H-059

行動予測を用いた犬の異常行動の検出手法の提案

高久優典† 田村仁†
日本工業大学大学院 工学研究科機械システム工学専攻†

1. はじめに

動物園やペットの飼育動物や野生動物の監視の作業を行ううえで，カメラ映像から自動的に状態を判別する技術が必要とされている．例えば嘔吐や失神などの非日常的な行動を判別した際に，従業員や獣医，飼い主に通知をすることで，可能な限り迅速に対応することが可能である．非日常的な状態を判別する方法として，機械学習による分類などがある．その一例である画像分類は，画像内の物体を検出し，カテゴリ分けを行う．

動物の監視においては，異常行動を検出したい．ここでは，「歩く」や「座る」などのような日常的にする行動を正常とし，「嘔吐」や「失神」などの非日常的な行動を異常とする．

機械学習による分類をするには，各カテゴリのデータ数をほぼ均等にする必要がある．各カテゴリのデータ数に偏りがある場合，学習精度がその偏りの影響を受けてしまう．しかし，異常行動は一般的には稀であるため，正常な行動と異常な行動でカテゴリ分けした場合，データ数に偏りが生じてしまう．

このようなデータに偏りがある場合には，一般的に異常検知を行う．異常検知とは，データの中から特徴的な外れ値を検出する手法である．そのため正常なデータのみで扱うことができ，異常なデータを必要としない．

この異常検知を機械学習で行ったものとして AnoGAN[1]という手法がある．AnoGAN には多数の画像から平均的な画像を生成する画像合成手法である敵対的生成ネットワーク(Generative Adversarial Networks: GAN)[2]を用いている．AnoGAN は部品の傷の検出などを目的としており，平均的な正常画像を 1 種類生成するため，動物のようにバリエーションの多い対象に対して扱うとすべての行動の平均的な画像を生成するため，利用することができない．

これの発展系として，GANomaly[3]がある．GANomaly は入力画像に対応した正常な再構築画像を複数種類生成することが可能である．

Proposal of Abnormal Dog Behavior Detection Method Using Action Prediction for Conference Presentation
†Yusuke Takaku, Hitoshi Tamura
Nippon Institute of Technology Graduate school Mechanical System Engineering Major

そのため動物に対応させることが可能であるが，動物の状態を検出するには静止画像だけでは限界がある．例えば，嘔吐する犬がいたとして，人間はその嘔吐を嗚咽する動作から予測することができる．しかし，この嗚咽する姿を静止画像でみると，前後の動きの情報がなくなり，単に下を向いて立ち止まっている姿勢に見えてしまう．[4]

そこで静止画像ではなく，フレーム前後の情報を含めて学習する動画認識や関節の動きなどの時系列データを用いる学習モデルを使うことで静止画像では取得できなかったような異常行動を取得する．動画認識は空港などの姿勢で不審な行動をしている人物を検出する手法として研究されている．対して時系列データを用いる場合は，1 人の対象の関節の動きからなどからデータを取得する．動画認識や時系列データによる学習モデルも，それぞれ得意不得意を持つ．本研究では，その特性を踏まえて動物の異常行動の検出に対応できる手法を提案する．

2. 各手法の検討

動画を用いて異常検知をする手法として大きく分けて 2 つある．それは動画認識と動画から必要な時系列データを取得して学習するモデルである．[5]

2.1 動画分類器による異常検知

一般的な動画分類器として C3D[6]がある．これは行動別に用意した動画を用いて学習する手法である．C3D は単に行動分類器であるため，異常検知を行うために異常検知手法を別途用意する必要がある．

正常な行動の動画を集め，行動別に分類し，学習する．その結果として C3D は正常な行動を分類する．この時，入力された行動が正常であれば，いずれかのクラスに対する確率が高くなる．対して入力された行動が異常であった場合，いずれのクラスに対しても確率が低迷する．このような結果になった行動を異常として検出する．

表 1　C3D の学習結果

	Loss	Accuracy
train	0.276106	0.894628
validation	0.476175	0.846774
test	0.570028	0.819355

学習を行った結果，C3D の行動分類器としての正解率は 0.81 となった．

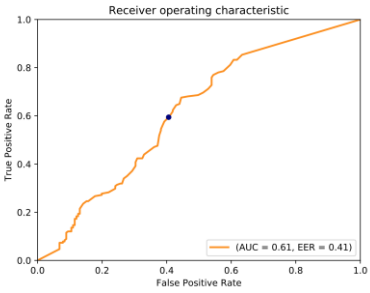

図 1　C3D の ROC

図 1 は，C3D で異常検知を行った結果の評価である．AUC は 0.61 となった．行動分類としての精度が高くないためそれに乗じて異常検知としての精度も低くなっている．また C3D の学習には正しく分類した動画が必要になるため目視による分類が必要であり，多くの時間を要する．

2.2　時系列データの分類器による異常検知

時系列データを扱う分類器としては Gated-Transformer Network(GTN)[7]がある．これは単変量時系列データを扱う Transformer[8]と異なり，多変量時系列データを扱うことができる．そのため，特徴点同士の関連性を考慮できる．犬の体の特徴点を MMpose[9]を用いて図 2 のように取得し，それを時系列データとして学習に用いる．

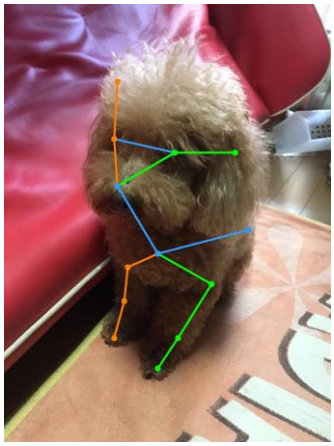

図 2　特徴点抽出の例

GTN も C3D と同じく分類器であるため，C3D と同様の方法で異常検知を行う．

表 2　GTN の学習結果

	正解率
max	53.12
average	30.63

表 2 は GTN の正解率である．その正解率は約 0.31 で留まった．この原因として GTN に用いたデータの中には時系列データのみでは姿勢が判別しにくいデータなどがあり，学習データに問題があったと考えられる．

2.3　フレーム予測による異常検知

フレーム予測を用いて異常検知を行う手法として，Future Frame Prediction for Anomaly Detection(FFPAD)[10]というものがある．これは GAN のネットワークを予測に用いている．学習した生成器による予測と実際の動作の差分を用いて異常を検知する．

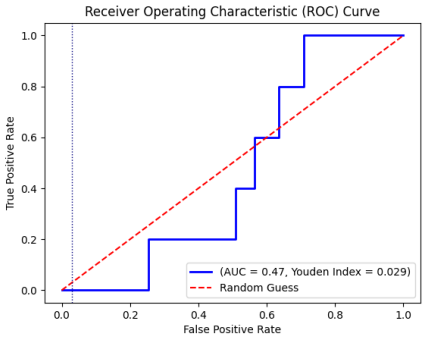

図 3　FFPAD の ROC

図 3 は FFPAD による異常検知の ROC である．結果から AUC の値は 0.47 と，低い結果であることがわかる．

図 4　誤検出の例[11]

図 4 は，FFPAD が正常な動画を異常と判定してしまった一例である．誤判定した原因として FFPAD は動画を用いており，犬とは関係のない要素も学習に含まれる．そのため，特徴が取りやすい直線状のものや揺れる木，動画撮影者の手振れの影響を大きく受けるためと考えられる．

3.　提案手法

2 章で触れたそれぞれの欠点を踏まえて私は Space-Time-Separable Graph Convolutional

Network(STSGCN)[12]を用いた異常検知手法を提案する．STSGCN は入力された時系列データの続きの値を予測する．その予測結果と実際の値を比較することで異常を判別する．STSGCN は時系列データを扱うため，FFPAD のように犬とは関係の情報に影響されることがない．また GTN は分類器であるのに対し，STSGCN は時系列データの予測を行う．そのため予め目視で行動別に分類する必要がない．本原稿では，STSGCN を犬の時系列データに適応させ，正しく動作するかの確認を行う．

4. 実験方法

本研究では，対象を犬として学習を行った．対象とした動画は GTN での結果などを踏まえて，犬が横を向いた動画のみとした．これにより不明な特徴点を可能な限り減らした．また犬の横向きに関しては目視で確認を行った．

収集した動画にシーンの切り替えが存在した場合は学習に大きな影響を与えるため，シーン切り替えが発生した場合はそのフレームで動画を分割した．

次に，収集した動画から Yolo[13]を用いて犬のバウンディングボックスを取得し，その中心を中央に固定する．これにより画角による犬の位置の影響や手振れの影響をなくすことができる．長尺の動画は複数の行動を含むため，30[fps]を 2 秒のフレーム数 60 枚の動画に分けてデータとした．

最後に，MMPose を用いて特徴点の座標を取得する．この際，フレームが 1 フレームまたは 2 フレーム連続で抜け落ちた時のみは平均をとることで補った．最後に 60 フレーム中 50 フレーム以上犬を検出できた時系列データのみを採用した．結果として集まったデータは 13,208 本である．

5. 実験の結果

本実験では，エポック数を 50，バッチサイズは 256 で学習を行った．

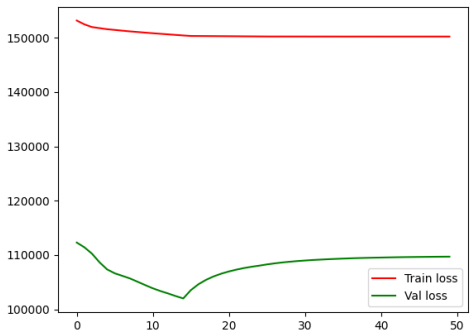

図 5　STSGCN の損失関数

図 5 は学習した際の Mean Per Joint Position Error(MPJPE)の損失関数である．MPJPE は関節点の推定した座標と実際の座標の距離を全ての関節点で平均をとることで算出される評価指標である．横軸はエポック数，縦軸はミリメートルである．図 5 から Train と Validation に 4 メートルほどのズレが発生している．このズレに関しては，6 章の考察にて後述する．

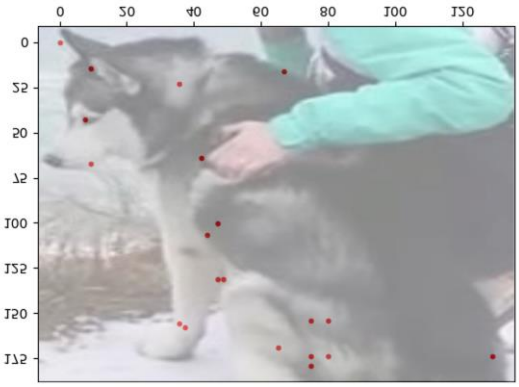

図 6　正解データの分布の確認[14]

図 6 は確認のため，20 ヶ所の特徴点と元となった動画のおおむね近いフレームと重ねたものである．Python スクリプトの都合上，グラフを上下反転させている．目の位置や，背中，尻，前足や後ろ足に座標が集まっていることから正しく分布できることがわかる．

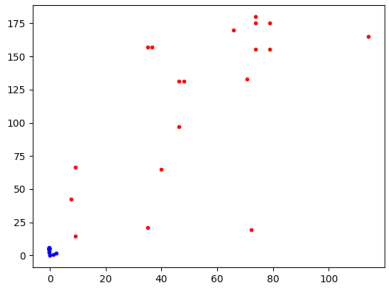

図 7　正解と予測の分布

図 7 は 1 フレームの正解データと予測データの分布である．赤い点が正解の座標，青い点が予測の座標である．横軸，縦軸ともに単位はメートルである．広く分布している正解データに対し，予測データは端に偏っており，十分な精度が出ていないのが分かる．

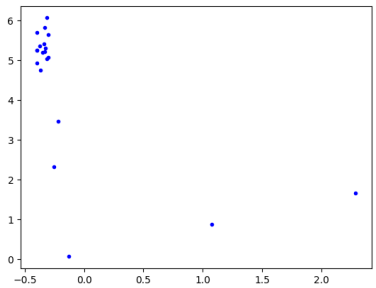

図 8　予測の分布

図 8 は図 7 のうち，青い予測データのみを拡大した図である．予測データは左上に多く集まっており，正解データとは大きく異なる形状をしている．

6. 考察

今回の実験から，正常なデータすら予測することができていないため異常検知には，現状適応することができない．加えて，学習に用いるデータを間違えていることが判明した．本来はメートル単位で学習するべきところを画像のピクセル単位の座標で学習してしまった．そのため，学習データに体長 200 メートルの犬とされる座標が含まれていることになる．よって MPJPE の結果と大きくずれており，参考にし難い結果となってしまった．

図 5 の Validation の損失関数を見るに値は明らかに変化し，収束しているため学習自体は問題なく行えていることがわかる．このことから正しくメートル単位の学習データを用いることで正しく予測できることが期待できる．しかし，図 7 を見るとそもそもの正解データにも誤差が存在するため，現状の時系列データから正しいメートル単位に修正をしても十分な精度が出ない可能性も考えられる．

7. 今後の予定

STSGCN に使う時系列データの正しい形状への加工と，現状のデータセットに含まれている正しくないデータへの対処などを予定している．加えて時系列データをベクトルに置き換えることで時系列予測や異常検知に対応できることから，そちらの使用も考えている．[15]また，昨今では膨大なデータで事前学習をしたのち，少量の正確なデータでファインチューニングを行うことでパラメータを調整する手法も注目されており，それを用いることで課題となっている学習データの不足を補うことも考えている．

参考文献

［1］ Thomas Schlegl , Philipp Seeböck , Sebastian M.Waldstein , Ursula Schmidt-Erfurth and Georg Langs , ”Unsupervised Anomaly Detection with [Schlegl]Generative Adversarial Networks to Guide Marker Discovery”, Lecture Notes in Computer Science（LNCS）. pp. 146-147, 2017

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. Advances in Neural Information Processing Systems 27 (2014)

［3］Samet Akcay，Amir Atapour-Abarghouei and Toby P. Breckon , ”GANomaly: Semi-Supervised Anomaly Detecion via Adversarial Training” , arXiv:1805.06725v3 [cs.CV] , 2018

[4] Yusuke Takaku and Hitoshi Tamura, “GANomaly と動物の画像を使った異常行動の検出”, Information Processing Society of Japan, 情報処理学会第 85 回全国大会講演論文集(2), pp. 243-244, 2023

[5]Yusuke Takaku and Hitoshi Tamura, “機械学習による動画解析を用いた 4 足動物の 行動の異常検知手法の検討”, FIT2023：第 22 回情報科学技術フォーラム：講演論文集　第 3 分冊, pp.143-146, 2023

[6] ran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp.4489–4497, 2015

[7] M. Liu, S. Ren, S. Ma, J. Jiao, Y. Chen, Z. Wang, and W. Song, “Gated Transformer Networks for Multivariate Time Series Classification 2021”.

[8] hish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[9] Arindam Sengupta,Feng Jin, Renyuan Zhang and Siyang Cao,”mm-pose: Real-Time Human Skeletal Posture Estimation using mmWave Radars and CNNs, IEEE Sensors Journal 20(17), 10032–10044 (2020)

[10] W. Liu, W. Luo, D. Lian and S. Gao, "Future frame prediction for anomaly detection-a new baseline", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[11]YouTube, https://www.youtube.com/watch?v=HlWhRYM3bbg, (2024/06/14 閲覧)

[12] Song, C., Lin, Y., Guo, S., & Wan, H. (2020). Spatial-Temporal Synchronous Graph Convolutional Networks: A New Framework for Spatial-Temporal Network Data Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 914-921.

[13] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. “You only look once: Unified, real-time object detection.”, CoRR, abs/1506.02640, 2015.

[14]YouTube, https://www.youtube.com/watch?v=eFIxrIN01Hg&t=576s, (2024/06/14 閲覧)

[15]Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, Boxiong Xu, “TS2Vec: Towards Universal Representation of Time Series”,Proceedings of the AAAI Conference on Artificial Intelligence, 2022