H-011

Two-Stream 3D Convolutional Neural Network（TS-3DCNN）による 万引き行動の自動検知
Automated Detection of Shoplifting with Two-Stream 3D Convolutional Neural Network (TS-3DCNN)

山下 裕之介† 檜作 彰良† 中山 良平†
Yunosuke Yamashita Akiyoshi Hizukuri Ryohei Nakayama

1. はじめに

我が国では，万引き被害が甚大であり，年間被害額は 4,615 億円に上る[1]．これに対処するため，多くの防犯カメラを設置した店舗もあるが，目視による監視は多大な時間と労力を要する．そこで，防犯カメラ映像から万引き行動を自動検知するシステムが切望されている．

行動認識分野では，これまでに深層学習を用いた手法が多数報告されている[2-4]．Karen らは，単一フレーム画像を解析する RGB-Stream Convolutional Neural Network（CNN）とオプティカルフローを解析する Optical Flow-Stream CNN の各出力を統合的に解析する Two-Stream ConvNets を提案し，少ない学習データでも高い認識精度を達成できることを示した[2]．また最近では，複数フレーム画像から動きを解析する 3DCNN（Three-Dimensional CNN）により，認識精度が改善できることも報告されている[3,4]．

そこで本研究では，Two-Stream ConvNets の CNN を時間軸方向へ拡張した Two-Stream 3DCNN（TS-3DCNN）を構築し，防犯カメラ映像における万引き行動の自動検知法を提案する．

2. 実験試料

実験試料として，UCF Anomaly Detection Dataset[5]及び動画共有サービスから，万引き行動が含まれる防犯カメラ映像を収集した．本研究では，万引き行動を「商品を鞄や服に隠す」動作とした．収集した映像から，万引き行動が含まれるシーン，含まれないシーンを抽出し，それぞれ異常映像，正常映像と定義した．そして，これらの映像から手動で人物を含む関心領域（ROI: Region of Interest）を切り出し，異常 ROI 映像（76 件）と正常 ROI 映像（76 件）を作成した．各 ROI 映像は，線形補間法により，画像サイズを 224×224 画素へリサイズした．最後に，各 ROI 映像を時間軸方向に 16 のセグメントに分割し，各セグメントから先頭の 1 フレームを抽出した 16 フレーム画像（224×224×16）を入力データとして用いた．

3. 方法

3.1 TS-3DCNN のネットワーク構造

図 1 に，TS-3DCNN のネットワーク構造を示す．TS-3DCNN の 2 入力のうち，1 入力は ROI 映像（224×224×16）とした．もう 1 つの入力として，ROI 映像のフレーム間差分映像（224×224×16）とフレーム間オプティカルフロー映像（224×224×16）を検討した．Two-stream モデルの入力としてフレーム間の違いに着目した画像を与えることにより，物体の消失および人物の動きに，より注目した学習を期待する．

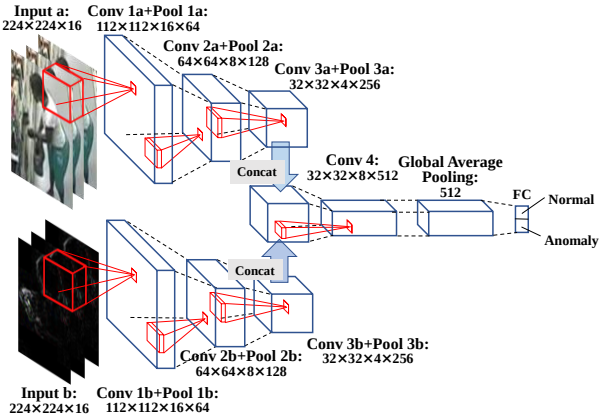

図 1　TS-3DCNN のネットワーク構造

TS-3DCNN は，ROI 映像とフレーム間差分（またはオプティカルフロー）映像を 3 層の畳み込み層と Max-Pooling 層で独立して処理し，それらの出力をマージ層において結合（Concat）する．そして，結合データを 1 層の畳み込み層で処理後，Global Average Pooling 層を経て，異常または正常に分類する．各畳み込み層の直後には，Batch Normalization 層と ReLU（Rectified linear unit）関数を用いる．ただし，出力層（FC）のみ softmax 関数とする．また，過学習を抑えるために，最終畳み込み層（Conv 4）と Global Average Pooling 層では，ドロップアウトを用いる．ここで，最終畳み込み層のドロップアウト率は 0.25，Global Average Pooling 層は 0.5 に設定する．

3.2 Kinetics-400 を用いた事前学習

3DCNN は，非常に多くのパラメータを有するため，それらのパラメータを最適化するためには，膨大な学習用データが必要である．しかし，実験試料の映像データは 152 件と少なく，パラメータの最適化に十分ではない．学習用データ数が十分でない場合に，CNN のパラメータを効果的に初期化する方法として知られるのが事前学習である[2]．そこで，人物の 400 種の動作に関する大規模データセット Kinetics-400（The Kinetics Human Action Video Dataset）[6]を用いて，TS-3DCNN の事前学習を行った．

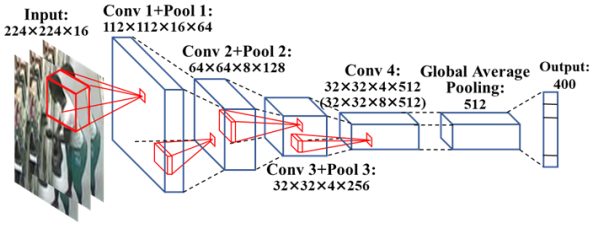

図 2　事前学習時のネットワーク構造

† 立命館大学大学院 理工学研究科, Graduate School of Science and Engineering, Ritsumeikan University

ただし本研究では，事前学習後，Cross Modality Pre-Training [3]に基づく TS-3DCNN のファインチューニングを行うため，ここでは，TS-3DCNN のフレーム間差分（またはオプティカルフロー）ROI 映像の解析に関する Stream を取り除き，出力層のノード数を 400 とした One-Stream の 3DCNN モデルを用いて事前学習を行った（図 2 参照）．

3.3 TS-3DCNN のファインチューニング

Kinetics-400 を用いた事前学習後，万引き映像データを用いた TS-3DCNN のファインチューニングを行った．ここでは，Cross Modality Pre-Training を用いて，事前学習した One-Stream 3DCNN モデルの畳み込み層（Conv 1, Conv 2, Conv 3）と Batch Normalization 層の重みを，TS-3DCNN の畳み込み層（Conv 1a, Conv 2a/2b, Conv 3a/3b）と Batch Normalization 層のパラメータの初期値として用いた．Conv 1b の初期パラメータは，Conv 1a のフィルタ 64 枚の重み平均により生成した 1 枚のフィルタを Conv 1b の 64 枚のフィルタすべてに与えた．また，Conv 1b 直後の Batch Normalization 層以外の Batch Normalization 層の平均と分散は学習時に更新されないよう凍結した．

TS-3DCNN のファインチューニング時の最適化アルゴリズムには SGD（Stochastic gradient descent），損失関数は cross entropy を用いた．出力層と最終畳み込み層の初期学習率を 0.001，それ以外の層を 0.0005 とした．また，モーメンタムは 0.9，減衰率は 0.0005，ミニバッチサイズは 3，エポック数は 20 と設定し，エポック数 8, 12, 15 で学習率を 1/10 に更新した．

3.4 評価方法

k-分割交差検証法を用いて，TS-3DCNN の学習および評価を行った．この検証法では，まず，152 ROI 映像を *k* 個のサブセットに分割し，一つのサブセットを評価用，残りの *k*-1 サブセットを学習用とする．そして，全てのサブセットが評価に用いられるまで学習と評価を繰り返す．本研究では，*k* の値を 4 とした．

また，万引き行動の検知精度の評価指標として，ROC（Receiver Operating Characteristics）分析に基づく ROC 曲線下の面積（AUC：Area Under the ROC Curve）を用いた．

4. 結果と考察

図 3 に，ROI 映像とフレーム間差分映像を入力した TS-3DCNN，ROI 映像とフレーム間オプティカルフロー映像を入力した TS-3DCNN，ROI 映像のみ入力した One-Stream 3DCNN および行動認識で良く用いられる ECO Lite（Efficient Convolutional Network for Online Video Understanding）[4]の ROC 曲線と AUC の結果を示す．ROI 映像とフレーム間差分映像を用いた TS-3DCNN の AUC（0.882）が最も高い結果となり，TS-3DCNN の有用性が示唆された．また，万引き行動は大きな動作を伴わないため，フレーム間の小さな変化を抽出できるフレーム間差分映像の方が，オプティカルフロー映像より適したと考える．

Cross Modality Pre-Training の有用性を評価するために，図 4 に事前学習なし，事前学習後の Cross Modality Pre-Training なし／ありの TS-3DCNN の ROC 曲線と AUC の結果を示す．ここで，TS-3DCNN の入力は，ROI 映像とフレーム間差分映像を用いた．ROI 映像のみで事前学習した One-Stream 3DCNN モデルのパラメータを TS-3DCNN のフレーム間差分映像の解析に関する Stream に転用すると認識精度が低下する結果となった．しかし，Cross Modality Pre-Training を用いることにより，事前学習なしの TS-3DCNN より認識精度が改善したことから，Cross Modality Pre-Training の有用性が示された．

5. おわりに

本研究では，TS-3DCNN を用いて防犯カメラ映像から万引き行動の自動検知を行った．提案手法の万引き行動の検知精度は，行動認識で良く用いられる ECO Lite よりも高く，その有用性が示唆された．

参考文献

[1] 全国万引犯罪防止機構 https://www.manboukikou.jp/ html/media.html
[2] K. Simonyan, A. Zisserman: “Two-Stream Convolutional Networks for Action Recognition in Videos”, Advances in Neural Information Processing System 27(NIPS), 2014
[3] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L.V. Gool: “Temporal Segment Networks for Action Recognition in Videos”, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Vol.41, No.11, pp. 2740-2775, 2018
[4] M. Zolfaghari, K. Singh, T. Brox: “ECO: Efficient Convolutional Network for online Video Understanding”, European Conference on Computer Vision (ECCV), pp. 695-712, 2018
[5] W. Sultani, C. Chen, M. Shah: “Real-world Anomaly Detection in Surveillance Videos”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.6479-6488, 2018
[6] W. Kay, J. Carreina, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman: “The Kinetics Human Action Video Dataset”, arXiv: 1705.06950, 2017

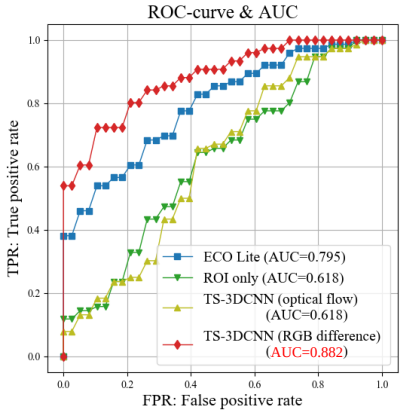

図 3　各手法における ROC 曲線の比較と AUC

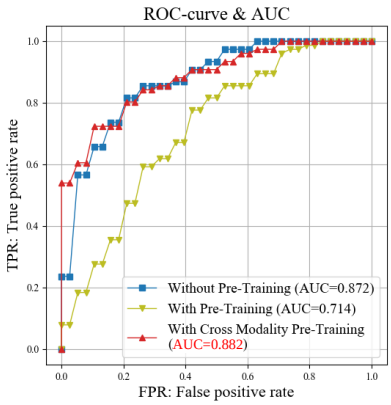

図 4　事前学習の有無による TS-3DCNN の精度比較