を検知する手法を検討した。

2. 提案手法

提案手法を図 1 に示す。なお、本稿では、NTU RGB+D 120 データセット[1]を用いて実験を行った。

骨格情報の抽出には、OpenPose[2]と AlphaPose[3]の精度を上記のデータセットで比較した結果から、OpenPose を用いている。提案手法では、まず、データセット中の全ての映像から得られた骨格情報を用い、後述する修正を施した ST-GCN[4]を学習させる。次に、学習済みのネットワークを用いて、正常行動及び不審行動の特徴量を取得し、取得された特徴量を各教師なし学習で比較し、不審行動を検出する。

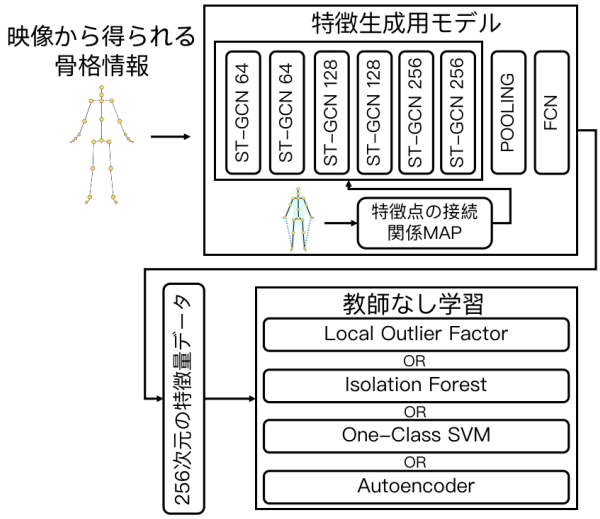

図 1　提案手法

† 早稲田大学大学院基幹理工学研究科，Waseda Univ.
‡ 早稲田大学理工学術院，Waseda Univ.

提案されている ST-GCN 中の畳み込み回数は 9 回であるが、予備実験にて、3 回、6 回、9 回、12 回の畳み込みを NTU RGB+D 120 データセットで比較したところ、6 回の AUC（Area Under the Curve）値が最も高かったため、本稿では、図 1 に示す 6 回の畳み込み層を利用する。

ST-GCN では骨格情報の時空間の接続関係を考慮しているが、身体的に隣接しているノード間の連結情報のみが考慮されている。例えば、腕と肘の連結設定はあるが、身体的に隣接していない腕と足の連結設定はない。しかしながら、人間の行動は、例えば、腕と足が完全に独立して動くことはなく、何らかの関係をもって両者が動くことが想定される。そのため、身体的に隣接していないノードでも関連性のあるものを連結設定に追加すると、より精度よく行動の特徴量を抽出できる可能性がある。

そこで、図 2 に示す 20 パターンの連結設定を用意し、比較を行った。図 2 で、黄色の丸は骨格ノード、黒の実線は ST-GCN でもともと考慮されているノード間連結情報、青の点線は新たに加えたノード間連結情報を示している。また、00 は ST-GCN のデフォルト設定である。データセット中の映像が膨大であるため、NTU RGB+D 120 の一部のデータ（正常行動から 631 映像を選択し、少なくとも 1 映像が後述する不審行動カテゴリに含まれるよう不審行動から 37 映像を選択）を使用し、後続の教師なし学習には Autoencoder を利用した予備実験の結果を表 1 に示す。なお、Autoencoder は 3 節に述べる設定で使用する。表 1 より、パターン 03、つまり、身体の同じ側の腕と足の連結設定を追加したものが最も良い結果であったことから、この連結設定を以下では用いた。

表 1　ノード連結パターンの比較

パターン	AUC 値	パターン	AUC 値
00	0.8089	10	0.8038
01	0.7356	11	0.6660
02	0.7882	12	0.7470
03	**0.8138**	13	0.6142
04	0.7542	14	0.7435
05	0.6731	15	0.7356
06	0.6022	16	0.6484
07	0.7584	17	0.7386
08	0.6256	18	0.6641
09	0.8047	19	0.7299

3. 不審行動検知

データセットのラベルを基に実際の映像内容を確認し、殴る（A50）、蹴る（A51）、人を押す（A52）、物を盗む（A57）、ものを使って殴る（A106）、ナイフで刺す（A107）、人にぶつかる（A108）、物を奪う（A109）、銃で撃つ（A110）を不審行動とし、それ以外は正常行動と

した。まず、全ての行動（114,480 映像）を 2 節で述べた修正 ST-GCN で学習し、学習済みのネットワークで、正常行動（105,888 映像）及び不審行動（8,592 映像）のそれぞれの特徴量を求める。

その後、教師なし学習として、Local Outlier Factor、Isolation Forest、One-Class SVM、Autoencoder を使用し、不審行動検知の比較を行う。ここで、それぞれの教師なし学習には、提案手法で得られた同一の 256 次元の特徴量データを入力する。それぞれに対する学習方法及びテスト方法を表 2 に示す。また、Autoencoder の実行条件を表 3 に示す。それ以外の手法では、scikit-learn のデフォルト値を使用して実行した。

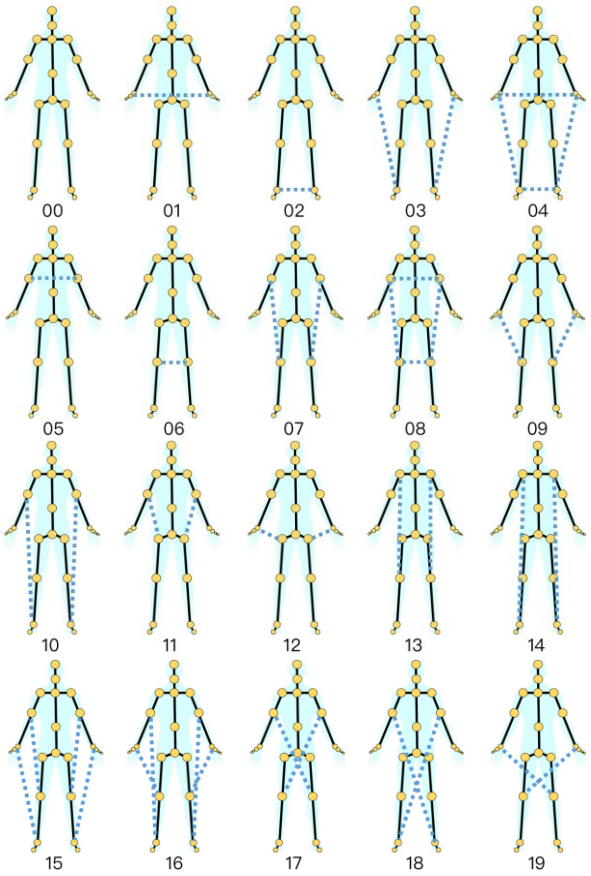

図 2　連結設定のパターン

表 2　学習データとテストデータ

検知手法	学習データ	テストデータ
Local Outlier Factor	80%の正常行動	20%の正常行動 100%の不審行動
Isolation Forest	—	100%の正常行動 100%の不審行動
One-Class SVM	80%の正常行動	20%の正常行動 100%の不審行動
Autoencoder	80%の正常行動	20%の正常行動 100%の不審行動

表 3　Autoencoder の実行条件

活性化関数	ELU
損失関数	MSE
最適化アルゴリズム	Adam
学習率	0.003
Epoch	600
層構造	256→128→64→16→8→16→64→128→256

Autoencoder では、入力と出力の二乗誤差に対して閾値を設定し、閾値を超えたものを不審行動（異常値）とする。ここで、閾値は検知結果の F 値が最も高くなる値とした。

5 分割交差検証による平均 AUC 値の比較を表 4 に示す。なお、Isolation Forest の場合は交差検証ではない。表 4 より、Autoencoder の平均 AUC 値が最も高く、約 80%の値が得られた。また、Isolation Forest が最も低い値となった理由としては、不審行動と正常行動が一部似ているため、学習を伴わない手法では、それらの判別が難しくなることが考えられる。

表 4　不審行動検知精度の比較

検知手法	平均 AUC 値
Local Outlier Factor	0.6664
Isolation Forest	0.5308
One-Class SVM	0.6896
Autoencoder	**0.7929**

4. まとめと今後の課題

本稿では、高品質ではない映像でも不審行動の判別ができるよう、人の骨格情報を利用した手法を検討した。既存の ST-GCN に対し、不審行動検知に適応した改良を加え、抽出された特徴量を教師なし学習で判定する手法を検討した。実験の結果、提案する ST-GCN の修正モデルと Autoencoder を組み合わせた場合、最も精度よく不審行動検知を行えることが分かった。

今後の課題としては、図 2 以外の接続関係を検討することが挙げられる。また、本稿で比較した教師なし学習以外の手法の検討も今後の課題である。

参考文献

[1] J. Liu, A. Shahroudy, M. Perz, G. Wang, L Duan, A. Kot, “NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding,” IEEE Tans. PAMI, Vol. 42, No. 10, pp.2684-2701, 2020.
[2] Z. Cao, G. Hidalgo, T. Simon, S. Wei, Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,” IEEE Tans. PAMI, Vol. 43, No. 1, pp.172-186, 2021.
[3] H. Fang, S. Xie, Y. Tai, C. Lu, “RMPE: Regional Multi-person Pose Estimation,” IEEE ICCV, DOI: 10.1109/ICCV.2017.256, 22-29 Oct. 2017, Venice, Italy.
[4] S. Yan, Y. Xiong, D. Lin “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in Proc. 32nd AAAI Conf. on Artificial Intelligence, 2018, pp. 7444-7452.