顕著な動画特徴の抽出による現実世界での異常検出

渡邉 祐大[1,a)]　岡部 誠[2,b)]　原田 泰典[3,c)]　鹿島 直二[3,d)]

概要

動画から異常を検出するための軽量で高精度な学習手法を提案する．近年成功している既存手法を分析したところ，動画全体から顕著な特徴を取り出すことが重要であることがわかった．そこで，我々は入力された全セグメント中から正常/異常の判定に重要な特徴量を自動抽出するための自己注意機構を導入する．その結果，既存手法の 1.3%のパラメータ数のニューラルネットワークでも，既存手法を上回るか同等の精度を達成できることが分かった．

1. はじめに

監視カメラの台数は世界中で年々増加を続けており，街中での防犯や工場や，発電所などの大規模施設での安全確認に利用されている．しかし，これらの映像全てを人間が目で見て確認することは難しいため，人間に代わって人工知能が映像を解析し，自動的に異常事象を検出できる技術の開発が急務である．

監視カメラ映像に基づく異常検出技術は長年研究されてきた．古くから研究されている手法の多くは異常が稀にしか起きないことを前提としている．こういった手法では，学習データとして正常なデータのみを与え，混合ガウスモデル [1]，スパースモデリング [2], [3]，オートエンコーダ（AE）[4] などを用いて正常なデータのパターンを学習し，推論時には，正常なデータのパターンとの乖離度を計算することで異常を検出する．こういった教師なし学習をもとにした手法は多く提案されているが，見た目や速度の違いなどの低レベルな特徴に基づいた異常検出しかできない．そこで近年，正常な動画と異常な動画の両方を含む弱い教師データセット [5], [6] を用いて異常検出器を学習する手法が提案されている [5], [6], [7], [8], [9], [10].

通常，フレーム単位の異常検出器を学習させる場合には全てのフレームに対してラベルを付け，学習させる必要があるため，ラベル付けのコストが高い．そこで Sultani ら [5] は動画単位で正常/異常のラベルが付いている弱い教師データセットを提案した．正常とラベル付けされた動画では全フレームを通して正常状態のみが記録されている．一方で，異常とラベル付けされた動画には正常状態と異常状態のフレームが混在している．このようなデータセットを用いることで，動画中のフレーム毎に正常/異常のラベル付けをする必要がなくなり，ラベル付けの労力を削減できる．また，Sultani ら [5] は，複数のフレームを 1 つの短期的なセグメントとして扱い，動画をセグメントのバッグと考え，そのバックの中からもっとも異常度を表すスコアが高いセグメントに対して学習することで，動画単位のラベル付けでも異常検出器を学習する多重インスタンス学習（MIL）をもとにした弱い教師ありデータセットのための異常検出手法を提案した．

しかし，近年成功している既存研究によれば，単一のセグメントだけに注目するのではなく，動画から得られる全セグメントをまとめて入力とし、セグメント間の時間的関係性を学習することが高精度を達成するために重要だと主張されている [7]．そこで我々は動画のセグメントを時間的にランダムに並び替えたデータセットを用い，これらの手法を分析した．すると，全セグメントをまとめて学習することは確かに重要だが，それらセグメント間の時間的関係性は高精度と無関係なことが分かった．

この知見に基づき，我々は MIL を用いず，代わりに入力の全セグメントから正常/異常の判定に重要な特徴量を自動抽出するための自己注意機構を導入した新たなモデルを提案する．その結果，既存手法 [7] の僅か 1.3%のパラメータ数のニューラルネットワークでも，既存手法 [7] を上回るか同等の精度を達成できることが分かった．ベンチマーク・データセット（UCF-Crime, ShanghaiTech）を用いたフレームレベルの検出精度を報告する．

2. 提案手法

我々は動画から異常を検出するための軽量で高精度な学習手法を提案する．提案手法は動画全体を解析し，正常/異常の判定に重要な特徴量を自動的に抽出して学習する．

[1] 静岡大学大学院総合科学技術研究科工学専攻　(現在，株式会社パナソニックシステムネットワークス開発研究所勤務)
[2] 静岡大学大学院総合科学技術研究科工学専攻
[3] 中部電力株式会社
[a)] watanabe.yudai.16@shizuoka.ac.jp
[b)] m.o@acm.org
[c)] Harada.Yasunori@chuden.co.jp
[d)] Naoji.Kashima@chuden.co.jp

データセットを $\mathcal{D} = \{(\mathbf{V}_i, y_i)\}$ とする．ただし，$\mathbf{V}_i$ は学習データセット中の i 番目の動画，y_i は $\mathbf{V}_i$ に付けられたラベルである．$y_i = \{0, 1\}$ で 0 ならば正常，1 ならば異常を表す．正常とラベル付けされた動画では全フレームを通して正常状態のみが記録されている．一方で，異常とラベル付けされた動画には正常状態と異常状態のフレームが混在している．$\mathbf{V}_i$ は T 個のセグメントに分割され，各セグメントは特徴抽出器によって D 次元の特徴ベクトル $\mathbf{F}_{i,j}$ に変換されている．ただし，$\mathbf{F}_{i,j}$ は $\mathbf{V}_i$ の j 番目の特徴ベクトルを表す．特徴抽出器は全ての実験を通して，Kineticts データセットによって学習済みの I3D[11] を用いた．

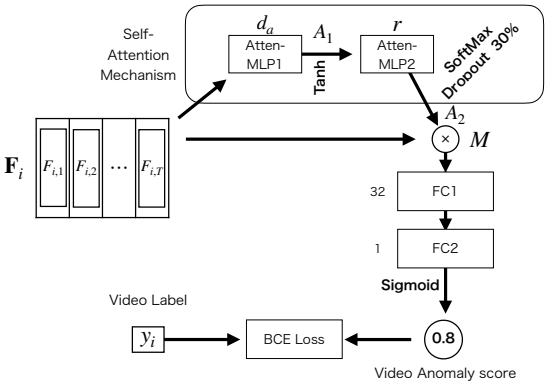

図 1 提案手法の概要図

提案手法は自己注意機構と 2 層の全結合層からなるシンプルなモデルである (図 1)．自己注意機構では，入力された $\mathbf{F}_i$ は Atten-MLP1 の多層パーセプトロンによって $d_a \times T$ の行列 $\mathbf{A}_1$ に変換される．この時，活性化には Tanh 関数を用いる．$\mathbf{A}_1$ は Atten-MLP2 の多層パーセプトロンによって $r \times T$ の行列 $\mathbf{A}_2$ に変換される．この時，活性化には softmax 関数を用いる．更に 30%のドロップアウト正則化も行う．

自己注意機構から得られた重み行列 $\mathbf{A}_2$ を用いて $\mathbf{M} = \mathbf{F}_i\mathbf{A}_2^t$ を計算する．$\mathbf{M}$ の形を $D \times r$ 次元のベクトルに変形した後，2 層の全結合層 FC1(32 ユニット) と FC2(1 ユニット）を経て動画の異常性を表すスコアに変換される．FC1 の活性化関数は恒等関数，FC2 の活性化関数は sigmoid 関数である．損失関数にはバイナリクロスエントロピー (BCE) 関数を用いた．

2.1 提案手法のモチベーション

いくつかの既存手法では動画内におけるセグメント間の時間的な関係性を考慮した学習が行われている [5], [7]. Sultani ら [5] は動画内で異常スコアは連続的に変動するはずであるとして，損失関数に異常スコアの連続性を課す項を導入した．また，Tian ら [7] は局所的な時間的特徴と大域的な時間的特徴を捉える目的で multi-scale temporal network (MTN) を導入した．これらの手法は共に高い異常検出精度を達成している．そこで我々は，セグメント間の時間的な関係性を捉えるための機構がどのように高精度に貢献しているのかを調査した．

具体的には，学習用の動画 $\mathbf{V_i}$ から得られた特徴ベクトルの集合 $\mathbf{F_i}$ では，デフォルトで特徴ベクトルは $\{\mathbf{F_{i,1}}, \mathbf{F_{i,2}}, \cdots, \mathbf{F_{i,T}}\}$ の順番に並んでいるが，この順番をランダムに並び替えたデータセットを作成し学習を行った．UCF-Crime データセット [5] を用いた実験結果を表 1 に示す．どちらの手法 [5], [7] でも特徴ベクトルのランダムな並び替えによる精度の低下は見られなかった．

Method	Reorder	AUC(%)
Sultani et al. [5]		81.39
	✓	81.54
RTFM [7]		84.30
	✓	84.26

表 1 UCF-Crime データセットにおいて，各動画の特徴ベクトルをランダムに並び替えたデータセットでの AUC 性能

この結果は，セグメント間の時間的関係性を捉えることが，異常検出器の精度に貢献しているわけではないことを示している．一方で，時間的な連続性を課す損失関数 [5] も，Tian らの手法で導入されている MTN や top-k 戦略 [7] も，より多くの特徴ベクトルを学習に関与させようと働きかける効果がある．MIL では学習に関与できるのは選ばれた少数の特徴ベクトルのみであるため，こういった工夫によって正常/異常の判定に重要な特徴量を抽出しようと試みていると思われる．我々は上記の観察から，これらの手法が高い異常検出精度を達成しているのは，動画から得られる全セグメントをまとめて学習の対象とし，それらの中から顕著な特徴量を効率良く抽出できる機構を持っているからではないかと考えた．

提案手法 (図 1) は以上の洞察に基づいてデザインされている．提案手法は MIL のフレームワークではなく，動画から得られる全セグメントを入力とし，その動画が正常か異常かを判定するモデルである．MIL を用いないので，顕著な特徴量の抽出はシンプルな自己注意機構で実現できる．自己注意機構には Lin ら [12] の手法に触発されたモデルを導入した．Lin ら [12] らの手法は文章分類を対象としており，可変長な入力に対応できる．我々は学習時は動画 $\mathbf{V}_i$ を T 個のセグメントに分割して学習するが、推論時の分割数は異なる値にしたいためこの機構を採用した．提案手法は時間的関係性を捉えるための機構を持たず，軽量かつ高精度な手法である (3 章).

2.2 動画分割による推論

提案手法を適用して異常検出したい動画を $\mathbf{V}^e$ とする．まず, 連続する 16 フレームを 1 つのセグメントとし, $\mathbf{V}^e$ を N 個のセグメント $\{\mathbf{V}_1^e, \mathbf{V}_2^e, \cdots, \mathbf{V}_N^e\}$ に分割する．各セグメント $\mathbf{V}_i^e$ は I3D[11] によって D 次元の特徴ベクトル $\mathbf{F}_i^e$ に変換される．特徴ベクトルの集合を $\mathbf{F}^e = \{\mathbf{F}_1^e, \mathbf{F}_2^e, \cdots, \mathbf{F}_N^e\}$

とする．次に分割サイズを l とし，$\mathbf{F}^e$ を l 個のセグメント毎に $m = N/l$ 個のバッグに分割する．即ち，

$$\mathbf{F}^e = \{\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_m\}$$
$$= \{\{\mathbf{F}^e_1, \cdots, \mathbf{F}^e_l\}, \{\mathbf{F}^e_{l+1}, \cdots, \mathbf{F}^e_{2l}\} \cdots \{\mathbf{F}^e_{N-l+1}, \cdots, \mathbf{F}^e_N\}\}$$

となる．次に，$\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_m$ を 1 つずつ異常検出器に入力とすることで推論結果 $\mathbf{S}^v = \{s^v_1, s^v_2, \cdots, s^v_m\}$ を得る．s^v_i は $\{\mathbf{V}^e_{(i-1)l+1}, \cdots, \mathbf{V}^e_{il}\}$ の各セグメントに対する異常スコアとして用いられる．フレームレベルの異常検出の結果を作る必要がある際は，得られたスコアを各セグメントの全フレームに割り当てることとする．

3. 実験

提案手法の評価を行うため，異常検出のための弱い教師データセット (UCF-Crime データセット [5] と ShanghaiTech データセット [3]) を用いて実験を行った．

3.1 データセットと評価指標

UCF-Crime データセットは，現実世界における監視カメラ映像を集めた大規模なデータセットである．このデータセットは 13 種類の異常を含んでいる．動画の本数は計 1900 本であり，動画の総時間は 128 時間である．1900 本の動画のうち，1610 本が学習用の動画で，290 本がテスト用の動画である．学習用の各動画には動画単位で正常/異常のラベルが付けられている．テスト用の各動画にはフレーム単位で正常/異常のラベルが付けられている．

ShanghaiTech データセットは，大学内に設置された 13 箇所の固定カメラで撮影された映像を集めた，中規模なデータセットである．このデータセットには，437 本の動画が含まれている．437 本の動画のうち，307 本が正常な動画で，130 本が異常を含んだ動画である．オリジナルのデータセットは教師なし学習による異常検出器の開発のためのデータセットであった．しかし，Zhong ら [8] は弱い教師あり学習のためのデータセットとして使えるように動画単位のラベル付けを行った．我々は Zhong ら [8] と同様の手順で弱い教師データセットを構築し，実験を行った．

評価指標．既存研究 [5], [6], [7], [8], [9], [10] と同様に，フレーム単位の異常検出精度に基づいて計算された ROC 曲線における AUC を評価指標とする．AUC は値が大きいほど高精度な異常検出器であることを表す．

3.2 実装の詳細

PyTorch を用いて開発し，評価実験を行った．最適化アルゴリズムには Radam を用い，学習率は 0.001 とした．バッチサイズは 64 とした．ただし，既存研究 [5] と同様に 1 つのミニバッチには，正常動画と異常動画が同数含まれるようにミニバッチを作る．ハイパーパラメータの T は 32 とした．また，既存研究 [7], [10] と同様に，各動画に対し 10-crop オーギュメンテーションを行った．

3.3 UCF-Crime データセットの実験結果

表 2 に UCF-Crime データセットにおけるフレーム単位での AUC 性能を示す．ただし，分割サイズ l は 28 とした．提案手法はシンプルなモデルながら，既存手法の中で精度が最も高い RTFM[7] に比べて 0.61%高い検出精度を達成している．

Method	Feature Type	AUC(%)
Sultani et al. [5]	C3D RGB	75.41
GCN-Anomaly [8]	TSN RGB	82.12
MIST [10]	I3D RGB(Fine)	82.30
CLAWS Net [9]	C3D RGB	83.03
Wu et al. [6]	I3D RGB	82.44
RTFM [7]	I3D RGB	84.30
Ours ($d_a = 64, r = 3$)	I3D RGB	84.74
Ours ($d_a = 128, r = 7$)		84.91

表 2 UCF-Crime データセットにおけるフレーム単位の AUC 性能

3.4 ShanghaiTech データセットの実験結果

表 3 に ShanghaiTech データセットにおけるフレーム単位での AUC 性能を示す．ただし，分割サイズ l は 21 とした．提案手法は既存手法の中で精度が最も高い RTFM[7] に比べて 1.49%劣る結果となったが，他の既存手法については上回る精度を達成している．また，95%以上の精度が達成できていることから，十分に実用的な異常検出器が学習できていることを示している．

Method	Feature Type	AUC(%)
GCN-Anomaly [8]	TSN RGB	84.44
CLAWS Net [9]	C3D RGB	89.67
RTFM [7]	I3D RGB	97.21
Ours ($d_a = 64, r = 3$)	I3D RGB	95.72

表 3 ShanghaiTech データセットにおけるフレーム単位の AUC 性能

3.5 モデルのパラメータ数の比較

表 4 に各手法のモデルにおける学習可能なパラメータ数を示す．提案手法は既存手法に比べてパラメータ数が非常に少ない極めて軽量なモデルであることが分かる．$d_a = 64, r = 3$ のときは RTFM[7] の僅か 1.3%のパラメータ数であるにもかかわらず，3.3 章では RTFM[7] よりも高い精度を達成し，3.4 章では匹敵する高精度を達成している．また，更に高い精度が達成できた $d_a = 128, r = 7$ のときでも既存手法の中で最も軽量なモデルである．

Method	Number of Parameters
Sultani et al. [5]	2,114,113
Wu et al. [6]	769,155
RTFM [7]	24,718,849
Ours ($d_a = 64, r = 3$)	328,004
Ours ($d_a = 128, r = 7$)	721,992

表 4 各手法のモデルの学習可能なパラメータ数

3.6 ハイパーパラメータと検出精度への影響

我々は 2 章でハイパーパラメータとして d_a と r を定義した．ハイパーパラメータの変化によって提案手法の学習可能なパラメータ数は変化し，検出精度も変化する．そこで UCF-Crime データセットを用いて，ハイパーパラメータの変化が検出精度にどのように影響を与えるかを調査した．ただし，分割サイズ l は 32 とした．結果を図 2 に示す．赤で囲まれた点は各 d_a における最も高い検出精度を表す．r のみに注目すると，r が 3 よりも大きければ検出精度の変化は概ね少なく安定していることがわかる．d_a と r の関係に注目すると，d_a を大きくするなら r もある程度大きくすることでより良い検出精度が得られることがわかる．しかし，r を大きくし過ぎると検出精度が低下する．また，256 までは d_a を大きくすることでより良い検出精度を得ることができたが，d_a を 512 まで大きくすると r の値に関わらず，全体的に検出精度が低下した．これはモデルを大きくし過ぎることで過学習が起こっている可能性が考えられる．

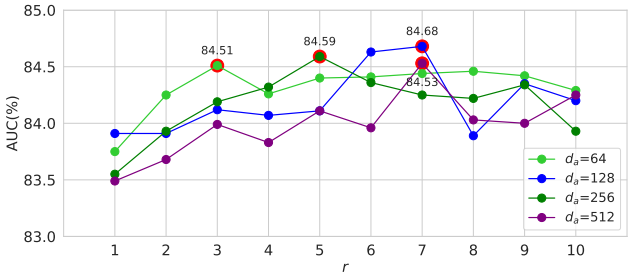

図 2 UCF-Crime データセットにおけるハイパーパラメータ d_a と r の変化と AUC 性能の関係

3.7 分割サイズの変化と検出精度への影響

我々は 2 章で分割サイズ l を定義した．提案手法は推論時に N 個の特徴ベクトルを $m = N/l$ 個のバッグに分けるため，各バッグには l 個の特徴ベクトルが入っている．分割サイズ l の値によって検出精度が変化するので，UCF-Crime データセットを用いて，l の変化が検出精度にどのように影響を与えるかを調査した．$d_a = 64, r = 3$ のときと $d_a = 128, r = 7$ のときの結果を図 3 に示す．赤で囲まれた点は各ハイパーパラメータにおける最も高い検出精度を表す．l が 16 前後になるまで検出精度が上がっている．提案手法は動画全体を解析する手法であり，ある程度の動画の長さを必要としているため，ある程度の分割サイズが必要なことがわかる．一方で，分割サイズ l が 16 よりも大きくなると安定した検出精度を出せており，提案手法は動画がある程度長ければ効率よく動画全体を解析できていることがわかる．

4. 結論

我々は，自己注意機構を用いて動画全体の特徴を解析して学習し，動画から異常を検出するための学習方法を提案した．提案手法は，既存手法で用いられていた MIL の代わりに入力の全セグメントから正常/異常の判定に需要な特徴量を自動抽出するための自己注意機構を導入する．また，提案手法のモデルは軽量でありながら既存手法と比較して遜色ない検出精度が達成できることを示した．

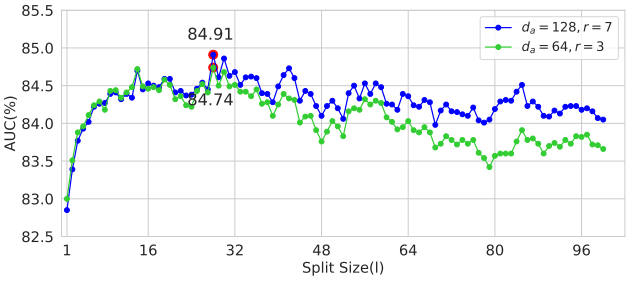

図 3 UCF-Crime データセットにおける推論時の動画分割サイズ l と AUC 性能の関係

参考文献

[1] A. Basharat, A. Gritai, and M. Shah, “Learning object motion patterns for anomaly detection and improved object detection,” in *CVPR*, pp. 1-8, 2008.

[2] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 fps in matlab,” in *ICCV*, pp. 2720-2727, 2013.

[3] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked rnn framework,” in *ICCV*, pp. 341-349, 2017.

[4] D. Gong, L. Liu, V. Le, B. Saha, M. Reda Mansour, S. Venkatesh, and A. van den Hengel, “Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection,” in *ICCV*, pp. 1705-1714, 2019.

[5] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in *CVPR*, pp. 6479-6488, 2018.

[6] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, “Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision,” pp. 322-339,in *ECCV* .Springer, 2020.

[7] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, “Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning,” in *ICCV*, 2021.

[8] J. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, “Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection,” in *CVPR*, pp. 1237-1246, 2019.

[9] M. Z. Zaheer, A. Mahmood, M. Astrid, and S. Lee, “Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection,” in *ECCV*. Springer, pp. 358-376, 2020.

[10] J. Feng, F. Hong, and W. Zheng, “Mist: Multiple instance self-training framework for video anomaly detection,” in *CVPR*, pp. 14009-14018, 2021.

[11] J. Carreira, and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Datase,” in *CVPR*, 2017.

[12] Z. Lin, M. Feng, C. N. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding,” in *arXiv* preprint, 2017.