6S-06

深層学習を用いた万引きの予兆検知

シュレスタ　スザン†　谷口 洋司§　田中　哲雄‡
神奈川工科大学大学院工学研究科†　第一工科大学§　神奈川工科大学‡

1. はじめに

近年，万引き犯罪の増加が懸念されている．政府発行の犯罪統計[1]によると2022年の万引き認知件数は 83,598 件である．監視カメラの普及にも関わらず，労力と時間の制約により，効果的な監視が難しくなっている．小売業界における万引き犯罪は経済的損失やセキュリティ上のリスクを引き起こしており，これに対処する新しいアプローチが求められている．近年，通常行動と万引きを分類する手法も多く提案されている．一方，万引き後の検知よりも，万引きする前や万引きしそうな行動の認識の方が事前に犯罪の対策を講じることができより有用である．

そこで，筆者らが行ってきた研究[2]の検知方式を改良し，本研究では，カメラ映像を活用し，万引き前の不審な行動を自動的に検知するシステムを提案する．具体的には，防犯カメラで撮られた実際の万引き映像データを対象とし，深層学習を用いて，映像に映っている人物の行動を通常，不審，万引きに分類する．これにより，人的リソースの効率的な利用が可能となり，小売業界におけるセキュリティの向上と経済的損失の軽減が期待される．

2. データセットの作成

2.1. UCF101-Crime

実験に使用したデータセットは，約128時間の通常行動の映像と犯罪映像の2種類のデータを含む，UCF101-Crime[3] データセットである．犯罪ラベリングされた映像は，虐待，逮捕，放火，暴行，交通事故，強盗，爆発，喧嘩，強盗，銃撃，窃盗，万引き，破壊行為などの犯罪を含む様々な違法行為の発生を録画した映像で構成されている．本研究では，万引き犯罪に着目する．特に様々な店で買い物をする人々の動画が含まれている，通常クラスの動画は,商品を閲覧している人々である．異常クラスの映像には，それぞれの店で盗みを働く人々の映像が含まれている．データセットには 50 件の異常クラスの万引きの映像が含まれている．

2.2. 追加データ

上記のデータにはクラスのアンバランスがあるため，動画共有サービス Youtube[4]からいろいろな言語で検索し，万引き行動が映っている動画を取得した．これらをデータセットに追加し，85 個の異常クラスの動画を含むデータセットを構築した．

2.3. データセットの前処理

収集した映像データセットから，万引きや不信な行動を含まないシーンを通常映像，万引き前の不審な行動を含むシーンを不審行動映像，万引きを含むシーンを万引き映像として，手動で抽出した．これらの動画から 10 フレームごとにフレームを取得し 128×128 画素にリサイズした．リサイズしたフレームに通常，不審，万引きのいずれかのラベルを付け3クラスのデータセットを作成した．

3. モデル構成

動画の分類に最も広く使われている2つの深層学習アーキテクチャは，畳み込みニューラルネットワーク（CNN）とリカレントニューラルネットワーク（RNN）である．CNN は主に動画から空間情報を学習するために使用され，RNN は動画から時間情報を学習するために使用される．実際の現場では万引き前の不審な行動は数秒であるため，数秒以内の不審な行動をリアルタイムで検知するモデルを構築する必要がある．したがって，構築した訓練データセットを用いて，リソース制約のある環境でも高い性能を発揮する MobilenetV2[5]と BiLSTM[6]で通常行動，不審行動と万引き行動の３つのクラスに分類を行う検知モデルを構成する．

3.1. MobileNetV2

MobileNetV2 は，Google が提案した軽量で高効率な畳み込みニューラルネットワークアーキテクチャで，主にモバイルやエッジデバイスのリアルタイム画像認識や分類に適している．MobileNetV2 は新しい畳み込み演算 Depthwise Separable Convolution の導入によりモデルのパラメータ数が減少し，推論速度が向上する．

Detection of Early Signs of Shoplifting using Deep Learning
†Sujan Shrestha・Graduate Institute of Engineering, Kanagawa Institute of Technology
§Yoji Taniguchi・Daiichi Institute of Technology
‡Tetsuo Tanaka・Kanagawa Institute of Technology

3.2. BiLSTM

BiLSTM は，時系列的な特徴を捉える能力と双方向性により，各フレームの予測で前後のフレームの情報を考慮できる．畳み込みニューラルネットワーク（CNN）と組み合わせることで，動画内の複雑な動作や時間的な変化に対処し，空間的な特徴と時系列的な動きを効果的に捉えて行動認識の性能を向上させることが可能である．

本モデルで MobilenetV2 によって映像の各フレームから特徴を抽出し，その特徴ベクトルを BiLSTM ネットワークに渡す．データセット量が少ないため，Keras を利用し MobilenetV2 を ImageNet[7]データセットで事前学習された特徴を抽出する転移学習を行う．BiLSTM は，MobilenetV2 アーキテクチャから得られた特徴ベクトルを前後のフレームにわたって考慮することで，より豊かな時空間の情報を捉える役割を果たす．BiLSTM からの出力ベクトルは，全結合層およびソフトマックス関数を通して処理され，最終的にフレームの分類予測が出力する．

4. 実験方法と結果

データ量が少ないため，構築したデータセットを 9:1 で訓練データとテストデータに分割し，検知モデルで学習を行った．学習にあたって，

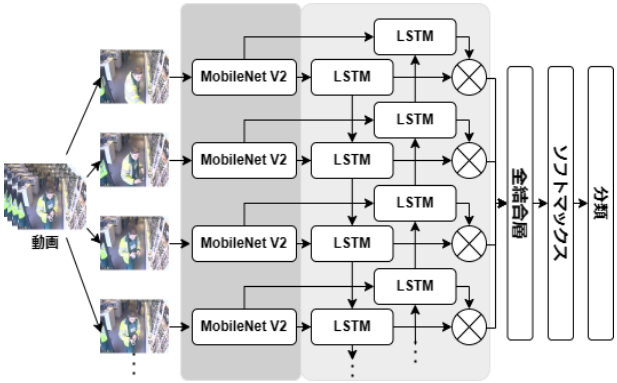

図 1 モデル構成図

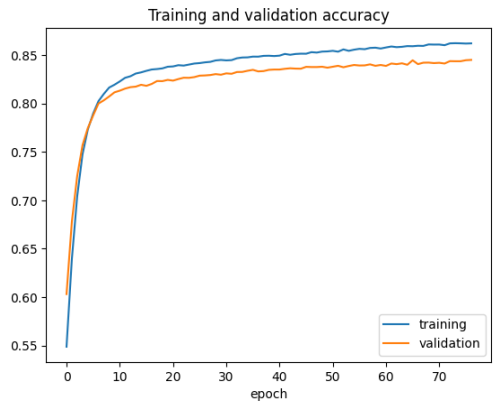

図 2 モデルの認識精度

77epoch が最適なエポック値だった為，77epoch まで学習した．77epoch で訓練データの認識精度は 86.21%，テストデータの認識精度は 84.54%となった．訓練データとテストデータの認識精度を図 2 から確認することができる．認識精度のグラフをみるとモデルはエポックごとに映像内の動的な情報を学習していることが分かる．

5. おわりに

本稿では，深層学習を用いて通常行動，不審行動，および万引き行動の動画を分類する行動認識モデルを提案した．このモデルでは，ImageNet データセットで事前学習された MobileNetV2 と，前後のフレームにわたって情報を考慮するために BiLSTM を組み合わせた．Youtube からの監視カメラ映像を含む UCF101-Crime データセットを使用し，通常行動，不審行動，万引き行動の 3 つのクラスに対するラベル付けデータセットを作成した．

実験の結果，提案モデルは映像内の動的な情報を学習できることが確認されたが，学習データが限られているため，その精度はまだ満足できるものではない．

今後は，映像のみでの実用レベルの万引き予兆検知を目指し，データセットの量を増やし，フレームごとに人物領域を検出し，特徴量の抽出を向上させるための方法を模索する．

参考文献

[1] 政府統計の総合窓口, https://www.e-stat.go.jp/, 2023 年 11 月 28 日参照.
[2] Shrestha Sujan 他，頭部方向推定を用いたキョロキョロ行動検知方式の提案，令和 5 年電気学会全国大会, 3-101, 2023.
[3] Sultani, Waqas, et al. “Real-World Anomaly Detection in Surveillance Videos.” ArXiv.org, 2018, arxiv.org/abs/1801.04264.
[4] Youtube,www.youtube.com
[5] Sandler, Mark, et al. “MobileNetV2: Inverted Residuals and Linear Bottlenecks.” ArXiv.org, 2018, arxiv.org/abs/1801.04381.
[6] Huang, Zhiheng, et al. “Bidirectional LSTM-CRF Models for Sequence Tagging.” ArXiv:1508.01991 [Cs], 9 Aug. 2015, arxiv.org/abs/1508.01991.
[7] Works CitedDeng, Jia, et al. “ImageNet: A Large-Scale Hierarchical Image Database.” 2009 IEEE Conference on Computer Vision and Pattern Recognition, June 2009, https://doi.org/10.1109/cvpr.2009.5206848.