2R-04

行動認識モデルを用いた監視カメラ映像での万引き検出

雑賀　智†　　中島　克人†

東京電機大学　未来科学部　情報メディア学科†

1 はじめに

長年，小売店は万引き被害による収益悪化の問題を抱えている．そこで我々は，監視カメラ映像からの万引き行動の自動検出を目指している．検出には人の行動検出に有効とされる 3D Convolutional Neural Network (3D CNN)[1]を用いる事とし，これに万引き行動を学習させて検出モデルを構築する．学習データに用いるための監視カメラ映像はプライバシーの問題もあり，殆ど公開されていない．そこで，3D CNN で構築された学習済みの人の行動認識モデルを元に，Web 上で少量ながら入手可能な万引き行動の監視カメラ映像を収集して追加の学習を行い，万引き，非・万引きの 2 値分類を試みた．本稿ではその手法およびその評価実験の結果を報告する．

2 関連技術

2.1 万引き検出システム

アースアイズ(株)が赤外線による深度センサー付きの監視カメラを用いた万引き検出システムを発表している[2]．しかし，実用上は多数の専用のカメラを設置する必要があるためコストは高くなる．我々は，既存の監視カメラからの映像だけで万引きの検出を行うことにより，低コストでのシステム実現を目指す．

2.2 3D CNN

深層学習では，一つの映像を自動解析して「走る」，「ジャンプする」などといった行動に分類する「行動認識」に関する研究が進んでいる．この分野で主要な技術である 3D CNN は Neural Network を構成するレイヤの一つであり，2D CNN から発展させたものである．2D CNN では一つの画像を 2 次元のカーネルで上下左右に畳み込むことにより空間特徴の学習ができるが，3D CNN では複数の映像フレームを 3 次元のカーネルを用いて，2 次元の空間に加えて時間方向にも同時に畳み込みを行う．これにより，時空間特徴の学習ができる（図 1）．本稿では 3D CNN の一種である後述の 3D ResNet[3]を用いて万引き検出モデルの構築を行う．

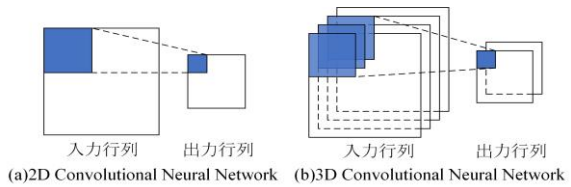

図 1 2D と 3D の Convolutional Neural Network

Shoplifting Detection in Surveillance Video based on Action Recognition Model

† Satoru Saika・Tokyo Denki University

† Katsuto Nakajima・Tokyo Denki University

3 提案手法

監視カメラ映像内で万引きを疑われる行動が検出されると店員に通知するシステムの構築を目指す．その実現のために，監視カメラ映像に対し，既存の一般物体認識器を用いて人物領域を矩形として切り出し，人の行動認識器でその領域内での「商品を服や鞄に入れる」という万引きらしき行動の検出を行う．

今回我々は，後者の万引き行動認識器のためのモデル構築とその評価を行う．

3.1 万引き行動データセット

本稿では Web 上から万引き動画を収集して時間と空間の切り取りを行い，「万引き（正例）」，「非・万引き（負例）」のラベル付けをして，356 個の動画を含むデータセットを構築した（図 2）．今回は，検出対象の万引きらしき行動は前述の「商品を服や鞄に入れる」という行動に限定し，それ以外を負例として扱うこととした．

図 2 作成した万引きデータセットの例

3.2 データセットの具体的な作成手順

最初に YouTube で「万引き」という単語を英語やロシア語などで検索することにより監視カメラ映像が映っている動画を 80 ほど収集した．

次に，収集動画の中で万引き行動とその前後の部分を，万引き行動にかかる時間として考えられる 10 秒単位で切り取った．そして，10 秒単位のそれぞれの動画に対し，物体検出の精度と速度の両方に優れた深層学習ベースの YOLOv3[4]の中で特に高い分類精度を誇る YOLOv3-spp を用いて人物領域を検出し，その周囲を 20%広げた領域を 300×300 画素にリサイズした．

最後に，時間と空間の切り取りが終わった動画に対して，手動で正例，負例のどちらかのラベルを付けた．今回の正例の数は 178 となった．負例の数は正例の数を超えたが，クラス間のデータの偏りは精度悪化につながると考えられるため負例の数を 178 になるまでランダムに削除した．

3.3 万引き行動の学習

構築したデータセットを用いて後述の 3D ResNet に「万引き」，「非・万引き」の 2 値分類を行う万引き検出モデルを構築する．

3.3.1 3D ResNet

静止画像分類タスクで高い精度を誇る Residual Network[5]の 2D CNN レイヤを 3D CNN レイヤに拡張させた 3D ResNet を今回の行動認識モデル構築に用いる．3D ResNet のベースモデルには，行動認識の汎化性能を高めるために数十万もの動画で構成された Kinetics データセット[6]で事前学習された物を使用する．

3.3.2 フレーム分割と間引き

動画を 3D CNN に入力するにはフレームへ分割する必要がある．フレーム数は本来任意数にできるが，今回実験で使う GPU のメモリ容量の制約から 64 フレームとする．今回 Web から入手した動画は約 30fps であり，時間が 10 秒であるため合計約 300 フレームから等間隔で 64 フレームを抜き出す．

3.3.3 オプティカルフロー

今回は映像内の動的な情報を学習するためにオプティカルフローを用いる．これは連続するフレーム間の物体の動きをベクトルとして表したものである．このベクトルの縦，横方向の動きを可視化したもの（図 3）を入力データとして用いる．今回はオプティカルフローの検出アルゴリズムとして，追跡対象の特徴点のみの移動ベクトルを求める「疎なオプティカルフロー」ではなく，フレーム内の全画素の移動ベクトルを求める「密なオプティカルフロー」を高速に検出できる Farneback[7]を用いる．

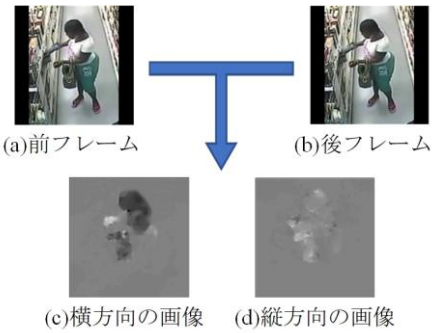

図 3 オプティカルフローの可視化画像

4 実験方法と結果

構築した万引きデータセットを 3:1 で訓練，テストデータに分割する事とし，万引き行動の学習と汎化性能の精度評価を 4 回の検証による交差検証で行った．入力データのチャンネルについては RGB の 3 チャンネルの場合と，オプティカルフローの縦，横方向の可視化画像をそれぞれ重ねた 2 チャンネルの場合で実験を行った．

100epoch までの訓練データにおける学習時の精度の推移を図 4 に，テストデータによる汎化性能の精度の推移を図 5 に示す．図 4 から RGB の 3 チャンネルでは学習の進みが速いのが分かる．図 5 から RGB での最高精度がオプティカルフローの 2 チャンネルより劣っていることが分かる．これは，今回作成したデータセットの 2 クラス間の空間情報の差が，ベースモデルでの行動認識データセットと比べると小さく，空間情報の学習がうまくできていない可能性を示している．一方，オプティカルフローの 2 チャンネルでは最終的に精度が 50%を完全に上回っており，これは映像内の動的な情報をわずかであるが学習できていることを示している．

今回の実験ではテストデータにおける最高精度が約 55%と実用にはまだまだ不十分となったが，RGB もオプティカルフローも学習データを増強することにより，更なる精度向上が期待できると考えている．

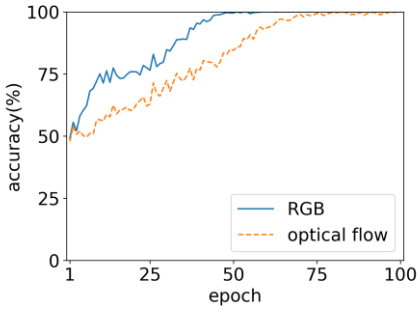

図 4 訓練データの認識精度(交差検証の平均)

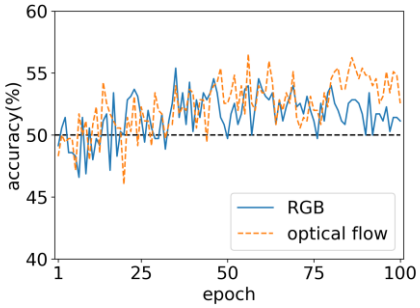

図 5 テストデータの認識精度(交差検証の平均)

5 おわりに

本稿では Kinetics データセットで事前学習された 3D ResNet に基づく行動認識モデルに，自作した万引きデータセットで追加学習を行い，万引き行動の学習と検出精度評価を行った．データセットの作成は，YouTube から監視カメラ映像の入手，YOLO による人物領域の切り出し，「万引き」と「非・万引き」の人手でのラベル付けの手順で行った．

実験では RGB 画像とオプティカルフローの可視化画像の 2 種の比較を行い，汎化性能において後者が優れており，映像内の動的な情報を学習できていることは確認できたが，その精度はまだまだ満足できるものではない．

今後も，映像のみでの実用レベルの万引き検出を目指し，映像内の人の手元の動きを有効に学習できる手法の探索を行いたい．

参考文献

[1] Du Tran, et al., “Learning Spatiotemporal Features with 3D Convolutional Net-works,” ICCV, pp.4489–4497, 2015.
[2] earth eyes, http://earth-eyes.co.jp/, 2018 年 12 月 22 日参照.
[3] 原 健翔，他，“3D Residual Network による行動認識のための時空間特徴の学習，” SSII，2018.
[4] Joseph Redmon, et al., “YOLOv3: An Incremental Improvement,” arXiv:1804.02767，2018.
[5] Kaiming He, et al., “Deep Residual Learning for Image Recognition,” CVPR, pp.770-778, 2016.
[6] Will Kay, et al., “The Kinetics Human Action Video Dataset,” arXiv:1705.06950, 2017.
[7] Gunnar Farneback, “Two-Frame Motion Estimation Based on Polynomial Expansion,” SCIA 2003: Image Analysis, pp.363-370, 2003.