情報処理学会論文誌 Vol.62 No.5 1193–1199 (May 2021)

Cross Modality Pre-Trainingを用いたTwo-Stream 3D Convolutional Neural Networksによる万引き行動の自動検知

山下 裕之介[1,a)] 檜作 彰良[1] 中山 良平[1]

受付日 2020年8月14日, 採録日 2021年2月2日

概要：わが国における万引き被害額は年間4,615億円にのぼり，その対策が喫緊の課題となっている．本研究では，Cross Modality Pre-Training を用いた TS-3DCNNs（Two-Stream 3D Convolutional Neural Networks）により，防犯カメラ映像から万引き行動を自動検知する手法を提案する．実験試料は，防犯カメラで撮影された万引き行動が含まれる異常映像76件，含まれない正常映像76件で構成された．提案手法では，まず，各映像とそのフレーム間差分映像を TS-3DCNNs の2つの入力層に入力し，3層の3次元畳み込み層とプーリング層により，2入力映像から特徴マップをそれぞれ抽出した．そして，それらの特徴マップを統合し，1層の畳み込み層と Global Average Pooling 層を経て，入力映像を異常/正常に分類した．TS-3DCNNs の学習では，まず，行動認識のデータセット（Kinetics-400）で Cross Modality Pre-Training による事前学習を行い，本実験試料で再学習した．提案手法の ROC 曲線下の面積（AUC）は0.918で，従来手法の Efficient Convolutional Network for Online Video Understanding（0.795）より高く，その有用性が示唆された．

キーワード：two-stream 3D convolutional neural networks，cross modality pre-training，万引き

Automated Detection Scheme of Shoplifting with Two-stream 3D Convolutional Neural Networks Based on Cross Modality Pre-training

Yunosuke Yamashita[1,a)] Akiyoshi Hizukuri[1] Ryohei Nakayama[1]

Received: August 14, 2020, Accepted: February 2, 2021

Abstract: The total financial damage of shoplifting in Japan becomes over 461.5 billion yen per year. The purpose of this study was to develop two-stream 3D convolutional neural networks (TS-3DCNNs) for automatically detecting shoplifting behavior in security camera videos. Our database consisted of 76 abnormal videos with shoplifting behavior and 76 normal videos without shoplifting behavior. Original video and the difference video between its frames were inputted to two input layers in TS-3DCNNs, respectively. The feature maps were extracted independently from each of two videos through three sets of 3D convolutional layer and pooling layer. Those feature maps were merged and then processed sequentially in a 3D convolutional layer, a global average pooling layer and a fully connected layer. The fully connected layer classified the input videos into abnormal or normal video. In the training of TS-3DCNNs, TS-3DCNNs was pre-trained using a behavior recognition dataset (Kinetics-400) based on cross modality pre-training and then was re-trained using our dataset. The area under the ROC curve with TS-3DCNNs was 0.918, showing substantially greater than that with the conventional method for behavior recognition, efficient convolutional network for online video understanding (0.795). The proposed TS-3DCNNs achieved high classification performance and would be useful in detecting shoplifting behavior in security camera videos.

Keywords: two-stream 3D convolutional neural networks, cross modality pre-training, shoplifting

1. はじめに

わが国における万引き被害は深刻であり，2017年度に検挙された窃盗事件204,296件のうち，万引きは75,257件

[1] 立命館大学大学院理工学研究科
Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525–8577, Japan
[a)] ri0075ii@ed.ritsumei.ac.jp

（36.8%）と高い割合を占める [1]．また，2010 年度の年間万引き被害額は 4,615 億円で，1 日あたり 12.6 億円にのぼったとの報告もある [2]．万引き被害を削減するために，多くの店舗では防犯カメラを設置しているが，目視による監視は多大な時間と労力を要する．そこで，コンピュータを用いて，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要望されている．

近年，人物の行動認識分野において，深層学習（Deep Learning）を用いた手法が多数報告されている [3], [4], [5], [6], [7], [8], [9], [10]．特に映像の時間軸方向の特徴抽出が議論となっており，人物の特徴点を追跡しベクトル化することでフレーム間の微小な変化を識別する Trajectories に基づく行動認識手法 [3]，自然言語処理で要素/単語間の関係性を学習する Attention 機構を応用した手法 [4] など様々な手法が提案されている．また，ECO Lite（Efficient Convolutional Network for Online Video Understanding）は，映像データの各フレーム画像から 2 次元 CNN（2D-CNN：Two-Dimensional Convolutional Neural Network）で特徴マップを抽出し，それらの特徴マップを 3 次元 CNN（3D-CNN：Three-Dimensional CNN）で解析することにより，高い認識精度を達成している [5]．

単一フレーム（RGB）画像だけでなく，フレーム間差分画像やオプティカルフロー画像を追加入力する複数入力 CNN モデルの有用性を示した研究もある [6], [7], [8]．空間情報として単一フレーム画像を解析する RGB-Stream CNN，時間情報としてオプティカルフロー画像を解析する Optical Flow-Stream CNN の各出力を統合的に解析する Two-Stream CNNs により，少ない学習データでも高い認識精度を達成できることが報告されている [6]．また，2 次元 CNN を，時間軸方向へ拡張した 3 次元 CNN により，時間軸方向の特徴を効果的に抽出でき，認識精度が改善されることを示した研究もある [5], [9]．万引き行動の検知においては，時間経過にともなう動作の遷移が重要な要素となるため，時間軸方向の特徴抽出に優れた 3 次元 CNN と，RGB 映像とフレーム間差分映像（またはオプティカルフロー映像）を入力できる機構を組み合わせた Two Stream 3DCNNs（TS-3DCNNs）を構築することにより，防犯カメラ映像から万引き行動をより高精度に検知できる可能性がある．

しかし，TS-3DCNNs は非常に多くのパラメータを有しており，それらのパラメータを最適化するためには，膨大な学習データが必要である．学習データが十分でない場合，実験対象とは異なるデータセットを用いて，ネットワークのパラメータを事前に調整する事前学習がしばしば用いられる [6]．事前学習で調整されたパラメータを実験対象の学習データで再調整することにより，ネットワークの効率的な学習が可能となる．しかし，複数入力モデルである TS-3DCNNs の事前学習は，多くの時間を要する．また，TS-3DCNNs のフレーム間差分映像（オプティカルフロー映像）を解析する Stream CNN では，時間軸方向の特徴を感度良く抽出するため，実験対象と動きが異なる映像を事前学習に用いた場合，パラメータを適切に調整できないという問題がある．Cross Modality Pre-Training は，2 画像を入力とする Two-Stream CNNs において，RGB 画像を解析する Stream CNN だけを事前学習し，その調整されたパラメータから新たなパラメータを生成し，別の Stream CNN の初期パラメータとして与える事前学習法である [8]．Cross Modality Pre-Training を用いることにより，多くの学習時間を要する問題，実験対象と異なる動作の映像を用いた事前学習でパラメータを適切に調整できないという問題を解決できる可能性がある．

従来研究において，人物検出に関する報告は多数あり，高精度な人物検出法が提案されている [10]．そこで本研究では，防犯カメラ映像上の人物が映る関心領域に対して，Cross Modality Pre-Training を用いた TS-3DCNNs により，万引き行動を自動検知する手法を提案する．

2. 実験試料

実験試料として，UCF Anomaly Detection Dataset [11] および動画共有サービス（YouTube）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」と定義した．収集した映像から，万引き行動が含まれるシーンを異常映像，含まれないシーンを正常映像として，手動で抽出した．正常映像には「商品棚の前で商品を見ている」，「商品を手に取る」，「商品を持って歩いている」動作が含まれる．抽出した映像時間は，万引き行動の開始から終了までの 1〜5 秒間である．各映像は，フレーム間の線形補間法 [12] によりフレームレートを 30 fps に統一した．そして，これらの映像から人物を含む関心領域（ROI：Region of Interest）を手動で切り出し，異常 ROI 映像（76 件）と正常 ROI 映像（76 件）の 152 映像を作成した．各 ROI 映像は，線形補間法 [12] により，各フレーム画像のサイズを 224 × 224 画素へリサイズした．**図 1** に ROI の例を示す．30 fps に統一した映像データでは，フレーム間の動きが小さい．そこで，全フレーム画像を CNN モデルに入力するのではなく，間隔を空けて選択したフレーム画像を CNN モデルに入力す

図 1 防犯カメラ映像から人物を含む ROI の切り出し例
Fig. 1 ROI for people extracted from security camera video.

る [5]．まず，各 ROI 映像を時間軸方向に 16 のセグメントに分割する．そして，各セグメントから先頭の 1 フレーム画像を抽出した 16 フレーム画像（224 × 224 × 16）を入力データとして用いた．以下，この 16 フレーム画像を ROI 映像とする．

本稿で示すすべての防犯カメラ映像のフレーム画像は，UCF Anomaly Detection Dataset [11] からの引用である．

3. 方法

3.1 TS-3DCNNs のネットワーク構造

図 2 に，TS-3DCNNs のネットワーク構造を示す．TS-3DCNNs 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）とオプティカルフロー映像（224 × 224 × 16）をそれぞれ検討した．**図 3** にフレーム間差分映像およびオプティカルフロー映像の例を示す．フレーム間差分映像は，ROI 映像の前後フレーム画像を差分することにより生成した．また，オプティカルフロー映像は，Farneback 法 [13] により ROI 映像中の全画素に対してオプティカルフローを推定し，HSV 空間上でベクトル表現したものを RGB にカラーコード化することにより生成した [14]．

TS-3DCNNs は，ROI 映像とフレーム間差分映像（またはオプティカルフロー映像）を 3 層の 3 次元畳み込み層と Max-Pooling 層で独立して処理し，それらの出力である特徴マップをマージ層で結合（Concat）した．そして，結合した特徴マップを 1 層の 3 次元畳み込み層で処理後，Global Average Pooling 層を経て，異常映像または正常映像に分類した．各畳み込み層では，フィルタサイズを 3 × 3 × 3，ストライドを 1 と設定し，畳み込み層の入力サイズと出力サイズが同じになるようにパディングを行った．図 2 の Pool 1a，Pool 1b の Max-Pooling 層はウインドウサイズを 2 × 2 × 1，それ以外の Max-Pooling 層は 2 × 2 × 2 とした．各畳み込み層の直後には Batch Normalization 層と ReLU（Rectified linear unit）関数を用い，出力層（FC）では活性化関数として softmax 関数を与えた．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，ドロップアウト率 0.5 でドロップアウトを用いた．

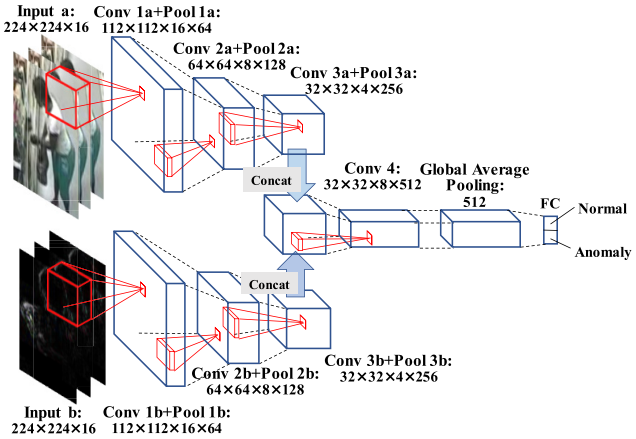

図 2 TS-3DCNNs のネットワーク構造
Fig. 2 Structure of TS-3DCNNs.

3.2 Kinetics-400 を用いた TS-3DCNNs の事前学習

実験試料の映像データは 152 件と少なく，TS-3DCNNs のパラメータの最適化に十分ではない．そこで本研究では，人物の 400 種類の動作映像の大規模データセットである Kinetics-400（The Kinetics Human Action Video Dataset）[15] を用いて，事前学習を行った．ここでは，Cross Modality Pre-Training を実施するため，TS-3DCNNs からフレーム間差分映像（またはオプティカルフロー映像）を解析する Stream CNN を取り除き，出力層のノード数を 400 とした One-Stream の 3DCNN（OS-3DCNN）モデルを用いて事前学習を行った（**図 4**）．また，Kinetics-400 データセットを学習データ（246,245 件）と検証データ（20,000 件）に分割し，OS-3DCNN の学習に用いた．

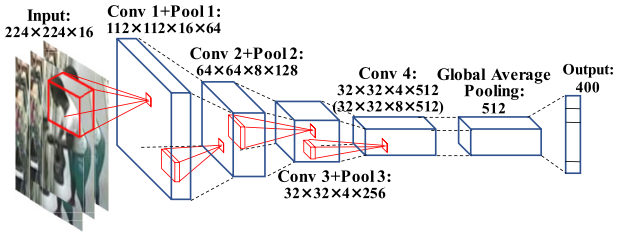

図 4 OS-3DCNN のネットワーク構造
Fig. 4 Structure of OS-3DCNN.

図 3 フレーム間差分映像とオプティカルフロー映像の例
Fig. 3 Difference video between frames and optical flow video.

OS-3DCNN の事前学習時の最適化アルゴリズムは，SGD（Stochastic gradient descent）を使用し，損失関数にはクロスエントロピー誤差を用いた．また，各学習パラメータとして，初期学習率：0.001，モーメンタム：0.9，減衰率：0.0005，ミニバッチサイズ：5，エポック数：15 を設定し，検証データに対する損失が 4 エポック間減少しないたびに学習率を 1/10 倍に更新した．

3.3 TS-3DCNNs のファインチューニング

本研究では，学習データ数が少ないため，Cross Modality Pre-Training に基づき，TS-3DCNNs の各 Stream CNN の初期パラメータを与えた．事前学習した RGB 映像を解析する OS-3DCNN の畳み込み層（Conv 1，Conv 2，Conv 3）と Batch Normalization 層のパラメータを，TS-3DCNNs の畳み込み層（Conv 1a，Conv 2a/2b，Conv 3a/3b）と Batch Normalization 層の初期パラメータとして与えた．TS-3DCNNs の Conv 1b の初期パラメータは，OS-3DCNN の Conv 1 が有するフィルタ 64 枚のフィルタ係数を各位置で平均し，1 枚のフィルタを生成して，Conv 1b の 64 枚のフィルタすべてに与えた．また，Conv 1b 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．出力層（FC）と最終畳み込み層（Conv 4）の初期パラメータは乱数で与えた．ここで，事前学習した OS-3DCNN のパラメータを TS-3DCNNs のフレーム間差分映像（またはオプティカルフロー映像）の Stream CNN に，そのまま転用していないことに注意されたい．そして，Cross Modality Pre-Training で初期パラメータを与えた TS-3DCNNs を実験試料である万引き映像データを用いてファインチューニングした．

TS-3DCNNs 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，損失関数はクロスエントロピー誤差を用いた．出力層（FC）と最終畳み込み層（Conv 4）の初期学習率は 0.01，それ以外の層は 0.005 とした．また，モーメンタム：0.9，減衰率：0.0005，ミニバッチサイズ：3，エポック数：20 と設定し，エポック数 8，12，15 で学習率を 1/10 倍に更新した．

3.4 評価方法

k–分割交差検証法 [16] に基づき，TS-3DCNNs の学習および評価を行った．本研究では，k の値を 4 と設定した．まず，152 ROI 映像を 4 つのサブセットに分割し，1 つのサブセットを評価用，残りの 3 サブセットを学習用とした．そして，すべ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し実施した．

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）[17] を用いた．

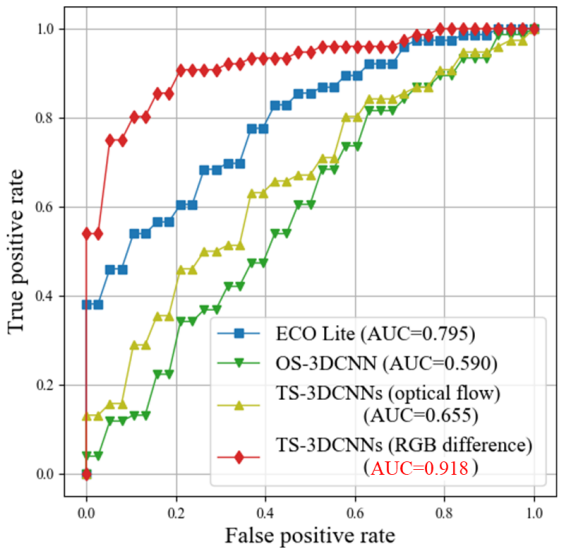

図 5 異なる CNN モデルによる ROC 曲線の比較
Fig. 5 Comparison of ROC curves for four CNN models.

4. 結果と考察

図 5 に，行動認識で頻繁に使用される ECO Lite [5]，ROI 映像のみ入力した OS-3DCNN，ROI 映像とオプティカルフロー映像を入力した TS-3DCNNs（TS-3DCNNs with RGB and Optical flow），および ROI 映像とフレーム間差分映像を入力した TS-3DCNNs（TS-3DCNNs with RGB and RGB Difference）による ROC 曲線と AUC の結果を示す．TS-3DCNNs with RGB and RGB Difference の AUC（0.918）は，ECO Lite（0.795），OS-3DCNN（0.590），TS-3DCNNs with RGB and Optical flow（0.655）より高い結果となった．行動認識に関する従来研究では，動きを補足するための情報としてオプティカルフロー映像を追加することの有用性が報告されている [6], [7]．本研究では，フレーム間差分映像を補足情報として追加した場合と比べ，オプティカルフロー映像の追加による AUC の改善は小さかった．本研究で対象とする万引き行動は大きな動作をともなわないうえに，防犯カメラ映像は解像度が低いため，オプティカルフローによる解析は困難であったと考える．一方，フレーム間差分処理は，解像度に関係なく，フレーム間の小さな動きも抽出できるため，万引き行動の検知に適すると考える．また，ROI 映像のみを入力する単一入力モデルである ECO Lite，OS-3DCNN より，フレーム間差分映像を追加入力した TS-3DCNNs の AUC が高かった．この結果は，ROI 映像に，万引き行動の微小な動き情報を補足できるフレーム間差分映像を追加し，それらを統合的に解析することの有用性を示すと考える．

Cross Modality Pre-Training の有用性を評価するために，事前学習をせずに初期パラメータを乱数で与えた TS-3DCNNs，事前学習された OS-3DCNN のパラメータを ROI 映像 Stream CNN の初期パラメータとして与えた

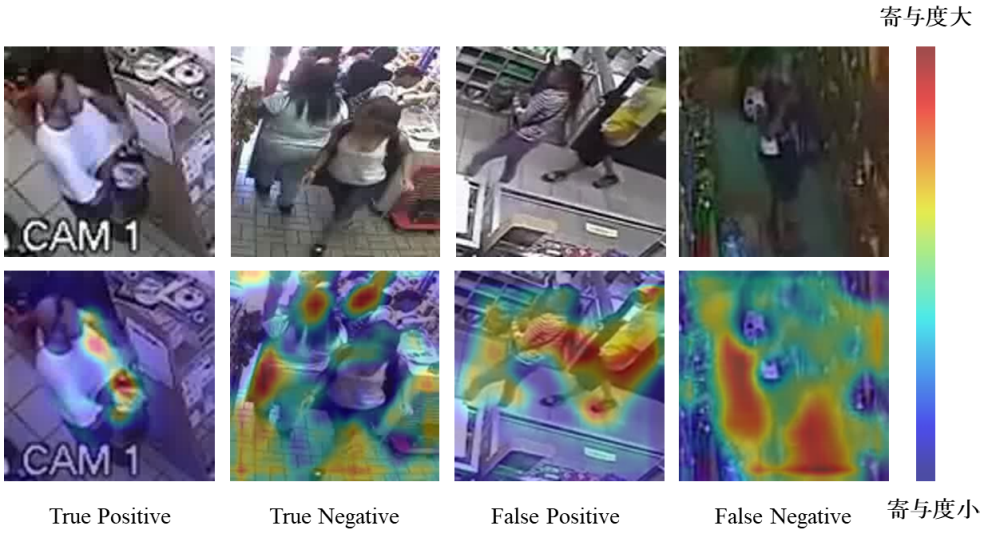

図 7 万引き行動検知の成功例と失敗例

Fig. 7 Example of true positive, true negative, false positive, and false negative.

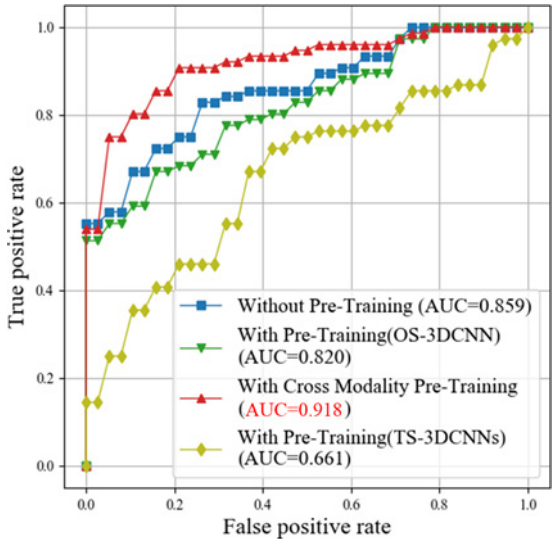

図 6 異なる事前学習法に基づく ROC 曲線の比較

Fig. 6 Comparison of ROC curves with different pre-training methods.

TS-3DCNNs，事前学習された OS-3DCNN のパラメータを Cross Modality Pre-Training に基づき初期パラメータとして与えた TS-3DCNNs，Two-Stream で事前学習した TS-3DCNNs による万引き検知の比較実験を行った．いずれの TS-3DCNNs も ROI 映像とフレーム間差分映像を入力とし，実験試料による再学習を実施した．**図 6** に異なる事前学習法に基づく ROC 曲線の比較を示す．Cross Modality Pre-Training を用いた TS-3DCNNs の AUC（0.918）は，事前学習なし TS-3DCNNs（0.859），One-Stream のみ事前学習を行った TS-3DCNNs（0.820），Two-Stream で事前学習を行った TS-3DCNNs（0.661）より高い結果となった．Two-Stream で事前学習を行った場合に AUC が最も高くなることを期待したが，最も低い結果となった．TS-3DCNNs のフレーム間差分映像 Stream CNN では動きに着目した解析を行うが，事前学習で用いた Kinetics-400 には，「スキー」「ボールを蹴る」など，万引き行動とは大きくかけ離れた動作も数多く含まれる．したがって，フレーム間差分映像 Stream CNN が抽出する特徴マップの分布が万引き動作とは大きく異なる事前学習が行われ，万引き動作の検知精度を低下させた可能性がある．また，One-Stream のみ事前学習を行った TS-3DCNNs では，ROI 映像 Stream CNN に事前学習で調整した初期パラメータを，フレーム間差分映像 Stream CNN には乱数を与えた．これにより再学習時に各 Stream CNN の最適化の進捗が不均衡となり，両 Stream CNNs の初期パラメータとして乱数を与えた事前学習なし TS-3DCNNs の AUC より低い結果になったと考える．Cross Modality Pre-Training では，One-Stream CNN の事前学習に基づき，両 Stream CNNs に調整した初期パラメータを与えることができるため，最適化の不均衡を発生させることなく，効率的な学習ができたと考える．

また，OS-3DCNN の事前学習時間は 134 時間 27 分であったのに対し，Two-Stream で事前学習した TS-3DCNNs は 437 時間 8 分であった．この結果から，Cross Modality Pre-Training が，分類精度の向上だけでなく，学習時間の短縮にも有用であることが確認できた．

TS-3DCNNs が本質的な判断根拠に基づいて万引き行動を検知しているか評価するため，2D-CNN の判断根拠の可視化手法である Grad-CAM（Gradient-weighted Class Activation Mapping）[18] を 3 次元に拡張し [19]，映像中で TS-3DCNNs の分類に寄与する領域を可視化した．可視化画像が赤いほど TS-3DCNNs の分類への寄与度が高いことを示す．**図 7** に，TS-3DCNNs with RGB and RGB Difference による万引き行動検知の成功例と失敗例およびその判断根拠の可視化画像を示す．正しく異常映像と判断した例（True Positive）では，商品を服に隠す手の動きに集中して着目していることが確認できた．正しく正常映像

と判断した例（True Negative）では，人物の足や頭の動きなど広範囲を着目し，万引き行動ではないと判断された．また，誤って異常映像と判断した例（False Positive）ではROI に複数人物が映っている映像が散見され，複数の人物の動作を広範囲に着目していた．誤って正常映像と判断した例（False Negative）では手と背景のコントラストが低い映像が含まれ，「商品を鞄や服に隠す」動作に着目していなかった．

今後の課題として，以下があげられる．

(1) 防犯カメラ映像のコントラスト改善

実験結果より，手と背景のコントラストが低い防犯カメラ映像において，提案手法は手の動きを正確に把握することができなかったと考えられる．今後，防犯カメラ映像のコントラスト改善手法を検討する必要がある．

(2) 人物検出手法の導入

提案手法を防犯カメラに導入するためには，個々の人物を前処理として検出する必要がある．Liu らは，映像から個々の人物を含む領域を高精度に抽出する手法を提案している [10]．今後，Liu らの手法と提案手法を組み合わせることにより，防犯カメラ映像上の人物検出から万引き行動の検知までを一貫して行うネットワークに拡張する．個々の人物を抽出することは，複数人物が映った映像において，個々人の動きのみに着目した解析を可能にし，検知精度の向上も期待できる．

5. おわりに

本研究では，Cross Modality Pre-Training を用いた TS-3DCNNs により，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，従来手法である ECO Lite よりも高く，その有用性が示唆された．また，Cross Modality Pre-Training が事前学習時間の削減と検知精度の向上に寄与することも明らかとなった．提案手法が検知に失敗した例として，複数人物が映っている映像，コントラストが低い映像があげられ，これらへの対策が，今後の課題である．

参考文献

[1] 警察庁：平成 30 年版警察白書　統計資料，入手先 ⟨https://www.npa.go.jp/hakusyo/h30/data.html⟩（参照 2020-07-25）．

[2] 全国万引犯罪防止機構，入手先 ⟨https://www.manboukikou.jp/html/media.html⟩（参照 2020-01-20）．

[3] Wang, H. and Schmid, C.: Action Recognition with Improved Trajectories, *The IEEE International Conference on Computer Vision* (*ICCV*), pp.3551–3558 (2013).

[4] Wang, X., Girshick, R., Gupta, A., et al.: Non-local neural networks, *The IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.7794–7803 (2018).

[5] Zolfaghari, M., Singh, K. and Brox, T.: ECO: Efficient Convolutional Network for Online Video Understanding, *European Conference on Computer Vision* (*ECCV*), pp.695–712 (2018).

[6] Simonyan, K. and Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos, *Advances in Neural Information Processing System 27* (*NIPS*) (2014).

[7] Wang, L., Xiong, Y., Wang, Z., et al.: Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, *European Conference on Computer Vision* (*ECCV*), pp.20–36 (2016).

[8] Wang, L., Xiong, Y., Wang, Z., et al.: Temporal Segment Networks for Action Recognition in Videos, *IEEE Trans. Pattern Analysis and Machine Intelligence* (*TPAMI*), Vol.41, No.11, pp.2740–2775 (2018).

[9] Tran, D., Bourdev, L., Fergus, R., et al.: Learning Spatiotemporal Features with 3D Convolutional Networks, *The IEEE International Conference on Computer Vision* (*ICCV*), pp.4489–4497 (2015).

[10] Liu, W., Anguelov, D., Erhan, D., et al.: SSD: Single Shot Multibox Detector, *European Conference on Computer Vision* (*ECCV*), pp.21–37 (2016).

[11] Sultani, W., Chen, C. and Shah, M.: Real-world Anomaly Detection in Surveillance Videos, *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.6479–6488 (2018).

[12] Meijering, E.: A chronology of interpolation: From ancient astronomy to modern signal and image processing, *Proc. IEEE*, pp.319–342 (2002).

[13] Farneback, G.: Two-Frame Motion Estimation Based on Polynomial Expansion, *Scandinavian Conference on Image Analysis* (*SCIA*), pp.363–370 (2003).

[14] Horn, B.K. and Schunck, B.G.: Determining optical flow, *Techniques and Applications of Image Understanding*, Vol.281, pp.319–331 (1981).

[15] Kay, W., Carreina, J., Simonyan, K., et al.: The Kinetics Human Action Video Dataset, arXiv: 1705.06950 (2017).

[16] Hizukuri, A. and Nakayama, R.: Computer-aided diagnosis scheme for determining histological classification of breast lesions on ultrasonographic images using convolutional neural network, *Diagnostics*, Vol.8, No.3, pp.1–8 (2018).

[17] Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern recognition*, pp.1145–1159 (1997).

[18] Selvaraju, R.R.M., Cogswell, M., Das, A., et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, *The IEEE International Conference on Computer Vision* (*ICCV*), pp.618–626 (2017).

[19] 山下裕之介，檜作彰良，中山良平：3 次元畳み込みニューラルネットワークを用いた万引き行動の自動検知と判断根拠の可視化，DIA2020 動的画像処理実利用化ワークショップ，pp.359–362 (2020).

山下 裕之介

2020 年立命館大学理工学部電子情報工学科卒業．同大学大学院理工学研究科電子システム専攻博士前期課程在学中．

檜作 彰良

2014 年三重大学大学院工学研究科博士後期課程修了，博士（工学）取得．2014 年みずほ情報総研情報通信研究部入社．2018 年みずほ情報総研情報通信研究部退社．2018 年立命館大学理工学部電子情報工学科助教，現在に至る．電子情報通信学会，電気学会，医学物理学会，日本医用画像工学会各会員．

中山 良平 （正会員）

1999 年宮崎大学工学部情報工学科卒業．2001 年同大学大学院工学研究科修士課程修了．2005 年三重大学大学院医学系研究科博士課程修了．同年三重大学医学部附属病院助教，2015 年立命館大学理工学部准教授，2020 年同教授．この間，2008 年シカゴ大学放射線科 visiting assistant professor．主に医用画像を対象とした画像認識，機械学習に関する研究に従事．博士（工学），博士（医学）．IEEE，SPIE，電子情報通信学会，日本医用画像工学会，日本画像情報学会，日本放射線技術学会各会員．