H-007

# 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

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

動物園や野生動物の監視において，作業の効率化にはカメラ映像から自動的に状態を判別すること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観測した場合に，従業員や獣医に通知することで，迅速に対応することができる．このような状態を判別する方法として，機械学習による画像分類がある．画像分類とは，画像内の物体を検出し，カテゴリ分けする技術のことである．

特に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定義する．

機械学習で画像分類をするには，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の偏りによって精度に影響が生じるためである．しかし，異常行動は一般的には稀な行動であるため，正常な行動と異常な行動のカテゴリ分けする際に，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，適した方法として異常検知という技術がある．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．この異常検知を得意とする機械学習手法として AnoGAN[1]という手法がある．AnoGAN 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の敵対的生成ネットワーク(Generative Adversarial Networks 略称:GAN)[2]を用いている．しかし，AnoGAN は部品の傷などの検出することを目的としており，平均的な正常画像を 1 種類だけ生成する．そのため，入力画像に動物の行動のようなバリエーションが多いものを使用すると，入力された行動全ての平均的な画像を生成してしまうため，個別の姿勢画像とは離れたものになってしまう可能性がある．

それに対して GANomaly[3]では，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できる．

しかし，動物の状態検出をするには静止画像だけでは限界がある．例えば，人間は動物が嘔吐する可能性を，嗚咽という動作から予測することができる．しかし，画像だけ見ると嗚咽の前後の動きの情報が欠落しており，単に下を向いて立ち止まる姿勢に見えてしまう．

姿勢を推定する手法以外に，連続するフレームを利用して行動を推定する動画認識がある．動画認識では，フレーム前後の動きも含めて学習するため，動作に関する情報が必要な場合にも対応できる．そのため，動物の行動検出において，誤認を減らすことができると考える．

## 2. 関連研究

先行研究として，GANomaly を用いて動物の正常と異常の判別をする研究[4]がある．この研究では，正常と異常の 2 つのカテゴリを持つ学習データを用いている．結果として，AUROC は 0.97 で正解率(Accuracy)は約 0. 91 であった．しかし，この研究では静止画像を使用しており，動画ではない．そのため，連続した情報は欠落しているため，行動というよりは姿勢による判定といえる．またこの研究では，嘔吐している姿勢と立ち止まった姿勢を誤認している．これは，連続した情報が欠落したことが判定を誤った原因と考えられる．

動画認識を用いた異常検知は，主に空港や雑踏などの人が多く集まる場所で，テロなどを防ぐために不審な行動をとった人物を検出する技術として盛んに研究されている．

動画認識を用いた異常検知として，Reem 氏らによる研究では，C3D[5]を用いてイスラム教の巡礼による人混みの中で立ち止まっている人や座り込んだ人などを異常として検出する試みが提案された[6]．しかし，これは予め異常とする行動を定義づけ，単に分類しているだけである．

### 2.1 動画認識による分類手法

動物などの動作の分類手法として，動画そのものを使用して学習する手法と，動物の関節位置座標を取得して学習する手法がある．

---

Investigation of Anomaly Detection Method for Quadruped Animal Behavior Using Machine Learning-based Video Analysis

†Yusuke Takaku，Hitoshi Tamura

Nippon Institute of Technology Graduate school Mechanical System Engineering Major

#### 2.1.1　動画を用いて学習する手法

動画をそのまま用いて学習する手法は，十分な数の動画データを集めることで学習することができる．代表的なものとして C3D や TimeSformer[7]などがある．

C3D は，Du 氏らによって提案された 3 次元の畳み込みニューラルネットワークを使用したビデオ解析のための特徴量抽出手法である．ただし，C3D は動画の分類手法であり，異常検知の手法ではない．そのため，正常行動と異常行動を事前に定義し，データを分類する必要がある．

TimeSformer は，Gedas 氏らによって提案された手法で，ビデオ分類を目的とした Transformer[8]である．TimeSformer は空間と時間の両方を用いて学習する．しかし，この TimeSformer もまたビデオ分類を目的とし，異常検知を目的としていない．また[7]は巨大なデータセットである Kinetics400[9]を使用しており，ローカル PC では実行が困難である．

#### 2.1.2　関節位置座標を用いて学習する手法

関節位置座標の取得に使われる代表的な手法として OpenPose[10]などがある．しかし OpenPose は人体を対象としており，動物を対象とした，姿勢推定手法やデータセットは比較的少数である．動物の関節位置座標は DeepLabCut[11]や DeepSnake[12]，MMpose[13]などを用いて取得することができる．関節位置座標のデータと動画データを組み合わせて学習を行う．

Minghao 氏らが提案した Gated Transformer[14]は，2 つの Transformer を用いた手法で，1 つは各関節の位置関係ごとの相関関係をモデル化し，もう 1 つは 1 つの関節の時間的な相関関係をモデル化している．複数の入力を組み合わせることで高い精度を実現する．しかし，現状ある Gated Transformer は，時系列と 1 軸方向のみのデータで学習する．今回の関節位置をデータとしたとき，時間軸とそれぞれ x，y，z の 3 つをとることになる．

図 1 は，MMpose を用いて犬の関節位置座標をデータとして取得したデータの例である．MMpose は 4 足歩行動物の学習済みモデルが提供されていて，任意の動画像を入力することで関節位置座標を取得できる．図 1 では，13 点の座標を取得されている．データ数が 13 点の 3 軸で 39 次元のデータを使用することになる．図 1 は正面からの視点で取得しているため，後ろ脚の情報が欠如している．このようにカメラの位置や画角などの影響を受けてしまうため，時系列データとして安定性に欠ける部分がある．

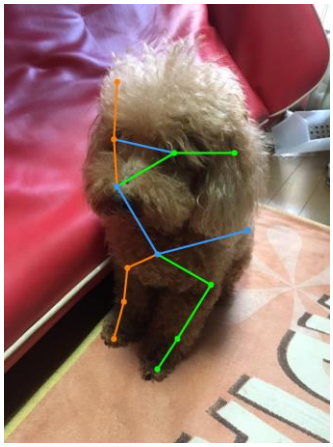

**図 1 関節位置座標の一例**

この Gated Transformer を野生動物に適応した例として，ShiChao 氏らの研究がある[15]．この研究では野生の哺乳類の 2D 画像を用いて，ギャロッピング，アンブリング，座る，歩くの 4 つの動作を対象として分類する．しかし，この研究非日常的な異常行動は対象とはしていない．そのためデータセットに大きな偏りが出ずに学習が可能である．また 2D 画像を用いているため，奥行き方向の情報は考慮していない．

## 3. 提案手法

C3D と Gated Transformer をベースに，異常検知を行う．C3D を使用する場合は，正常と異常で分類するのではなく，正常行動と異常行動を定義して分類する．Gated Transformer で分類する場合は，現状 1 軸しか用いていないため Gated Transformer を x，y，z の 3 次元に拡張しなければならない．一般に監視カメラ映像では，画角に対して動物の行動の方向は固定されず，3 次元的に解釈できなければならないためである．

今回は，関節位置座標などを使用せず動画のみで学習が可能な C3D で十分な精度が出るかを確認する．C3D で十分な精度が出ない場合は，Gated Transformer での検知を試みる．

### 3.1　対象とする行動

本研究では，4 足歩行動物として犬を用いる．犬は世界中で飼われており，データの収集が容易である．学習に用いた動画の条件を，表 1 に示す．

**表 1　入力に用いた動画の条件**

| 条件 |
|---|
| 2匹以上いない |
| 対象が隠れていない |
| 見切れていない |
| 人間が動体に干渉していない |

表 1 の人間が動体に干渉していないというのは，人間が強くリードを引いて歩かせる，無理やり伏せさせるといった強制的にその姿勢にさせることを指す．犬の動作は，表 2 のように 4 つの動作とその他の合計 5 つに定義する．

**表 2　動作の定義**

| 動作 | 定義 |
|---|---|
| 歩く | 前脚、後ろ脚が連続的に入れ違う。 |
| 座る | 臀部が地面についている。 |
| 伏せる | 胴体が地面についている。 |
| 立ち止まる | 脚が全て地面についている。 |
| その他 | |

### 3.2　作成するデータセット

既存のデータセットとして Kaggle が提供している Dog Behavior Analysis Dataset[16]がある．しかし[16]は動物の関節位置座標を csv ファイルのみで動画ファイルが存在しないため，C3D に適応することができない．他にも Stanford Dogs Dataset[17]がある．しかし[17]は画像データであるため，C3D と Gated Transformer の学習に用いることができない．

そのため，データセットを作るうえで YouTube や飼い犬の動画などを用いて作成した．

犬を正確に認識するために，Yolov5[18]を用いて犬が映り込んだフレームのみを抽出した．動画内にシーン切り替えが存在する場合，連続画像を学習する C3D にとって問題となるため，シーンが切り替わったタイミングを動画の区切りとして avi ファイルで分割し保存した．図 2 に，その分割例を示す．


**図 2　動画の分割**

図 2 で分割した動画を目視で確認し，表 1 と表 2 の定義に当てはまるように分類した．

最終的に集まったデータの数は，表 3 のとおりである．

学習回数は 200 エポック，バッチサイズは 8 とした．また評価指標として Accuracy を用いた．

## 4.　実験結果

各エポックでの損失関数と正解率を測定した．その結果を図 3 と図 4 に示す．

**表 3　データセット**

| 動作 | 動画数 | 総フレーム数 |
|---|---|---|
| 歩く | 162 | 18923 |
| 座る | 142 | 19045 |
| 伏せる | 154 | 18690 |
| 立ち止まる | 144 | 18893 |
| その他 | 215 | 18886 |

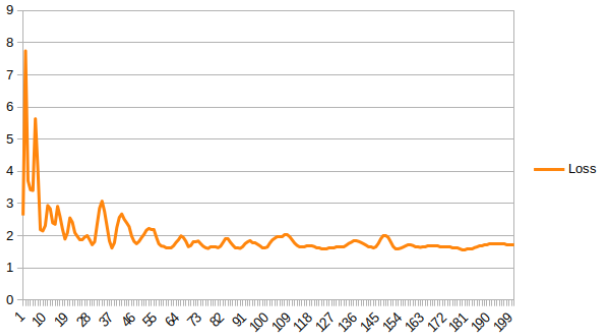


**図 3　損失関数**

図 3 は，各エポック数における損失関数である．損失関数は緩やかに減少しているのが分かるが，1.5 を下回ることはなかった．

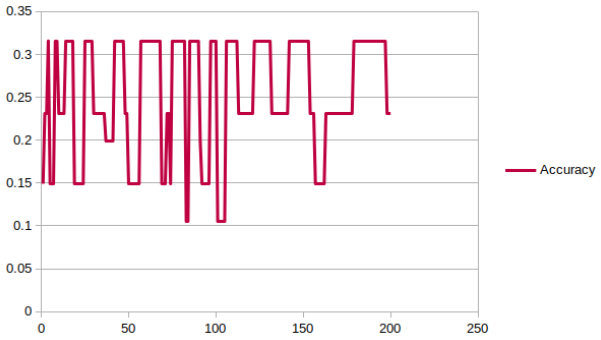


**図 4　正解率**

図 4 は各エポック数における正解率の変位である．0.1 から 0.32 の間を不規則に上下しているのが分かる．

## 5.　考察

C3D の学習が不十分である結果になった．多種多様なデータが入っているその他が正解率を落とす原因と考え，その他を除外した 4 択で学習させた結果，200 エポック目の時の正解率は 0.26 であった．5 択の時の 200 エポック目の正解率は 0.23 であり，その他の有無による影響はなく，ランダムな回答をしていると考えられる．その他は異常行動が含まれている可能性があるラベルであり，その他の有無に関わらず認識率が低いことから，異常行動が認識率低下の原因であるとは考えられない．異常行動が学習に影響していないことが考えられる．原因はデータセット

に問題があると考え，入力に使われる tfrecord ファイルを中から 1 フレームを抽出し正しく犬が映っているかを確認した．

**図 5　test.tfrecord 中の 1 フレーム**

図 5 は，学習に使われた test.ftrecord の分割された動画[19]のうちの 1 フレームである．実際に中央に犬が映っているのが確認できる．そのため，学習の際に犬の情報に欠落があるとは考えにくい．しかし全てのカテゴリの動画内の多くに歩いている人間が映り込んでいた．これが原因で C3D は動物ではなく，歩いている人間を判定の要因としてしまっている可能性がある．今回の結果では学習にされた動画そのものに問題があったため，C3D の性能を十分に発揮できなかった可能性が考えられる．そのため，データセットを改めて見直し C3D で学習させる必要がある．

## 7. 今後の予定

現状のデータセットでは十分な精度が出なかったため，データセットを見直し，再加工して正しく判定ができるようにしようと考えている．また，C3D は単純な動画分類を行っており，正常と異常で判定しているわけではない．そのため，人間が直接動画を見て手作業で正解データを分類する必要がある．設けた行動以外をその他として扱っているため，カテゴリ分けされていない正常行動を異常として判定する可能性もある．分類問題は各カテゴリの正解データの数を均等にする必要があり，発生率が少ない異常な行動を集めるのは困難である．そのため，GANomaly の手法を C3D や Gated Transformer などに対応させることを予定している．

## 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] 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

[6] Reem Alharthi, Areej Alhothali, Bander Alzahrani, Sahar Aldhaheri, “Massive Crowd Abnormal Behaviors Recognition Using C3D”, 2023 IEEE International Conference on Consumer Electronics(ICCE),2023

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In ICML, 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] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017.

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7291–7299, 2017.

[11] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience, 2018.

[12] Peng, S., Jiang, W., Pi, H., Li, X., Bao, H., Zhou, X.: Deep snake for real-time instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8533–8542, 2020

[13] 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)

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

[15] ShiChao Deng, Guizhong Tang, Lei Mei, “Wild Mammal Behavior Recognition Based on Gated Transformer Network”, 2022 International Conference on Cyber-Physical Social Intelligence(ICCSI), 2022

[16] Kaggle, Dog Behavior Analysis Dataset, URL: Dog Behavior Analysis Dataset | Kaggle, 2023/06/15 閲覧

[17] Kaggle, Stanford Dogs Dataset, Stanford Dogs Dataset | Kaggle, 2023/06/15 閲覧

[18] 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.

[19] YouTube, National Dog Show 2022: Working Group (Full Judging) | NBC Sports, URL, https://www.youtube.com/watch?v=2K623IdQaVk&t=593s, 2023/06/14 閲覧